In the electronic version use the bookmarks / table of contents in the Navigation Panel in order to help you find what you want.
You may find it helpful to print out selected portions, such as the Table of Contents.

Information in bold is more important to pass your exam. Information presented in italics (and subsections whose titles are in italics) should not be needed to directly answer exam questions. It is provided to aid the reader’s overall understanding of the subject.

I have doubled underlined highly recommended questions to do on your first pass through the material, underlined recommended questions to do on your second pass, and starred additional questions to do on a third pass through the material. No questions were labeled from the 2011 exam or later, in order to allow you to use them as practice exams.

Changes to the Exam 8 syllabus for 2019:

Exam 8 will not contain multiple-choice questions. All questions will require an essay response.

My solutions to questions are intended to be model solutions. Often they are more detailed and contain more explanation than would be needed in order to get full credit. This was done in order to give you a clearer and better understanding of the subject material.

Make sure you write legibly; the more organized your work the better.

After you have done one of the recent exams, be sure to look closely at the CAS Examiner’s Report. See the sample solutions in the Examiner’s Reports, and read the comments of the examiners.

On numerical questions clearly label your final answer and show enough intermediate steps so someone can follow what you did. If there are a series of parallel calculations you can show just one of them. Of course, if writing down more helps you to do the problem, do so.

In the case of verbal questions, do not concentrate on grammar or complete sentences. Feel free to list or outline your ideas. The selected use of abbreviations can save some time.

In any case, remember that spending considerable additional time to increase 80% credit to 100% credit on a single question is usually not a good use of your limited exam time. You can come back later to a question, if you have the time.

---

1 Obviously feel free to do whatever questions you want. This is just a guide for those who find it helpful.
2 I have removed the multiple choices from most past exam questions that had them.
3 In some cases, I even quote a reading word for word. This does not mean you need to be able to do so!
4 Sometimes much of what I say is directed at those students who did not answer a question correctly and need to learn more. In any case, you want to know as much as possible to help answer the question that will be on your exam, as opposed to whatever they happened to have asked on some past exam.
As stated in the CAS Syllabus: “The model response to the typical essay question is brief, less than one-half of a written page. Be concise — candidates do not need to answer in complete sentences when a well-composed outline format is more appropriate. Candidates should not waste time on obscure details. They should show that they have learned the relevant material and that they understand it. They should state the obvious, if it is part of the answer.”

Also read “The Importance of Adverbs on Exams,” in which the Exam Committee notes the difference between: briefly discuss, discuss, and fully discuss. Brief descriptions, discussions, etc., are worth 1/4 point. (Unmodified) discussions or descriptions are worth 1/2 point. Full descriptions or discussions are worth at least 1 point. Please look carefully for these word choices and point values on all CAS upper-level exams. Most importantly, answer the question in accordance with the amount of information being asked.

The CAS is gradually moving towards an integrative testing framework. **Integrative Questions (IQs)** will require candidates to understand multiple facets of the syllabus material and concepts in addressing complex business problems in a single exam question. IQs will differ from a typical exam question in three significant ways.

1. An IQ will be worth more points. One IQ could be worth 10-15% of the total exam.
2. Each IQ will require candidates to draw from multiple syllabus learning objectives in order to answer the question.
3. IQs will test at a higher average Bloom’s Taxonomy level than a standard exam question.

To assist candidates with preparing to answer an IQ, the CAS released sample IQs and responses. It should be noted that while the samples were constructed in parallel with the IQ that will appear on the exam, they may not be structured in the same manner nor cover the same learning objectives as the exam question. It is advised that candidates use the samples to validate preparation and identify potential areas for improvement after completing the majority of their study, rather than using them during their initial study as one might use textbook exercises.

**Exam 8 featured one IQ on the Fall 2017 exam, and two on the Fall 2018 exam.**

It is expected that Exams 7, 8, and 9 will continue to include IQs in future sittings, and the number of IQs that will appear on the exams will gradually increase over time. At the same time, there will be fewer exam questions overall to account for the presence of IQs in order to avoid any increase in the time length of the exam. There will be no change to the normal grading process, as described in the Syllabus, for IQs.

My study guide includes question written by me, and some by Sholom Feldblum. In addition, the former exam questions are arranged in chronological order. The more recent exam questions are on average more similar to what you will be asked on your exam, than are less recent questions.

---


6 See the CAS webpage for updated information.

7 I thank Sholom Feldblum for the kind permission to use his material. Any mistakes are my responsibility.
It is important that you do problems when learning a subject and then some more problems a few weeks later.

As you get closer to the exam, the portion of time spent doing problems should increase.

There are two manners in which you should be doing problems. First you can do problems in order to learn the material. Take as long on each problem as you need to fully understand the concepts and the solution. Reread the relevant syllabus material. Carefully go over the solution to see if you really know what to do. Think about what would happen if one or more aspects of the question were revised. This manner of doing problems should be gradually replaced by the following manner as you get closer to the exam.

The second manner is to do a series of problems under exam conditions, with the items you will have when you take the exam. Take in advance a number of points to try based on the time available. For example, if you have an uninterrupted hour, then one might try $60 / 4 = 15$ points of problems. Do problems as you would on an exam in any order, skipping some and coming back to some, until you run out of time. Leave time to double check your work.

It is important that you develop the skill of quickly and clearly writing down what you know. Many of you will benefit by giving some of your solutions to questions to someone else to “grade”. They should give you feedback on whether they were able to follow what you did. They should point out where you wrote more than was necessary or not enough.

Read the “Hints on Study and Exam Techniques” in the CAS Syllabus.

The CAS has posted a pdf on Bloom’s Taxonomy of question writing. You might want to look at it.

I thank Sholom Feldblum for the kind permission to use his material.
I thank Roger Wilk for kindly providing me with his notes on Clark, and Grossi & Kunreuther. Finally, thanks to the many past students who have helped me to improve these study guides.

**Sold separately are my seminar style slides.** They are electronic and you can print them out.

**Feel free to send me any questions or suggestions: hmahler@mac.com**
Please send me any suspected errors by Email. (Please specify as carefully as possible the page and Exam number.)

I will post a list of errata on my webpage: www.howardmahler.com/Teaching

---

8 Someone else taking this exam or who has just passed this exam would be a good choice.  
9 On average you get less credit on essay questions when graded by someone else then when you self-grade.
Reading Period:

“A 15-minute reading period will be given immediately prior to the start of the exam. The reading period will begin at the published start time of the exam. After candidates have read the instructions, the proctor will give candidates 15 minutes to silently read the questions. Candidates may wish to use this time to ensure that they do not have a defective exam and to develop a strategy for which questions they will attempt first. Writing will NOT be permitted during this time and candidates will not be permitted to hold pens or pencils or use calculators. At the end of 15 minutes, the proctor will give the instruction for candidates to begin the exam. This will mark the beginning of the timed section of the exam.”

Your exam will be 4 hours in addition to the 15 minute reading period.10

Preparing for a CAS Exam--what to do with hard material
by Dr. J. Eric Brosius, FCAS

The syllabus for a typical CAS exam includes both easy and hard material. Many students learn the easy material well, but adopt less-than-optimal strategies for learning the hard material. Some spend a lot of time trying to understand syllabus readings that are nearly incomprehensible. Others ignore the more difficult readings altogether. Neither approach is a good idea, not if you hope to pass! I will suggest a better way to approach these readings. Your goal in studying is not to understand the material in general but to be able to answer the questions. Do not study the syllabus readings in a vacuum; consider also what types of questions are likely to be asked. Each exam contains both easy problems and hard problems.

10 Check the CAS webpage for any updated information.
We can divide the problems into four categories based on the difficulty of the material and the difficulty of the problem, as follows:

Box 1 contains easy problems on easy material. These are easy to answer; unfortunately, there are not enough of them!

Box 2 contains hard problems on easy material. You can prepare for these by practicing problems from old tests and other sources of sample problems.

Box 3 contains hard problems on hard material. Few students can afford to spend the time required to answer all of these. Fortunately, the Examination Committee does not ask many of these question: even if they understand the reading well enough to do so, there isn't much point in a question that no one can answer. Be prepared to skip Box 3 problems if necessary.

Box 4 contains easy problems on hard material. These problems can supply the extra points you need to change a "5" into a “6”. They appear often, because the Examination Committee tends to ask easy questions about hard readings. When a reading is technically difficult, and especially if it was recently added to the syllabus, even the simplest question poses a challenge. Study these readings with an eye to answering the obvious questions. It is a shame not to get points for a question that could have been answered if only you had read the first paragraph of the reading.

Plan for your exam in such a way that you focus on Box 2 and Box 4. Prepare for Box 2 questions by studying the easy material in detail, and by doing many sample problems. Prepare for Box 4 questions by outlining the high points of the material, and by trying to guess, alone or with other students, what questions on this material might appear on the exam.
Since 2011, the points on exam questions are similar to the present. Going back a few more years further in time, a 5 point exam question might only be worth 3 points today.\textsuperscript{11}

\begin{tabular}{|c|c|c|c|c|}
\hline
Exam 8 & Points & Number of Questions & Integrated Questions & Average \% of Exam per Integrated Question \\
\hline
2011 & 59 & 25 & & \\
\hline
2012 & 54.75 & 23 & & \\
\hline
2013 & 57.5 & 25 & & \\
\hline
2014 & 60.25 & 25 & & \\
\hline
2015 & 59.5 & 23 & & \\
\hline
2016 & 53.25 & 21 & & \\
\hline
2017 & 53.75 & 20 & 1 & 15.8\% \\
\hline
2018 & 52 & 17 & 2 & 17.8\% \\
\hline
\end{tabular}

\begin{tabular}{|c|c|c|c|c|}
\hline
Exam 8 & Pass Mark & Percent of Available Points & 95th Percentile & 75th Percentile \\
\hline
2011 & 43.75 & 74.15\% & 47.38 & 43.00 \\
\hline
2012 & 37.75 & 68.95\% & 44.25 & 39.75 \\
\hline
2013 & 40.75 & 70.87\% & 47.50 & 43.63 \\
\hline
2014 & 37.50 & 62.24\% & 44.50 & 40.63 \\
\hline
2015 & 40.75 & 68.49\% & 48.50 & 43.13 \\
\hline
2016 & 37.25 & 69.95\% & 42.88 & 38.88 \\
\hline
2017 & 37.5 & 69.77\% & 43.00 & 39.50 \\
\hline
2018 & 33.75 & 64.91\% & 47.38 & 43.00 \\
\hline
\end{tabular}

\textsuperscript{11} For my problems, it depends on when I wrote them. My older ones are probably more like the older exam questions as far as points go. I am sorry that my study guides are not more consistent with respect to “points”.
<table>
<thead>
<tr>
<th>Exam 8</th>
<th>Exams Taken</th>
<th>Passed</th>
<th>Raw Pass Ratio</th>
<th>Effective Pass Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>418</td>
<td>93</td>
<td>22.2%</td>
<td>23.9%</td>
</tr>
<tr>
<td>2012</td>
<td>519</td>
<td>218</td>
<td>42.0%</td>
<td>43.7%</td>
</tr>
<tr>
<td>2013</td>
<td>592</td>
<td>283</td>
<td>47.8%</td>
<td>49.3%</td>
</tr>
<tr>
<td>2014</td>
<td>729</td>
<td>350</td>
<td>48.0%</td>
<td>50.2%</td>
</tr>
<tr>
<td>2015</td>
<td>771</td>
<td>313</td>
<td>40.60%</td>
<td>42.18%</td>
</tr>
<tr>
<td>2016</td>
<td>791</td>
<td>301</td>
<td>38.05%</td>
<td>40.13%</td>
</tr>
<tr>
<td>2017</td>
<td>945</td>
<td>376</td>
<td>39.8%</td>
<td>41.7%</td>
</tr>
<tr>
<td>2018</td>
<td>953</td>
<td>314</td>
<td>32.9%</td>
<td>35.1%</td>
</tr>
</tbody>
</table>

One measure of the difficulty of an exam is the ratio of the 75\textsuperscript{th} percentile to the available points:

<table>
<thead>
<tr>
<th>Exam 8</th>
<th>Points</th>
<th>75th Percentile</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>59</td>
<td>43.00</td>
<td>72.9%</td>
</tr>
<tr>
<td>2012</td>
<td>54.75</td>
<td>39.75</td>
<td>72.6%</td>
</tr>
<tr>
<td>2013</td>
<td>57.5</td>
<td>43.63</td>
<td>75.9%</td>
</tr>
<tr>
<td>2014</td>
<td>60.25</td>
<td>40.63</td>
<td>67.4%</td>
</tr>
<tr>
<td>2015</td>
<td>59.5</td>
<td>43.13</td>
<td>72.5%</td>
</tr>
<tr>
<td>2016</td>
<td>53.25</td>
<td>38.88</td>
<td>73.0%</td>
</tr>
<tr>
<td>2017</td>
<td>53.75</td>
<td>39.50</td>
<td>73.5%</td>
</tr>
<tr>
<td>2018</td>
<td>52</td>
<td>35.00</td>
<td>67.3%</td>
</tr>
</tbody>
</table>

The lower the ratio of the 75\textsuperscript{th} percentile to the available points, the harder the exam.
Mahler’s Guide to Advanced Ratemaking
Copyright ©2019 by Howard C. Mahler.

Information in bold or sections whose title is in bold are more important for passing the exam. Larger bold type indicates it is extremely important. Information presented in italics (including subsections whose titles are in italics) should not be needed to directly answer exam questions and should be skipped on first reading. It is provided to aid the reader’s overall understanding of the subject, and to be useful in practical applications.

I have doubled underlined highly recommended questions to do on your first pass through the material, underlined recommended questions to do on your second pass, and starred additional questions to do on a third pass through the material.¹ No questions were labeled from the 2011 exam or later, in order to allow you to use them as practice exams.

Solutions to problems are at the end of each section.²

¹ Obviously feel free to do whatever questions you want. This is just a guide for those who find it helpful.
² Note that problems include both some written by me and some from past exams. The latter are copyright by the Casualty Actuarial Society and are reproduced here solely to aid students in studying for exams. The solutions and comments are solely the responsibility of the author; the CAS bears no responsibility for their accuracy. While some of the comments may seem critical of certain questions, this is intended solely to aid you in studying and in no way is intended as a criticism of the many volunteers who work extremely long and hard to produce quality exams. There are also some past exam questions copyright by the Society of Actuaries.
<table>
<thead>
<tr>
<th>Section #</th>
<th>Pages</th>
<th>Section Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9-101</td>
<td>Mahler, An Example of Credibility and Shifting Risk Parameters</td>
</tr>
<tr>
<td>2</td>
<td>102-201</td>
<td>Bailey &amp; Simon, Credibility of a Single Car</td>
</tr>
<tr>
<td>3</td>
<td>202-516</td>
<td>Goldburd, Khare and Tevet, Generalized Linear Models</td>
</tr>
<tr>
<td>4</td>
<td>517-547</td>
<td>ASOP 12: Risk Classification</td>
</tr>
<tr>
<td>5</td>
<td>548-634</td>
<td>Robertson, NCCI's 2007 Hazard Group Mapping</td>
</tr>
<tr>
<td>6</td>
<td>635-715</td>
<td>Couret &amp; Venter, Class Frequency Vectors</td>
</tr>
<tr>
<td>7</td>
<td>716-1022</td>
<td>Clark, Reinsurance Pricing</td>
</tr>
<tr>
<td>8</td>
<td>1023-1107</td>
<td>Bernegger, Exposure Curves</td>
</tr>
<tr>
<td>9</td>
<td>1108-1268</td>
<td>Grossi &amp; Kunreuther, Catastrophes</td>
</tr>
<tr>
<td>10</td>
<td>1269-1384</td>
<td>Experience Rating</td>
</tr>
<tr>
<td>11</td>
<td>1385-1461</td>
<td>NCCI Experience Rating Plan</td>
</tr>
<tr>
<td>12</td>
<td>1462-1559</td>
<td>ISO Experience Rating Plan</td>
</tr>
<tr>
<td>13</td>
<td>1560-1656</td>
<td>Frequency and Loss Distributions</td>
</tr>
<tr>
<td>14</td>
<td>1657-1972</td>
<td>Bahnemann, Distributions for Actuaries</td>
</tr>
<tr>
<td>15</td>
<td>1973-2096</td>
<td>Lee Diagrams, Loss Distributions</td>
</tr>
<tr>
<td>16</td>
<td>2097-2171</td>
<td>Retrospective Rating</td>
</tr>
<tr>
<td>17</td>
<td>2172-2249</td>
<td>NCCI Retrospective Rating Plan</td>
</tr>
<tr>
<td>18</td>
<td>2250-2336</td>
<td>Table M Construction</td>
</tr>
<tr>
<td>19</td>
<td>2337-2431</td>
<td>Table L</td>
</tr>
<tr>
<td>20</td>
<td>2432-2529</td>
<td>Lee Diagrams, Retrospective Rating</td>
</tr>
<tr>
<td>21</td>
<td>2530-2555</td>
<td>Limited Table M</td>
</tr>
<tr>
<td>22</td>
<td>2556-2581</td>
<td>Other Loss Sensitive Plans</td>
</tr>
<tr>
<td>23</td>
<td>2582-2697</td>
<td>Pricing Large Dollar Deductible Policies</td>
</tr>
<tr>
<td>24</td>
<td>2698-2712</td>
<td>Concluding Remarks, Individual Risk Rating</td>
</tr>
</tbody>
</table>
### Past Exam Questions by Section

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mahler, Shifting Risk Parameters</td>
<td>10, 31</td>
<td>20</td>
<td>44, 45, 46</td>
</tr>
<tr>
<td>2</td>
<td>Bailey &amp; Simon, Cred. Single Car</td>
<td>6, 30, 32</td>
<td>50</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>Goldburd, Khare and Tevet, GLMs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ASOP 12: Risk Classification</td>
<td></td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>Robertson, Hazard Group Mapping</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Couret &amp; Venter, Class Freq.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Clark, Reinsurance Pricing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Bernegger, Exposure Curves</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Grossi &amp; Kunreuther, Catastrophes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Experience Rating</td>
<td>20, 40, 42</td>
<td>4, 27, 28c&amp;d</td>
<td>31a, 32</td>
</tr>
<tr>
<td>11</td>
<td>NCCI Experience Rating Plan</td>
<td>16, 41</td>
<td>24, 25</td>
<td>10, 34</td>
</tr>
<tr>
<td>12</td>
<td>ISO Experience Rating Plan</td>
<td>17</td>
<td>1, 21, 22, 23</td>
<td>9, 33</td>
</tr>
<tr>
<td>13</td>
<td>Frequency and Loss Distributions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Bahnemann, Distrib. for Actuaries</td>
<td>11, 33, 35</td>
<td>36, 38, 41, 42</td>
<td>13, 36a, 40a</td>
</tr>
<tr>
<td>15</td>
<td>Lee Diagrams, Loss Distributions</td>
<td></td>
<td>39</td>
<td>37</td>
</tr>
<tr>
<td>16</td>
<td>Retrospective Rating</td>
<td>21, 22, 24, 44</td>
<td>29, 31, 32, 34</td>
<td>1, 21, 27</td>
</tr>
<tr>
<td>17</td>
<td>NCCI Retro. Rating Plan</td>
<td>46, 47</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>Table M Construction</td>
<td>45</td>
<td>10</td>
<td>22, 23</td>
</tr>
<tr>
<td>19</td>
<td>Table L</td>
<td>25</td>
<td>30, 35</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Lee Diagrams, Retro. Rating</td>
<td>50</td>
<td></td>
<td>4, 26</td>
</tr>
<tr>
<td>21</td>
<td>Limited Table M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Other Loss Sensitive Plans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pricing LDD Policies</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some questions are based on more than one syllabus reading, particularly on recent exams. In any case, sometimes it is unclear what is the best section in which to put a question. In those cases, I have made one of the possible reasonable choices of where to put a question.

---

3 Integrated questions involve several different syllabus readings.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mahler, Shifting Risk Parameters</td>
<td>48</td>
<td>34</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Bailey &amp; Simon, Cred. Single Car</td>
<td>1</td>
<td>32</td>
<td>2, 22</td>
</tr>
<tr>
<td>3</td>
<td>Goldburd, Khare and Tevet, GLMs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ASOP 12: Risk Classification</td>
<td>2, 43b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Robertson, Hazard Group Mapping</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Couret &amp; Venter, Class Freq.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Clark, Reinsurance Pricing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Bernegger, Exposure Curves</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Grossi &amp; Kunreuther, Catastrophes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Experience Rating</td>
<td>12, 13, 31</td>
<td>1, 4, 40</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>NCCI Experience Rating Plan</td>
<td>28</td>
<td>17, 42</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>ISO Experience Rating Plan</td>
<td>30</td>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>13</td>
<td>Frequency and Loss Distributions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Bahnemann, Distrib. for Actuaries</td>
<td>35, 38, 40, 41</td>
<td>39</td>
<td>11, 35, 37c</td>
</tr>
<tr>
<td>15</td>
<td>Lee Diagrams, Loss Distributions</td>
<td>34, 39</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Retrospective Rating</td>
<td>5, 6, 21</td>
<td>5, 6, 44</td>
<td>8, 10, 31, 32</td>
</tr>
<tr>
<td>17</td>
<td>NCCI Retro. Rating Plan</td>
<td>8, 9, 10, 22, 23, 25</td>
<td>9, 33, 34</td>
<td>35, 40</td>
</tr>
<tr>
<td>18</td>
<td>Table M Construction</td>
<td>19, 48</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>19</td>
<td>Table L</td>
<td>26</td>
<td>45</td>
<td>38, 39</td>
</tr>
<tr>
<td>20</td>
<td>Lee Diagrams, Retro. Rating</td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>21</td>
<td>Limited Table M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Other Loss Sensitive Plans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pricing LDD Policies</td>
<td>42</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>1</td>
<td>Mahler, Shifting Risk Parameters</td>
<td>21</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Bailey &amp; Simon, Cred. Single Car</td>
<td>22</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Goldburd, Khare and Tevet, GLMs</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ASOP 12: Risk Classification</td>
<td></td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Robertson, Hazard Group Mapping</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>Couret &amp; Venter, Class Freq.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Clark, Reinsurance Pricing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Bernegger, Exposure Curves</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Grossi &amp; Kunreuther, Catastrophes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Experience Rating</td>
<td>2, 6, 26, 28</td>
<td>15, 16, 39</td>
<td>26</td>
</tr>
<tr>
<td>11</td>
<td>NCCI Experience Rating Plan</td>
<td>27</td>
<td></td>
<td>24, 27</td>
</tr>
<tr>
<td>12</td>
<td>ISO Experience Rating Plan</td>
<td>3, 4, 5</td>
<td>14, 41</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>Frequency and Loss Distributions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Bahnemann, Distrib. for Actuaries</td>
<td>13, 37, 38, 43</td>
<td>5, 6, 19</td>
<td>25, 26</td>
</tr>
<tr>
<td>15</td>
<td>Lee Diagrams, Loss Distributions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Retrospective Rating</td>
<td>7, 32, 33</td>
<td>47</td>
<td>30, 32</td>
</tr>
<tr>
<td>17</td>
<td>NCCI Retro. Rating Plan</td>
<td>31</td>
<td>18, 20, 45</td>
<td>31</td>
</tr>
<tr>
<td>18</td>
<td>Table M Construction</td>
<td></td>
<td>43</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>Table L</td>
<td>30</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Lee Diagrams, Retro. Rating</td>
<td>8, 9, 29</td>
<td>4, 17</td>
<td>33</td>
</tr>
<tr>
<td>21</td>
<td>Limited Table M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Other Loss Sensitive Plans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pricing LDD Policies</td>
<td>35</td>
<td>46, 48</td>
<td>34, 36</td>
</tr>
<tr>
<td>Sec.</td>
<td>2007 Exam 9</td>
<td>2008 Exam 9</td>
<td>2009 Exam 9</td>
<td>2010 Exam 9</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>1</td>
<td>Mahler, Shifting Risk Parameters</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Bailey &amp; Simon, Cred. Single Car</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Goldburd, Khare and Tevet, GLMs</td>
<td>4a</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>ASOP 12: Risk Classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Robertson, Hazard Group Mapping</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Couret &amp; Venter, Class Freq.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Clark, Reinsurance Pricing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Bernegger, Exposure Curves</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Grossi &amp; Kunreuther, Catastrophes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Experience Rating</td>
<td>26</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>NCCI Experience Rating Plan</td>
<td>25, 28</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>12</td>
<td>ISO Experience Rating Plan</td>
<td>27</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>13</td>
<td>Frequency and Loss Distributions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Bahnemann, Distrib. for Actuaries</td>
<td>7, 8, 10</td>
<td>26, 27</td>
<td>17, 18, 26</td>
</tr>
<tr>
<td>15</td>
<td>Lee Diagrams, Loss Distributions</td>
<td></td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Retrospective Rating</td>
<td>32, 35</td>
<td></td>
<td>28, 31</td>
</tr>
<tr>
<td>17</td>
<td>NCCI Retro. Rating Plan</td>
<td></td>
<td>36</td>
<td>30</td>
</tr>
<tr>
<td>18</td>
<td>Table M Construction</td>
<td>30, 34</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>19</td>
<td>Table L</td>
<td></td>
<td>32, 33</td>
<td>32</td>
</tr>
<tr>
<td>20</td>
<td>Lee Diagrams, Retro. Rating</td>
<td>31</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Limited Table M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Other Loss Sensitive Plans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pricing LDD Policies</td>
<td>33, 36</td>
<td>30, 31</td>
<td>29a</td>
</tr>
<tr>
<td>24</td>
<td>Conclud. Remarks, Indiv. Risk Rat.</td>
<td></td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>Sec.</td>
<td>Title</td>
<td>2011 Exam 8</td>
<td>2012 Exam 8</td>
<td>2013 Exam 8</td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------------------------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>1</td>
<td>Mahler, Shifting Risk Parameters</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Bailey &amp; Simon, Cred. Single Car</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Goldburd, Khare and Tevet, GLMs</td>
<td>3</td>
<td>2, 4</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>ASOP 12: Risk Classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Robertson, Hazard Group Mapping</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Couret &amp; Venter, Class Freq.</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Clark, Reinsurance Pricing</td>
<td>7, 8</td>
<td>7, 10</td>
<td>21, 23, 25</td>
</tr>
<tr>
<td>8</td>
<td>Bernegger, Exposure Curves</td>
<td>9</td>
<td>8</td>
<td>20, 22</td>
</tr>
<tr>
<td>9</td>
<td>Grossi &amp; Kunreuther, Catastrophes</td>
<td>5, 6</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>10</td>
<td>Experience Rating</td>
<td>15, 16b&amp;c</td>
<td>11, 16a&amp;c</td>
<td>9, 10b</td>
</tr>
<tr>
<td>11</td>
<td>NCCI Experience Rating Plan</td>
<td>12</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>ISO Experience Rating Plan</td>
<td>14</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>Frequency and Loss Distributions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Bahnemann, Distrib. for Actuaries</td>
<td>10, 17</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>Lee Diagrams, Loss Distributions</td>
<td>11</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Retrospective Rating</td>
<td>20, 25</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>17</td>
<td>NCCI Retro. Rating Plan</td>
<td>21</td>
<td>19, 23</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Table M Construction</td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>19</td>
<td>Table L</td>
<td></td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>20</td>
<td>Lee Diagrams, Retro. Rating</td>
<td>22</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>Limited Table M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Other Loss Sensitive Plans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pricing LDD Policies</td>
<td>18, 19</td>
<td>20</td>
<td>16, 19</td>
</tr>
<tr>
<td>24</td>
<td>Conclud. Remarks, Indiv. Risk Rat.</td>
<td>23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Added for the 2011 Exam: Bernegger, Robertson, Couret & Venter, Grossi & Kunreuther. Clark Reinsurance Pricing was on Exam 6 prior to 2011.

### Advanced Ratemaking

**2019-CAS8**

<table>
<thead>
<tr>
<th>Sec.</th>
<th>2015 Exam 8</th>
<th>2016 Exam 8</th>
<th>2017 Exam 8</th>
<th>2018 Exam 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mahler, Shifting Risk Parameters</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Bailey &amp; Simon, Cred. Single Car</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Goldburd, Khare and Tevet, GLMs</td>
<td>3</td>
<td>4, 5, 6, 7</td>
<td>4, 5, 6</td>
</tr>
<tr>
<td>4</td>
<td>ASOP 12: Risk Classification</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Robertson, Hazard Group Mapping</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Couret &amp; Venter, Class Freq.</td>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Clark, Reinsurance Pricing</td>
<td>21, 23</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>Bernegger, Exposure Curves</td>
<td>20</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>Grossi &amp; Kunreuther, Catastrophes</td>
<td>22</td>
<td>18, 19</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>Experience Rating</td>
<td>10, 11, 12</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>NCCI Experience Rating Plan</td>
<td></td>
<td>9, 10</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>ISO Experience Rating Plan</td>
<td>9</td>
<td>9, 10</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>Frequency and Loss Distributions</td>
<td></td>
<td></td>
<td>1*</td>
</tr>
<tr>
<td>14</td>
<td>Bahnemann, Distrib. for Actuaries</td>
<td>8a</td>
<td>7, 8, 14</td>
<td>8, 13</td>
</tr>
<tr>
<td>15</td>
<td>Lee Diagrams, Loss Distributions</td>
<td>7</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>16</td>
<td>Retrospective Rating</td>
<td>15, 17</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>17</td>
<td>NCCI Retro. Rating Plan</td>
<td>16</td>
<td>15, 17</td>
<td>14</td>
</tr>
<tr>
<td>18</td>
<td>Table M Construction</td>
<td>12</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>19</td>
<td>Table L</td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Lee Diagrams, Retro. Rating</td>
<td></td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Limited Table M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Other Loss Sensitive Plans</td>
<td></td>
<td></td>
<td>1*</td>
</tr>
<tr>
<td>23</td>
<td>Pricing LDD Policies</td>
<td>13, 14, 18, 19</td>
<td>15, 16</td>
<td>2*, 12</td>
</tr>
</tbody>
</table>

ASOP No. 12 Risk Classification was added to the syllabus for 2017. It replaced American Academy of Actuaries “Risk Classification Statement of Principles.” Some of the past exam questions on Risk Classification no longer apply.

For the 2017 exam, many previous readings were replaced by: a CAS Study Note “Individual Risk Rating,” by Fisher, McTaggart, Petker, and Pettingell, and a CAS Monograph “Loss Distributions for Actuaries,” by Bahnemann.

Integrated questions (which cover material in more than one section) are marked with a star.\(^4\)

\(^4\) The 2017 Exam 8 Sample Integrative Question is in section 23.
Section 3, Generalized Linear Models, Goldburd, Khare, and Tevet

Generalized Linear Models are widely used by actuaries in ratemaking, loss reserving, etc.

**GLMs can be thought of as a generalization of multiple linear regressions.**
However, the distribution of random errors need not be Normal.
Common distributions for the errors are:
Normal, Poisson, Gamma, Binomial, Negative Binomial, Inverse Gaussian, and Tweedie.

Also there is a link function that connects the linear combination of variables and the thing to be modeled.
Common link functions are: identity, inverse, logarithmic, logit, and inverse square.
In a linear model, the link function is equal to the identity function.
In a multiplicative model, the link function is logarithmic; this is analogous to an Exponential regression.

**Generalized Linear Models are fit via maximum likelihood.**

Our goal in modeling is to find the right balance where we pick up as much of the systematic effects (called the signal) as possible and as little of the randomness in the data (called the noise).

Based on the syllabus reading, I do not expect you to be asked to fit a model. Rather you should concentrate on how to set up a GLM, choose between different models, and how to interpret computer output.

Therefore, do not get bogged down in the mathematical details of some of the examples I give, which are provided for those who find that concrete examples help them to learn the material.

This CAS Study Note also discusses some things that apply to most modeling and actuarial work, rather than just to GLMs.

---

1 “Generalized Linear Models for Insurance Rating”, by Mark Goldburd, Anand Khare, and Dan Tevet, CAS monograph series number 5, added to the syllabus for 2016.
Only Sections 1 to 8 are on the syllabus.
Section 9 “Variations on the Generalized Linear Models” is not on the syllabus.
You might also find it useful to glance at “Predictive Models, A Practical Guide for Practitioners and Regulators”, by Don Closter and Caryn Carmean, a short 2019 CAS White Paper, not on the syllabus.
Types of Variables:

Variables can be continuous: size of loss, height, weight, Body Mass Index (BMI), etc.

Variables can be discrete: number of children, number of claims in the last three years, etc.

Variables can be categorical: there are a discrete number of categories. The different possible values that a categorical variable can take on are called its levels.

In the case of nominal variables, the categories do not have a natural order. For example, type of vehicle: sedan, SUV, truck, van.

Sometimes however, the categories have a natural order; such variables are called ordinal. For example injuries may be categorized as: minor, serious, catastrophic, and fatal. This also occurs when a continuous variable is grouped into categories.

Additive and Multiplicative Models:

When one uses the identity function, the model is additive:
\[ \mu = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p. \]

This is analogous to a linear regression.

For example, \( \mu = 100 + 5x_1 - 3x_2. \)
Each increase of 1 in \( x_1 \) results in an increase of 5 in \( \mu \).
Each increase of 1 in \( x_2 \) results in a decrease of 3 in \( \mu \).

When one uses the log link function, the model is multiplicative:
\[ \ln [\mu] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p. \iff \mu = \exp [\beta_0 + \beta_1 x_1 + ... + \beta_p x_p]. \]

This is analogous to an exponential regression.

For example, \( \mu = \exp [5 + 0.2x_1 - 0.1x_2]. \)
Each increase of 1 in \( x_1 \) results in \( \mu \) being multiplied by \( e^{0.2} = 1.221 \).
Each increase of 1 in \( x_2 \) results in \( \mu \) being multiplied by \( e^{-0.1} = 0.905 \).
Advantages of Multiplicative Rating Structures:

1. **A multiplicative plan guarantees positive premium.**
   Having additive terms in a model can result in negative premiums, which doesn’t make sense; you may have to implement clunky patches like minimum premium rules.

2. **A multiplicative model has more intuitive appeal.**
   “It doesn’t make much sense to say that having a violation should increase your auto premium by $500, regardless of whether your base premium is $1,000 or $10,000. Rather it makes more sense to say that the surcharge for having a violation is 10%.”

   “For these and other reasons, **log link models, which produce multiplicative structures, are usually the most natural model for insurance risk.**”

Nevertheless, sometimes a multiplicative model (model using a log link function) does not do a good job of modeling the data, while a different link function does a better job. This is an empirical issue. Most factors in insurance rating algorithms are multiplicative, however it is not uncommon to also have additive elements as well.

Even if one uses a log link function, when interaction terms are included in a model, the structure of the model will no longer have all of the nice features of a multiplicative model. For example: $\mu = \exp[5 + 0.2x_1 - 0.1x_2 + 0.03x_1 x_2]$.

Now the effect of a change in $x_1$ depends on the value of $x_2$, while the effect of a change in $x_2$ depends on the value of $x_1$. For example, in private passenger auto insurance, the effect on expected pure premiums of gender varies by age.

Also keep in mind that for a binary or binomial target variable, for example whether or not a policy is renewed, a logit link function is commonly used as will be discussed.

---

2 See page 5 of Goldburd, Khare, and Tevet.
3 This is an empirical question. For example, a more complicated surcharge such as $100 plus 10\%$ of base premium might be a better prediction of the extra future expected costs.
4 It would be extremely unusual to pay $10,000 or more as a base premium for private passenger automobile. Perhaps they are referring to commercial automobile. In any case, the $10,000 is just for illustrative purposes.
5 Chapter 2 of “Basic Ratemaking” by Werner and Modlin has some examples.
Other Uses of GLMs

While GLMs are commonly used for classification ratemaking, the benefits of GLMs are not restricted to the application of pricing.

The following are a few of the other applications for which insurance companies are using GLMs:

• Practitioners are using GLMs to reduce a variety of risk variables into one score. This has obvious application in regards to creating underwriting tiers, credit scores, fire protection scores, vehicle symbols, etc.

• Many companies have begun to perform elasticity modeling. By building elasticity models for new and renewal business, companies can predict the impact of various actions on market share. A few companies are already linking the profitability and elasticity models to find the optimal pricing decision.

• Claims handlers are starting to see the advantages of GLMs and are using them to help set more accurate reserves and to provide early identification of claims that may be fraudulent or are most likely to end up in a lawsuit.

• Competitive analysis units are using GLMs to reverse-engineer competitors' rates given a large sample of rating quotes.

---

6 Quoted from “GLM Basic Modeling: Avoiding Common Pitfalls,” by Geoff Werner and Serhat Guven, CAS Forum Winter 2007, not on the syllabus.
Common Link Functions:

\[ g(\mu) = \sum \beta_i x_i \iff \mu = g^{-1}(\sum \beta_i x_i). \]

The \( x_i \) are the predictor or explanatory variables.

The \( \beta_i \) are the coefficients, which are to be fit.

\( \beta x = \sum \beta_i x_i \), is the linear predictor.

\( g \) is the link function, whose form needs to be specified.

The link function must satisfy the condition that it be differentiable and monotonic (either strictly increasing or strictly decreasing). Common link functions to use include:

<table>
<thead>
<tr>
<th>Link Function</th>
<th>( g(\mu) )</th>
<th>( g^{-1}(y) )</th>
<th>( \mu = \beta x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td>( g(\mu) = \mu )</td>
<td>( g^{-1}(y) = y )</td>
<td>( \mu = \beta x )</td>
</tr>
<tr>
<td>Log</td>
<td>( g(\mu) = \ln(\mu) )</td>
<td>( g^{-1}(y) = e^y )</td>
<td>( \mu = e^{\beta x} )</td>
</tr>
<tr>
<td>Logit</td>
<td>( g(\mu) = \ln[\mu/(1 - \mu)] )</td>
<td>( g^{-1}(y) = \frac{e^y}{e^y + 1} )</td>
<td>( \mu = \frac{e^{\beta x}}{e^{\beta x} + 1} )</td>
</tr>
<tr>
<td>Reciprocal</td>
<td>( g(\mu) = 1/\mu )</td>
<td>( g^{-1}(y) = 1/y )</td>
<td>( \mu = 1 / (\beta x) )</td>
</tr>
</tbody>
</table>

With more than one variable, the use of the log link function results in a familiar multiplicative model for classification relativities.

One can also use other powers as a link function, such as \( g(\mu) = 1/\mu^2 \) or \( g(\mu) = \sqrt{\mu} \).

Let \( p \) be the probability of policy renewal. Then \( 0 < p < 1 \).
Thus, \( 0 < p / (1 - p) < \infty \).

Applying the logit link function, \(-\infty < \ln[p / (1 - p)] < \infty\).
So we have converted the domain from 0 to 1 to a range of minus infinity to infinity.

The inverse of the logit link function, \( \frac{e^y}{e^y + 1} \), converts the interval from minus infinity to infinity to the interval from zero to one, which would be appropriate for probabilities.\(^7\)

Exercise: \( \mu = \frac{e^{\beta x}}{e^{\beta x} + 1} \). Determine \( \mu \) for \( \beta x = -2 \), \( \beta x = 0 \), and \( \beta x = 2 \).

[Solution: \( e^{-2} / (e^{-2} + 1) = 0.119 \). \( e^0 / (e^0 + 1) = 0.5 \). \( e^2 / (e^2 + 1) = 0.881 \).
Comment: These all make sense as probabilities.]

\(^7\) Note that \( F(x) = e^x / (e^x + 1) \), \(-\infty < x < \infty\) is the logistic function.
Here is a graph of the logit link function, $\ln\left(\frac{x}{1-x}\right)$, for $0 < x < 1$, with range $-\infty$ to $\infty$:

Here is a graph of the inverse of the logit link function, the logistic function $\frac{e^x}{e^x + 1}$, $-\infty < x < \infty$, with range 0 to 1:
It is common to pick the form of the variable $X$, to be a member of an exponential family. *In that case, there are corresponding “canonical link functions”.*

**Canonical Link Functions:**

<table>
<thead>
<tr>
<th>Distribution Form</th>
<th>Canonical Link Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Identity</td>
</tr>
<tr>
<td>Poisson</td>
<td>Log: ln($\mu$)</td>
</tr>
<tr>
<td>Gamma</td>
<td>Reciprocal: $1/\mu$</td>
</tr>
<tr>
<td>Binomial</td>
<td>Logit: ln[$\mu/(1-\mu)$]</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>$1/\mu^2$</td>
</tr>
</tbody>
</table>

Using the canonical link function makes the estimate from the GLM unbiased.

The Normal Distribution is used for example in ordinary linear regression. The Poisson Distribution could be used to model claim frequencies or claim counts. The Gamma Distribution could be used to model claim severities. The Binomial Distribution could be used to model probability of policy renewal.

---

8 While these choices result in some nice mathematica properties, they are not required.
9 If used to model claim severities, one could use the log link function, ln($\mu$).
10 The use of the logit link function with the Binomial or special case Bernoulli is the idea behind logistic regression.
Structure of Generalized Linear Models:

One can state the assumptions of a Generalized Linear Model as:

1. **Random component**: Each component of Y, the target variable is independent and is from one of the exponential family of distributions.$^{11}$

2. **Systematic component**: The p explanatory variables are combined to give the linear predictor X $\beta$.

3. **Link function**: The relationship between the random and systematic components is specified via a link function, g, that is differentiable and monotonic such that:

$$E[Y] = \mu = g^{-1}(X \beta). \iff X \beta = g(\mu).$$

The target variable, also called the dependent variable, Y, is the thing being modeled; it may be: frequency, severity, pure premiums, loss ratios, or something like the probability of policy renewal.

The predictor variables, also called response variables or independent variables, x's, the things being used as inputs to the model, can be things like: age, gender, amount of insurance, etc.

The linear predictor has an intercept $\beta_0$ plus p slopes:

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$  

$$\eta = g(\mu).$$

Several different models may be fit to the same data, with one or more of the above features differing. Then the models would be compared using the output diagnostics, in order to determine the best model to use for the purpose.$^{12}$

---

11 Y is a vector; “each component” of Y refers to the elements of that vector.
12 Similar diagnostics are available as for a multiple linear regression.
A One Dimensional Example of Generalized Linear Models:\textsuperscript{13}

Let us assume a set of three observations: (1, 1), (2, 2), (3, 9).

The predictor variable $x$ takes on the values 1, 2, and 3 for the observations.\textsuperscript{14}

The target variable $Y$ takes on the values 1, 2 and 9 for the observations.

In a generalized linear model, $Y$ will have some distributional form. The mean of the distribution
will vary with $x$. However, any other parameters will be constant.

For now let us assume the identity link function, $g(\mu) = \mu$, so that $\mu = \sum \beta x_i = \beta_0 + \beta_1 x$.\textsuperscript{15}

Thus for now we are fitting a straight line. In general, the identity link function leads to a linear
model.

Assume that $Y$ is Poisson, with mean $\mu$.\textsuperscript{16}

$$
\mu = \beta_0 + \beta_1 x.
$$

For the Poisson Distribution as per Loss Models, $f(y) = e^{-\lambda} \frac{\lambda^y}{y!}$.

$$
\ln f(y) = -\lambda + y \ln(\lambda) - \ln(y!).
$$

The loglikelihood is the sum of the contributions from the three observations:

$$
-(\beta_0 + \beta_1) - (\beta_0 + 2\beta_1) - (\beta_0 + 3\beta_1) + \ln(\beta_0 + \beta_1) + 2\ln(\beta_0 + 2\beta_1) + 9\ln(\beta_0 + 3\beta_1)
$$

$$
- \ln(1) - \ln(2) - \ln(9!).
$$

To maximize the loglikelihood, we set its partial derivatives equal to zero.

Setting the partial derivative with respect to $\beta_0$ equal to zero:

$$
0 = -3 + 1/(\beta_0 + \beta_1) + 2/(\beta_0 + 2\beta_1) + 9/(\beta_0 + 3\beta_1).
$$

Setting the partial derivative with respect to $\beta_1$ equal to zero:

$$
0 = -6 + 1/(\beta_0 + \beta_1) + 4/(\beta_0 + 2\beta_1) + 27/(\beta_0 + 3\beta_1).
$$

Solving these two equations in two unknowns: $\beta_0 = -12/5 = -2.4$ and $\beta_1 = 16/5 = 3.2$.\textsuperscript{17}

$$
\mu = -2.4 + 3.2x. \text{ For } x = 1, \mu = 0.8. \text{ For } x = 2, \mu = 4.0. \text{ For } x = 3, \mu = 7.2.\textsuperscript{18}
$$

\textsuperscript{13} I do not expect you to have to go into this level of detail on your exam.

\textsuperscript{14} It is not clear in this example whether $x$ can take on values other than 1, 2 and 3. These may be the only possible values, or they might be the three values for which we happen to have had an observation. In practical applications, when $x$ is discrete, we would expect to have many observations for each value of $x$.

\textsuperscript{15} I have treated $x_0$ as the constant 1 and $x_1$ as the predictor variable $x$.

\textsuperscript{16} In the case of a Poisson, there are no additional parameters beyond the mean.

\textsuperscript{17} I used a computer to solve these two equations. One can confirm that these values satisfy these equations.

\textsuperscript{18} This differs from what would be obtained if one assumed $Y$ was Normal rather than Poisson.
This model should be interpreted as follows. For a given value of \( x \), \( Y \) is Poisson Distributed with mean \( \mu = -2.4 + 3.2x \). For example, for \( x = 3 \), the mean = 7.2. However, due to random fluctuation, for \( x = 3 \) we will observe values of \( Y \) varying around the expected value of 7.2.\(^{19}\)

If we make a very large number of observations of individuals with \( x = 3 \), then we expect to observe a Poisson Distribution of outcomes with mean 7.2.

As discussed, another important decision is the choice of the link function. In this example, let us maintain the assumption of a Poisson Distribution, but instead of the identity link function let us use the log link function.

\[
\ln(\mu) = \Sigma \beta_i x_i = \beta_0 + \beta_1 x. \quad \Rightarrow \quad \mu = \exp[\Sigma \beta_i x_i] = \exp[\beta_0 + \beta_1 x].
\]

\[
f(y) = e^{-\lambda} \frac{\lambda^y}{y!}.
\]

\[
\ln f(y) = -\lambda + y\ln(\lambda) - \ln(y!) = -\mu + y\ln(\mu) - \ln(y!) = -\exp[\beta_0 + \beta_1 x] + y(\beta_0 + \beta_1 x) - \ln(y!).
\]

The loglikelihood is the sum of the contributions from the three observations:

\[
\begin{align*}
-\exp[\beta_0 + \beta_1] &- \exp[\beta_0 + 2\beta_1] - \exp[\beta_0 + 3\beta_1] + \beta_0 + \beta_1 + 2(\beta_0 + 2\beta_1) + 9(\beta_0 + 3\beta_1) \\
- \ln(1) - \ln(2) - \ln(9!) &
\end{align*}
\]

To maximize the loglikelihood, we set its partial derivatives equal to zero. Setting the partial derivative with respect to \( \beta_0 \) equal to zero:

\[
0 = -\exp[\beta_0 + \beta_1] - \exp[\beta_0 + 2\beta_1] - \exp[\beta_0 + 3\beta_1] + 12.
\]

Setting the partial derivative with respect to \( \beta_1 \) equal to zero:

\[
0 = -\exp[\beta_0 + \beta_1] - 2\exp[\beta_0 + 2\beta_1] - 3\exp[\beta_0 + 3\beta_1] + 32.
\]

Thus we have two equations in two unknowns:

\[
\begin{align*}
\exp[\beta_0 + \beta_1](1 + \exp[\beta_1] + \exp[2\beta_1]) &= 12. \\
\exp[\beta_0 + \beta_1](1 + 2\exp[\beta_1] + 3\exp[2\beta_1]) &= 32.
\end{align*}
\]

Dividing the second equation by the first equation:

\[
\frac{1 + 2\exp[\beta_1] + 3\exp[2\beta_1]}{1 + \exp[\beta_1] + \exp[2\beta_1]} = \frac{32}{12} = \frac{8}{3}.
\]

\[
\Rightarrow \exp[2\beta_1] - 2\exp[\beta_1] - 5 = 0.
\]

\(^{19}\) For the Poisson Distribution, the variance is equal to the mean.
Letting $v = \exp[\beta_1]$, this equation is: $v^2 - 2v - 5 = 0$, with positive solution $v = 1 + \sqrt{6} = 3.4495$. 
$\exp[\beta_1] = 3.4495. \Rightarrow \beta_1 = 1.238.

\Rightarrow \exp[\beta_0] = 12/(\exp[\beta_1] + \exp[2\beta_1] + \exp[3\beta_1]) = 12/(3.4495 + 3.4495^2 + 3.4495^3) = 0.2128.
\Rightarrow \beta_0 = -1.547.

$\mu = \exp[\beta_0 + \beta_1 x] = \exp[\beta_0] \exp[\beta_1]^x = (0.2128)(3.4495^x)$.

For $x = 1$, $\mu = 0.734$. For $x = 2$, $\mu = 2.532$. For $x = 3$, $\mu = 8.735$. 
This differs from the result obtained previously when using the identity link function:

<table>
<thead>
<tr>
<th>$x$</th>
<th>Observed</th>
<th>Poisson, Identity Link</th>
<th>Poisson, Log Link Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.8</td>
<td>0.734</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4.0</td>
<td>2.532</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>7.2</td>
<td>8.735</td>
</tr>
</tbody>
</table>

Here is the same information in the form of a graph, with the data shown as dots:

In general, the choice of a link function makes a difference. 
Using the log link function we got an exponential model rather than a linear model. With more explanatory variables, the log link function gives a multiplicative rather than an additive model.
Exponential Families:

Linear Exponential Families include:
Bernoulli, Binomial (m fixed), Poisson, Geometric, Negative Binomial (r fixed),
Exponential, Gamma (α fixed), Normal (σ fixed), Inverse Gaussian (θ fixed),
and the Tweedie Distribution.

Confusingly, when working on GLMs, “Exponential Family” means “Linear Exponential Family.” This is how the syllabus reading refers to them, and thus from now on I will do the same.

Exponential Families have two parameters, µ the mean, and φ the dispersion parameter. The dispersion parameter is related to the variance. In a GLM φ is fixed across the observations and is treated as a nuisance parameter, in the same way that σ is treated in multiple regression.

It turns out that the relationship between the mean and variance uniquely identifies which linear exponential family we have.

\( \text{Var}[Y] = \phi \, V(\mu) \), where the form of \( V(\mu) \) depends on which exponential family we have.

If the variance does not depend on the mean, then we have a Normal Distribution.
If the variance is proportional to the square of the mean, then we have a Gamma Distribution.
If the variance is proportional to the cube of the mean, then we have an Inverse Gaussian Distribution.
If the variance is proportional to the mean and we have a discrete distribution, then we have a Poisson Distribution.

For the Gamma Distribution, \( f(y) = (y/\theta)^{\alpha} \exp[-y/\theta] / (y \Gamma[\alpha]) \). \( E[Y] = \alpha \theta \). \( \text{Var}[Y] = \alpha \theta^2 \).

If used in a GLM, then we are assuming that we have a Gamma Distribution with \( \alpha \) fixed. Then, \( \text{Variance} = \alpha \theta^2 = (\alpha \theta)^2 / \alpha = (\text{mean})^2 / \alpha \).

Thus for the Gamma Distribution (with \( \alpha \) fixed) the variance is proportional to the square of the mean.

For the Gamma Distribution: \( V(\mu) = \mu^2 \) and \( \phi = 1/\alpha \).

For the following members of the exponential family of distributions, where \( m \) is their mean, their variance is proportional to \( \mu^p \):

- Normal distribution, \( p = 0 \).
- Poisson distribution, \( p = 1 \).
- Gamma distribution, \( p = 2 \).
- Tweedie distribution, \( 1 < p < 2 \).
- Inverse Gaussian distribution, \( p = 3 \).

---

\(^{20}\) Linear Exponential families are defined via the form of their density; however, this definition is not on the syllabus of this exam.
The syllabus reading gives a list of $V(\mu)$ for different exponential families.\textsuperscript{21} The syllabus reading does not go into detail on how to relate the parameterization of exponential families using $\mu$ and $\phi$ to that which you may already be familiar from for example Loss Models. However, in order to make things a little more concrete here is a table.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\mu$</th>
<th>$\phi$</th>
<th>$V(\mu)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>1</td>
</tr>
<tr>
<td>Poisson</td>
<td>$\lambda$</td>
<td>1</td>
<td>$\mu$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$\alpha \theta$</td>
<td>$1/\alpha$</td>
<td>$\mu^2$</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>$\mu$</td>
<td>$1/\theta$</td>
<td>$\mu^3$</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>$\beta/\kappa$</td>
<td>1</td>
<td>$\mu(1 + \kappa \mu)$</td>
</tr>
<tr>
<td>Binomial</td>
<td>$mq$</td>
<td>1</td>
<td>$\mu(1 - \mu/m)$</td>
</tr>
<tr>
<td>Tweedie</td>
<td></td>
<td></td>
<td>$\mu^p$</td>
</tr>
</tbody>
</table>

As discussed subsequently, for the overdispersed Poisson $\phi > 1$.

Gamma as per Loss Models, with mean $= \alpha \theta$ and variance $= \alpha \theta^2$, with $\alpha$ fixed.

Inverse Gaussian as per Loss Models, with mean $= \mu$ and variance $= \mu^3/\theta$, with $\theta$ fixed.

For the Negative Binomial, $\kappa = 1/r$, fixed. $\kappa$ is called the overdispersion parameter. As per Loss Models, the Negative Binomial has mean $= r \beta$ and variance $= r \beta (1+\beta)$.

Binomial, as per Loss Models with $m$ fixed, with mean $= mq$ and variance $= mq(1-q)$. $m = 1$ is a Bernoulli. Goldburd, Khare, and Tevet give $V(\mu)$ for the case where $m = 1$.

The Tweedie Distribution will be discussed subsequently.

\textsuperscript{21} See Table 1 in Goldburd, Khare, and Tevet.
Gamma Distribution:

\[ f(x) = \frac{(x/\theta)^{\alpha}}{\Gamma(\alpha)} \exp[-x/\theta], \quad x > 0. \]

Mean = \( \alpha \theta \).

Variance = \( \alpha \theta^2 \).

\( \phi = 1/\alpha \).

\( V(\mu) = \mu^2 \).

The Gamma Distribution is commonly used to model severity.

Here are graphs of the densities of Gamma Distributions with \( \mu = 100 \) and \( \phi = 1/5 \) or 1/2:

The Gamma Distribution has support from 0 to infinity. The Gamma Distribution is right-skewed (has positive skewness), with a sharp peak and a long tail to the right.

Exercise: Determine the variance for a Gamma Distribution with \( \mu = 20 \) and \( \phi = 1/4 \).

[Solution: Variance = \( \phi \cdot V(\mu) = \phi \cdot \mu^2 = (1/4)(20^2) = 100 \).]

[Comment: The coefficient of variation is: \( \sqrt{100}/20 = 1/2 = \sqrt{\phi}. \)]

---

22 Parameterized as per Loss Models, not on the syllabus of this exam.
I do not expect you to need to know the density.

23 The first has \( \alpha = 5 \) and \( \theta = 20 \), while the second has \( \alpha = 2 \) and \( \theta = 50 \).
Inverse Gaussian Distribution: 

As per Loss Models: 

\[ f(x) = \frac{\theta^2}{2\pi x^{1.5}} \exp\left(-\frac{\theta (\frac{x}{\mu} - 1)^2}{2x}\right), \quad x > 0. \]

Mean = \(\mu\).

Variance = \(\mu^3 / \theta\).

\(\phi = 1/\theta\).

\(V(\mu) = \mu^3\).

The Inverse Gaussian Distribution can be used to model severity. The Inverse Gaussian Distribution is appropriate when the severity has a larger skewness than for a Gamma.

Exercise: Determine the variance for an Inverse Gamma Distribution with \(\mu = 20\) and \(\phi = 1/5\).

[Solution: Variance = \(\phi V(\mu) = \phi \mu^3 = (1/5)(20^3) = 1600\).]

Graphs of the densities of Inverse Gaussian Distributions with \(\mu = 100\) and \(\phi = 0.04\) or 0.01:

---

24 While the Gaussian (normal) describes a Brownian Motion's level at a fixed time, the Inverse Gaussian describes the distribution of the time a Brownian Motion with positive drift takes to reach a fixed positive level.

The cumulant generating function is the natural log of the moment generating function.

The cumulant generating function of an Inverse Gaussian is the inverse function of that of a Gaussian (Normal).

25 I do not expect you to need to know the density.

26 The first has \(\theta = 25\), while the second has \(\theta = 100\).
For the Gamma the variance is proportional to the square of the mean, while for the Inverse Gaussian the variance is proportional to the cube of the mean. The Inverse Gaussian and Gamma are similar, but the Inverse Gaussian has larger skewness and a higher peak.\textsuperscript{27}

For example, a Gamma Distribution with $\mu = 10$ and $\alpha = 2$ has mean $= 10$, and variance $= 10^2/2 = 50$. An Inverse Gaussian Distribution with $\mu = 10$ and $\phi = 20$ has mean $= 10$, and variance $= 10^3/20 = 50$. Thus these two distributions have the same mean and variance.

Here is a graph comparing these two densities:

![Graph Comparing Gamma and Inverse Gaussian Distributions](image)

The Inverse Gaussian Distribution has a higher peak than the Gamma Distribution.

\textsuperscript{27} The skewness for the Gamma distribution is always twice times the coefficient of variation. The skewness for the Inverse Gaussian distribution is always three times the coefficient of variation.
The Inverse Gaussian has more probability in the extreme righthand tail. With the aid of a computer, for this Gamma Distribution the survival function at 40 is $S(40) = 0.30\%$, while for this Inverse Gaussian Distribution, $S(40) = 0.58\%$. 
A Two Dimensional Example of Generalized Linear Models

Let us assume we have two types of drivers, male and female, and two territories, urban and rural. Then there are a total of four combinations of gender and territory. We assume an equal number of claims in each of the four combinations.

Let us assume that we have the following observed severities:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>800</td>
<td>500</td>
</tr>
<tr>
<td>Female</td>
<td>400</td>
<td>200</td>
</tr>
</tbody>
</table>

Let us assume the following generalized linear model:
Gamma Function
Reciprocal link function

Define male and rural as the base level, which introduces a constant term. Then the constant, β₀, applies to all observations.

Let \( X_1 = 1 \) if female and 0 if male.
Let \( X_2 = 1 \) if urban and 0 if rural.

\[
\frac{1}{\mu} = \sum \beta_i x_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 . \Rightarrow \mu = \frac{1}{\beta_0 + \beta_1 x_1 + \beta_2 x_2}.
\]

Therefore, the modeled means are:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>(1/({\beta_0+\beta_2}))</td>
<td>(1/\beta_0)</td>
</tr>
<tr>
<td>Female</td>
<td>(1/({\beta_0+\beta_1+\beta_2}))</td>
<td>(1/({\beta_0+\beta_1}))</td>
</tr>
</tbody>
</table>

For the Gamma Distribution as per Loss Models, \( f(y) = (y/\theta)^\alpha \exp[-y/\theta] / (y \Gamma(\alpha)) \).

\[
\ln f(y) = (\alpha-1)\ln(y) - y/\theta - \alpha\ln(\theta) - \ln[\Gamma(\alpha)] = (\alpha-1)\ln(y) - y/(\mu/\alpha) - \alpha\ln(\mu/\alpha) - \ln[\Gamma(\alpha)]
\]

\[
= (\alpha-1)\ln(y) - \alpha y/\mu - \alpha\ln(\mu) + \alpha\ln(a) - \ln[\Gamma(\alpha)]
\]

\[
= (\alpha-1)\ln(y) - \alpha y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \alpha\ln(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)].
\]

---

28 I do not expect you to have to go into this level of detail on your exam.
29 One could instead use the log link function, and obtain somewhat different results.
30 Since we have taken male as the base level, the covariate has to involve not male.
31 Since we have taken rural as the base level, the covariate has to involve not rural.
The loglikelihood is the sum of the contributions from the four observations:

\((\alpha-1)(\ln(800) + \ln(400) + \ln(500) + \ln(200))\)
\[-\alpha(800(\beta_0 + \beta_2) + 400(\beta_0 + \beta_1 + \beta_2) + 500\beta_0 + 200(\beta_0 + \beta_1))\]
\[+ \alpha(\ln(\beta_0 + \beta_2) + \ln(\beta_0 + \beta_1 + \beta_2) + \ln(\beta_0) + \ln(\beta_0 + \beta_1)) + 4\alpha\ln(\alpha) - 4\ln[\Gamma(\alpha)].\]

To maximize the loglikelihood, we set its partial derivatives equal to zero.

Setting the partial derivative with respect to \(\beta_0\) equal to zero:
\[0 = -\alpha(800 + 400 + 500 + 200) + \alpha\{1/(\beta_0 + \beta_2) + 1/(\beta_0 + \beta_1 + \beta_2) + 1/\beta_0 + 1/(\beta_0 + \beta_1)\}.\]
\[\Rightarrow 1/(\beta_0 + \beta_2) + 1/(\beta_0 + \beta_1 + \beta_2) + 1/\beta_0 + 1/(\beta_0 + \beta_1) = 1900.\]

Setting the partial derivative with respect to \(\beta_1\) equal to zero:
\[0 = -\alpha(400 + 200) + \alpha\{1/(\beta_0 + \beta_1 + \beta_2) + 1/(\beta_0 + \beta_1)\}.\]

Setting the partial derivative with respect to \(\beta_2\) equal to zero:
\[0 = -\alpha(800 + 400) + \alpha\{1/(\beta_0 + \beta_2) + 1/(\beta_0 + \beta_1 + \beta_2)\}.\]

Solving these three equations in three unknowns:\(^{32}\)
\[
\beta_0 = 0.00223811, \beta_1 = 0.00171142, \text{ and } \beta_2 = -0.00106605.
\]
\[
\mu = \frac{1}{0.00223811 + 0.00171142x_1 - 0.00106605x_2}.
\]

For Male and Urban: \(x_1 = 0, x_2 = 1,\) and \(\mu = 1 / (0.00223811 - 0.00106605) = 853.20.\)

For Female and Urban: \(x_1 = 1, x_2 = 1,\)
and \(\mu = 1 / (0.00223811 + 0.00171142 - 0.00106605) = 346.80.\)

For Male and Rural: \(x_1 = 0, x_2 = 0,\) and \(\mu = 1/0.00223811 = 446.81.\)

For Female and Rural: \(x_1 = 1, x_2 = 0,\) and \(\mu = 1/(0.00223811 + 0.00171142) = 253.20.\)

The fitted severities by cell are:\(^{33}\)

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>853.20</td>
<td>446.81</td>
<td>650.01</td>
</tr>
<tr>
<td>Female</td>
<td>346.80</td>
<td>253.20</td>
<td>300.00</td>
</tr>
<tr>
<td>Average</td>
<td>600.00</td>
<td>350.01</td>
<td>475.00</td>
</tr>
</tbody>
</table>

\(^{32}\) I used a computer to solve these three equations.

There is no need to solve for \(\alpha\) in order to calculate the fitted pure premiums by cell.

However, using a computer, the maximum likelihood alpha is 45.6.

\(^{33}\) The averages were computed assuming the same number of claims by cell.
This compares to the observed severities by cell:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>800</td>
<td>500</td>
<td>650</td>
</tr>
<tr>
<td>Female</td>
<td>400</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>Average</td>
<td>600</td>
<td>350</td>
<td>475</td>
</tr>
</tbody>
</table>

Notice how the averages for male, female, urban, and rural are equal for the fitted and observed. The overall experience of each class and territory has been reproduced by the model.

In general, the estimates will be in balance as they were here, when one uses the canonical link function; the canonical link function for the Gamma is the reciprocal link function.\(^\text{34}\)

Exercise: For the Urban territory, what the relativity of male compared to female indicated by the GLM?
[Solution: 853.20/346.80 = 2.460.]

Exercise: For the Rural territory, what the relativity of male compared to female indicated by the GLM?
[Solution: 446.81/253.20 = 1.765.]

The relativities are different in the different territories. In general for a particular GLM, the relativities for one predictor variable can depend on the level(s) of the other predictor variable(s).

Here we have used the reciprocal link function. If instead the log link function had been used, the model would have been multiplicative, and the indicated multiplicative relativities would not have depended on territory. If instead the identity link function had been used, the model would have been additive, and the indicated additive relativities would not have depended on territory.

We could instead change the definitions of the covariates, and have a model without an intercept:
\[x_1 = 1 \text{ if male.}\]
\[x_2 = 1 \text{ if female.}\]
\[x_3 = 1 \text{ if urban and } x_3 = 0 \text{ if rural.}\]

Then \(1/\mu = \sum \beta_i x_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\). \(\Rightarrow \mu = \frac{1}{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}\).

---

\(^{34}\) See “A Systematic Relationship Between Minimum Bias and Generalized Linear Models”, by Stephen J. Mildenhall, PCAS 1999, not on the syllabus.
Therefore, the modeled means are:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>$1/(\beta_1 + \beta_3)$</td>
<td>$1/\beta_1$</td>
</tr>
<tr>
<td>Female</td>
<td>$1/(\beta_2 + \beta_3)$</td>
<td>$1/\beta_2$</td>
</tr>
</tbody>
</table>

For the Gamma Distribution as per Loss Models, $f(y) = (y/\theta)^\alpha \exp[-y/\theta] / (y \Gamma[\alpha])$.

\[
\ln f(y) = (\alpha-1)\ln(y) - y/\mu - \alpha \ln(\mu) + \alpha \ln(\alpha) - \ln[\Gamma[\alpha]].
\]

The loglikelihood is the sum of the contributions from the four observations:

\[
(\alpha-1)(\ln(800) + \ln(400) + \ln(500) + \ln(200)) - \alpha(800(\beta_1 + \beta_3) + 400(\beta_2 + \beta_3) + 500\beta_1 + 200\beta_2) + \alpha(\ln(\beta_1 + \beta_3) + \ln(\beta_2 + \beta_3) + \ln(\beta_1) + \ln(\beta_2)) + 4\alpha \ln(\alpha) - 4 \ln[\Gamma[\alpha]].
\]

To maximize the loglikelihood, we set its partial derivatives equal to zero.

Setting the partial derivative with respect to $\beta_1$ equal to zero:

\[
0 = -\alpha(800 + 500) + \alpha(1/(\beta_1 + \beta_3) + 1/\beta_1). \Rightarrow 1/(\beta_1 + \beta_3) + 1/\beta_1 = 1300.
\]

Setting the partial derivative with respect to $\beta_2$ equal to zero:

\[
0 = -\alpha(400 + 200) + \alpha(1/(\beta_2 + \beta_3) + 1/\beta_2). \Rightarrow 1/(\beta_2 + \beta_3) + 1/\beta_2 = 600.
\]

Setting the partial derivative with respect to $\beta_3$ equal to zero:

\[
0 = -\alpha(800 + 400) + \alpha(1/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3)). \Rightarrow 1/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3) = 1200.
\]

Solving these three equations in three unknowns:\[35\]

$\beta_1 = 0.00223811$, $\beta_2 = 0.00394952$, and $\beta_3 = -0.00106605$.

$\mu = \frac{1}{0.00223811x_1 + 0.00394952x_2 - 0.00106605x_3}$.

For Male and Urban: $x_1 = 1$, $x_2 = 0$, $x_3 = 1$, and $\mu = 1 / (0.00223811 - 0.00106605) = 853.20$.

For Female and Urban: $x_1 = 0$, $x_2 = 1$, $x_3 = 1$, and $\mu = 1 / (0.00394952 - 0.00106605) = 346.80$.

For Male and Rural: $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, and $\mu = 1/0.00223811 = 446.81$.

For Female and Rural: $x_1 = 0$, $x_2 = 1$, $x_3 = 0$, and $\mu = 1/0.00394952 = 253.20$.

The modeled means are the same as in the other version of the model with a base level.

---

[35] I used a computer to solve these three equations.
Instead fit an Inverse Gaussian with the inverse square link function to this same data.\textsuperscript{36}\textsuperscript{37}

For the Inverse Gaussian: \( f(x) = \frac{\theta}{\sqrt{2\pi}} \exp\left[ -\frac{\theta}{2x} \left( \frac{x}{\mu} - 1 \right)^2 \right] \), \( \text{mean} = \mu \), \( \text{variance} = \frac{\mu^3}{\theta} \).

Ignoring terms that do not involve \( \mu \),

\[
\ln f(x) = -\frac{\theta}{2x} \left( \frac{x}{\mu} - 1 \right)^2 = -\frac{\theta}{2x} \left( \frac{x^2}{\mu^2} - 2 \frac{x}{\mu} + 1 \right) = -\frac{\theta x}{2\mu^2} + \frac{\theta}{\mu} - \frac{\theta}{2x}.
\]

Use \( x_1 = 1 \) if male.
\( x_2 = 1 \) if female.
\( x_3 = 1 \) if urban and \( x_3 = 0 \) if rural.

Using the squared reciprocal link function: \( 1/\mu^2 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \).

Thus ignoring terms that do not include \( \mu \), the loglikelihood is:

\[
-\frac{\theta}{2} \left\{ 800(\beta_1 + \beta_3) + 500(\beta_1) + 400(\beta_2 + \beta_3) + 200(\beta_2) \right\} + \theta \left( \sqrt{\beta_1 + \beta_3} + \sqrt{\beta_1} + \sqrt{\beta_2 + \beta_3} \right).
\]

Setting the partial derivative with respect to \( \beta_1 \) equal to zero:

\[
0 = -\frac{\theta}{2} \{800 + 500\} + \frac{\theta}{2} \{1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_1}\}. \Rightarrow 1300 = 1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2}.
\]

Setting the partial derivative with respect to \( \beta_2 \) equal to zero:

\[
0 = -\frac{\theta}{2} \{400 + 200\} + \frac{\theta}{2} \{1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2}\}. \Rightarrow 600 = 1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2}.
\]

Setting the partial derivative with respect to \( \beta_3 \) equal to zero:

\[
0 = -\frac{\theta}{2} \{800 + 400\} + \frac{\theta}{2} \{1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_2 + \beta_3}\}. \Rightarrow 1200 = 1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_2 + \beta_3}.
\]

Solving these three equations in three unknowns:\textsuperscript{38}

\[
\beta_1 = 0.0000054693, \beta_2 = 0.0000134722, \text{and} \beta_3 = -0.00000415544.
\]

\( 1/\mu^2 = 0.0000054693x_1 + 0.0000134722x_2 - 0.00000415544x_3. \)

---

\textsuperscript{36} Again assuming equal claims per cell.
\textsuperscript{37} While the inverse square is the canonical link function for the Inverse Gaussian, one could use a different link function.
\textsuperscript{38} I used a computer to solve these three equations.

There is no need to solve for \( \theta \) in order to calculate the fitted pure premiums by cell.
For Male and Urban: \(x_1 = 1, x_2 = 0, x_3 = 1,\) and
\[\mu = 1 / \sqrt{0.0000054693 - 0.00000415544} = 872.42.\]

For Female and Urban: \(x_1 = 0, x_2 = 1, x_3 = 1,\) and
\[\mu = 1 / \sqrt{0.0000134722 - 0.00000415544} = 327.62.\]

For Male and Rural: \(x_1 = 1, x_2 = 0, x_3 = 0,\) and \(\mu = 1 / \sqrt{0.0000054693} = 427.60.\)

For Female and Rural: \(x_1 = 0, x_2 = 1, x_3 = 0,\) and \(\mu = 1 / \sqrt{0.0000134722} = 272.45.\)

The fitted severities by cell differ from the previous model and are as follows:\(^{39}\)

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>872.42</td>
<td>427.60</td>
<td>650.01</td>
</tr>
<tr>
<td>Female</td>
<td>327.62</td>
<td>272.45</td>
<td>300.04</td>
</tr>
<tr>
<td>Average</td>
<td>600.02</td>
<td>350.03</td>
<td>475.02</td>
</tr>
</tbody>
</table>

This compares to the observed severities by cell:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>800</td>
<td>500</td>
<td>650</td>
</tr>
<tr>
<td>Female</td>
<td>400</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>Average</td>
<td>600</td>
<td>350</td>
<td>475</td>
</tr>
</tbody>
</table>

Notice how subject to rounding, again the averages for male, female, urban, and rural are equal for the fitted and observed. The overall experience of each class and territory has been reproduced by the model.

In general, the estimates will be in balance as they were here, when one uses the canonical link function; the canonical link function for the Inverse Gaussian is the inverse square link function.\(^{40}\)

When the weights differ by cell, this balance involves weighted averages.

\(^{39}\) The averages were computed assuming the same number of claims by cell.

\(^{40}\) See “A Systematic Relationship Between Minimum Bias and Generalized Linear Models,” by Stephen Mildenhall, PCAS 1999, not on the syllabus.
Design Matrix:

As is the case for multiple regression, it is common in GLMs to work with a design matrix.
Each row of the design matrix corresponds to one observation in the data.\footnote{When we have more than one exposure or claim in a cell, a row may correspond to several observations grouped.}
Each column of the design matrix corresponds to a covariate in the model.
If there is an intercept or constant term in the model, then the first column refers to it;
the first column of the design matrix will then consist of all ones.

A one dimensional example, with one covariate plus an intercept, was discussed previously:
Three observations: (1, 1), (2, 2), (3, 9).

\[ Y = \beta_0 + \beta_1 X. \]

Then the design matrix is:

\[
\begin{pmatrix}
1 & 1 \\
1 & 2 \\
1 & 3 \\
\end{pmatrix}
\]

Since the intercept applies to each observation, the first column is all ones.
The second column contains the observed values of the only covariate \( X \).

Note that the design matrix depends on the observations and the definitions of the covariates.
The design matrix does not depend on the link function or the distributional form of the errors.

The response vector would contain the observed values of \( Y \):

\[
\begin{pmatrix}
1 \\
2 \\
9 \\
\end{pmatrix}
\]

The vector of parameters is:

\[
\begin{pmatrix}
\beta_0 \\
\beta_1 \\
\end{pmatrix}
\]

If one used the identity link function, then this model can be written as: \( E[Y] = X \beta \),
where \( X \) is the design matrix and \( \beta \) is the vector of parameters.
If instead one used the log link function, then this model can be written as: \( E[Y] = \exp[X \beta] \).

In general, with a link function \( g \), a GLM can be written as: \( E[Y] = g^{-1}[X \beta] \).

With more covariates, things get a little more complicated. There is not a unique way to define
the covariates. The important thing is to have the design matrix be consistent with the chosen
definitions of the covariates.
A two dimensional model was previously discussed:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>800</td>
<td>500</td>
</tr>
<tr>
<td>Female</td>
<td>400</td>
<td>200</td>
</tr>
</tbody>
</table>

Usually on your exam, one would define a base level, which introduces a constant term. For example, as before we could define male/rural as the base level. Then the constant, $\beta_0$, would apply to all observations. Let $X_1 = 1$ if female and 0 if male. Let $X_2 = 1$ if urban and 0 if rural.

Then with link function $g$, the GLM is: $g(\text{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

If we order the observations as follows, then the design matrix is:

$$
\begin{pmatrix}
\text{Male/Urban} \\
\text{Male/Rural} \\
\text{Female/Urban} \\
\text{Female/Rural}
\end{pmatrix} \leftrightarrow
\begin{pmatrix}
1 & 0 & 1 \\
1 & 0 & 0 \\
1 & 1 & 1 \\
1 & 1 & 0
\end{pmatrix}.
$$

The first column of ones corresponds to the constant term which applies to all observations. The first row of the design matrix corresponds to male/urban: $X_1 = 0, X_2 = 1$. The second row corresponds to male/rural: $X_1 = 0, X_2 = 0$. The third row corresponds to female/urban: $X_1 = 1, X_2 = 1$. The last row corresponds to female/rural: $X_1 = 1, X_2 = 0$.

---

42 One could define any of the four combinations as the base level.
43 Since male is the base level, the covariate has to involve not male.
44 Since rural is the base level, the covariate has to involve not rural.
As stated previously, on your exam the model is likely to be defined with a base level. Nevertheless, one could instead define:

\[ X_1 = 1 \text{ if male. (0 if female)} \]
\[ X_2 = 1 \text{ if female. (0 if male)} \]
\[ X_3 = 1 \text{ if urban and 0 if rural.} \]

Then with link function \( g \), the GLM is:

\[ g(E[Y]) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3. \]

Then if we order the observations as before, then the design matrix is:

\[
\begin{pmatrix}
\text{Male/Urban} \\
\text{Male/Rural} \\
\text{Female/Urban} \\
\text{Female/Rural}
\end{pmatrix}
\leftrightarrow
\begin{pmatrix}
1 & 0 & 1 \\
1 & 0 & 0 \\
0 & 1 & 1 \\
0 & 1 & 0
\end{pmatrix}.
\]

The first row corresponds to male/urban: \( X_1 = 1, X_2 = 0, \) and \( X_3 = 1. \)
The second row corresponds to male/rural: \( X_1 = 1, X_2 = 0, \) and \( X_3 = 0. \)
The third row corresponds to female/urban: \( X_1 = 0, X_2 = 1, \) and \( X_3 = 1. \)
The last row corresponds to female/rural: \( X_1 = 0, X_2 = 1, \) and \( X_3 = 0. \)

The response vector would contain the observed values of \( Y \), in the same order as the rows of the design matrix:

\[
\begin{pmatrix}
\text{Male/Urban} \\
\text{Male/Rural} \\
\text{Female/Urban} \\
\text{Female/Rural}
\end{pmatrix}
\leftrightarrow
\begin{pmatrix}
800 \\
500 \\
400 \\
200
\end{pmatrix}.
\]

The vector of parameters is:

\[
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3
\end{pmatrix}.
\]

---

\(^{45}\) One can put the observations in any order, as long as one is consistent throughout.
This definition of covariates is not unique. For example instead define:

\[ X_1 = 1 \text{ if urban. (0 if rural)} \]
\[ X_2 = 1 \text{ if rural. (0 if urban)} \]
\[ X_3 = 1 \text{ if female and 0 if male.} \]

Exercise: For these definitions, what are the design matrix and the response vector?

[Solution: If we order the observations as before, then the design matrix is:

\[
\begin{pmatrix}
\text{Male/Urban} \\
\text{Male/Rural} \\
\text{Female/Urban} \\
\text{Female/Rural}
\end{pmatrix}
\leftrightarrow
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
1 & 0 & 1 \\
0 & 1 & 1
\end{pmatrix}.
\]

The response vector would contain the observed values of Y, in the same order as the rows of the design matrix:

\[
\begin{pmatrix}
\text{Male/Urban} \\
\text{Male/Rural} \\
\text{Female/Urban} \\
\text{Female/Rural}
\end{pmatrix}
\leftrightarrow
\begin{pmatrix}
800 \\
500 \\
400 \\
200
\end{pmatrix}.
\]

Comment: While the design matrix is different than before, this version of the model is just as valid as the previous ones, as long as everything is handled consistently.

The first row of the design matrix corresponds to male/urban: \( X_1 = 1, X_2 = 0, \) and \( X_3 = 0. \)

The second row corresponds to male/rural: \( X_1 = 0, X_2 = 1, \) and \( X_3 = 0. \)

The third row corresponds to female/urban: \( X_1 = 1, X_2 = 0, \) and \( X_3 = 1. \)

The last row corresponds to female/rural: \( X_1 = 0, X_2 = 1, \) and \( X_3 = 1. \)
Generalized Linear Models, An Example of Adding Dimensions:

Assume we have a one-dimensional model with two territories: Urban and Rural. While there are several different ways to set up this model, let us define: Urban is the base level, $\beta_0$ is the intercept, $X_1 = 1$ if Rural.

Let us now add another dimension, gender: Male or Female.

We can either let Female/Urban be the base level and $X_2 = 1$ if Male, or let Male/Urban be the base level and $X_2 = 1$ if Female. In either case, we add only one more variable to the model we had for one dimension.

We could now add another dimension such as age: Young, Senior, Other. Regardless of which model we had for two dimensions, we would add two more variables to include age. Age has three levels, and in order to add it to our model we need to add $3 - 1 = 2$ variables to the model.

Assume our model for two dimensions had: Female/Rural as the base level, $\beta_0$ is the intercept, with $X_1 = 1$ if Urban, $X_2 = 1$ if Male.

Then for example we could take:
Female/Rural/Other as the base level, $\beta_0$ is the intercept, with $X_3 = 1$ if Young and $X_4 = 1$ if Senior.

If the model has a base level and corresponding constant term, then each categorical variable introduces a number of covariates equal to the number of its levels minus 1.

In this example, the number of covariates is: (constant term) + (2-1) + (2-1) + (3-1) = 5.

In practical applications, it is important to choose the base level of each category to be one with lots of data. If the chosen base level has little data, then the standard errors of the coefficients will be larger than if one had chosen a base level with lots of data.\footnote{See Figure 2 in Goldburd, Khare, and Tevet. Even when the base level has lots of data, the standard errors of coefficients corresponding to levels with little data will be wider than those levels with more data.}
For example, assume instead our model for two dimensions had:
No base level with $X_1 = 1$ if Urban, $X_2 = 1$ if Rural, and $X_3 = 1$ if Male.
Then for example we could take: $X_4 = 1$ if Young and $X_5 = 1$ if Senior.

Without a base level and corresponding constant term, then one and only one of the categorical variables has a number of covariates equal to the number of its levels.
Each of the other categorical variables introduces a number of covariates equal to the number of its levels minus one.

For this example, without a base level, territory has a number of covariates equal to its number of levels, while gender and age each have a number of covariates equal to their number of levels minus one. The total number of covariates is: $2 + (2-1) + (3-1) = 5$, the same as before.

**Design Matrices. Continuous Variables:**

We have looked at discrete categorical variables such as territory. GLMs can also use continuous variables such as amount of insurance and time living at current residence.\(^ {47}\) With continuous variables, determining the design matrix is somewhat different than it is with discrete variables.

Let us assume we are modeling pure premiums for homeowners and observe five policies:

<table>
<thead>
<tr>
<th>Policy</th>
<th>Amount of Insurance ($000)</th>
<th>Time at Residence</th>
<th>Pure Premium ($000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>130</td>
<td>11</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>180</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>250</td>
<td>7</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>16</td>
<td>0</td>
</tr>
</tbody>
</table>

If $X_1 =$ Amount of Insurance and $X_2 =$ Time at Residence,
then the design matrix and response vector are:

$$X = \begin{pmatrix} 100 & 3 \\ 130 & 11 \\ 180 & 0 \\ 250 & 7 \\ 400 & 16 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 \\ 30 \\ 0 \\ 80 \\ 0 \end{pmatrix}.$$

The GLM is: $g(E[Y]) = \beta X$.

\[^ {47}\text{In some cases, the model will perform better if such continuous variables are grouped into several categories.}\]
Poisson Distribution:

\[ f(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \]

Mean = \( \lambda \).
Variance = \( \lambda \).
\( \phi = 1 \).
\( V(\mu) = \mu \).

The Poisson Distribution is commonly used to model frequency.

Overdispersion:

\[ \text{Var}[Y_i] = \phi \ E[Y_i]. \] Since for the Poisson \( \phi = 1 \), the variance is equal to the mean.

When the variance is greater than the mean, one could use a Negative Binomial Distribution, which has a variance greater than its mean.\(^{48}\)

We can instead use an overdispersed Poisson with \( \phi > 1 \).
\[ \text{Var}[Y_i] = \phi \ E[Y_i]. \] For \( \phi > 1 \), variance is greater than the mean.
While this does not correspond to the likelihood of any exponential family, otherwise the GLM mathematics works.\(^{49} \)\(^{50}\)

Using an overdispersed Poisson (ODP), we get the same estimated betas as for the usual Poisson regression.\(^{51}\)
However, the standard errors of all of the estimated parameters are multiplied by \( \sqrt{\hat{\phi}} \).\(^{52}\)

Although not mentioned in the syllabus readings, the usual estimator of the dispersion parameter \( \phi \) is:
\[ \hat{\phi} = \frac{1}{n - p} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\mu_i}. \]

\(^{48}\) One way the variance can be greater than the mean is if frequency is Poisson for each insured, but the means of the Poissons vary between insureds. If the Poisson means follow a Gamma Distribution, then the mixed distribution is a Negative Binomial Distribution.

\(^{49}\) This is called using a quasi-likelihood, although the syllabus reading does not use that term.

\(^{50}\) Often using a Negative Binomial Distribution or an overdispersed Poisson approach to fit a GLM will produce similar results.

\(^{51}\) This is the same reason we can fit the betas in a Normal regression without fitting \( \sigma \).

\(^{52}\) The variance of the estimated parameter is multiplied by \( \phi \).
Negative Binomial Distribution:

\[ f(x) = \frac{\Gamma(x+r)}{x! \Gamma(r)} \frac{\beta^x}{(1+\beta)^{x+r}} = \frac{\Gamma(x+1/\kappa)}{x! \Gamma(1/\kappa)} \frac{(\kappa \mu)^x}{(1+\kappa \mu)^{x+r}}, \quad x = 0, 1, 2, \ldots \]

Mean = \( r \beta = \beta / \kappa \).

Variance = \( r \beta (1+\beta) = (\beta / \kappa) (1+\beta) \).

\( \phi = 1 \).

\( V(\mu) = \mu (1 + \kappa \mu) \).

\( \kappa = 1/r \) is called the overdispersion parameter.

As \( \kappa \) approaches zero while keeping the mean constant, the Negative Binomial Distribution approaches a Poisson Distribution.\(^{53}\)

The Negative Binomial Distribution has its variance greater than its mean.

One way a Negative Binomial Distribution arises is as a Gamma mixture of Poisson Distributions.

The Negative Binomial Distribution is used to model frequency.

Here is a graph comparing the densities of a Poisson with mean 5, and a Negative Binomial with mean 5 and \( \kappa = 1/2 \) (\( r = 2 \)):

---

\(^{53}\) This is mathematically equivalent to letting \( \beta \) approach zero while keeping the mean constant.
One Dimensional Poisson Example with Exposures:

Exposures are a measure of how much insurance protection has been provided. Car years are an example. If one insures three cars each for two years, that is 6 car years of exposure.

Assume the same three observations: (1, 1), (2, 2), (3, 9). However, let us assume 2, 3, and 4 exposures respectively.

Let us again fit a GLM using a Poisson with a log link function.

\[ \lambda_i = \exp[\beta_0 + x \beta_1] \]

We assume that \( Y_i \) is Poisson, with mean \( n_i \lambda_i \), where \( n_i \) is the number of exposures for observation \( i \).

For example, the third observation is Poisson with mean: \( 4 \exp[\beta_0 + 3 \beta_1] \).

For the Poisson Distribution, \( \ln f(y) = -\lambda + y \ln(\lambda) - \ln(y!) \).

Thus the contribution to the loglikelihood from the third observation is:

\[-4 \exp[\beta_0 + 3 \beta_1] + 9 \{\ln4 + (\beta_0 + 2 \beta_1)\} - \ln[9!].\]

The loglikelihood is the sum of the contributions from the three observations:

\[-2 \exp[\beta_0 + \beta_1] - 3 \exp[\beta_0 + 2 \beta_1] - 4 \exp[\beta_0 + 3 \beta_1] + (\beta_0 + \beta_1) + 2(\beta_0 + 2 \beta_1) + 9(\beta_0 + 3 \beta_1) + \ln[2] + 2 \ln[3] + 9 \ln[4] - \ln(1) - \ln(2) - \ln(9!).\]

To maximize the loglikelihood, we set its partial derivatives equal to zero. Setting the partial derivative with respect to \( \beta_0 \) equal to zero:

\[ 0 = -2 \exp[\beta_0 + \beta_1] - 3 \exp[\beta_0 + 2 \beta_1] - 4 \exp[\beta_0 + 3 \beta_1] + 12. \]

Setting the partial derivative with respect to \( \beta_1 \) equal to zero:

\[ 0 = -2 \exp[\beta_0 + \beta_1] - 6 \exp[\beta_0 + 2 \beta_1] - 12 \exp[\beta_0 + 3 \beta_1] + 32. \]

Solving these two equations in two unknowns: \( \beta_0 = -1.97234 \) and \( \beta_1 = 0.91629 \).

\[ \mu_i = n_i \exp[-1.97234 + 0.91629 x_i]. \]

For \( x = 1, \mu = 2 \exp[-1.97234 + 0.91629] = 0.696. \)
For \( x = 2, \mu = 3 \exp[-1.97234 + (2)(0.91629)] = 2.609. \)
For \( x = 3, \mu = 4 \exp[-1.97234 + (3)(0.91629)] = 8.696. \)

\[ ^{54} \text{I used a computer to solve these two equations. One can confirm that these values satisfy these equations.} \]
Offsets, Poisson Model with Log Link Function:

When fitting a Poisson Distribution with a log link function, it is common to state the model with an offset term which is \( \ln[\text{exposures}] \). Offset terms are used to adjust for group size or differing time periods of observation.

With the log link function: \( \lambda_i = \exp[\eta_i] \).

We assume that \( Y_i \) is Poisson, with mean \( n_i \lambda_i \),
where \( n_i \) is the number of exposures for observation \( i \).

\[
\mu_i = n_i \lambda_i = n_i \exp[\eta_i]. \quad \Leftrightarrow \quad \ln[\mu_i] = \ln[n_i] + \eta_i.
\]

Thus we have rewritten the usual equation relating the mean to the linear predictor, \( \eta = X\beta \), with an additional term, \( \ln[n_i] \) which is called the offset. Note that the offset involves a vector of known amounts, the number of exposures corresponding to each observation.

In the previous example: \( \ln[\mu_i] = \ln[n_i] + \beta_0 + \beta_1 x_i \). \( \Leftrightarrow \mu_i = n_i \exp[\beta_0 + \beta_1 x_i] \).

Thus the use of an offset term will produce an equivalent model and the same result as obtained previously.

Computer software to fit GLMs will have an option to include an offset term.
Offsets, When Updating Only Part of the Rating Plan\textsuperscript{55}

“When updating deductible factors, it is frequently desirable to calculate them using traditional loss elimination-based techniques, while the GLM is used for factors other than deductible.”

Assume for example, one is updating other parts of the rating algorithm, but is leaving the deductible credits the same\textsuperscript{56}. The current deductibles and credits are as follows:

<table>
<thead>
<tr>
<th>Deductible</th>
<th>Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$500</td>
<td>Base</td>
</tr>
<tr>
<td>$1000</td>
<td>8% credit</td>
</tr>
<tr>
<td>$2500</td>
<td>14% credit</td>
</tr>
</tbody>
</table>

Then in a GLM for pure premium using a log link function:

\[ \mu = \exp[X\beta] f_D, \]

where \( X\beta \) is the linear predictor (not taking into account deductible), and \( f_D \) is the appropriate deductible factor of: 1, 0.92, or 0.86.

\[ \ln[\mu] = X\beta + \ln[f_D] = X\beta + \text{offset}. \]

This mathematical term is identical to the use of an offset in the case of a Poisson frequency. However, there the offset was \( \ln[\text{exposures}] \) while here the offset is \( \ln[1 \text{- deductible credit}] \).

If an observation is from a policy with a $500 deductible, then the offset is \( \ln[1] = 0 \).
If an observation is from a policy with a $1000 deductible, then the offset is \( \ln[1 - 0.08] = -0.0834 \).
If an observation is from a policy with a $2500 deductible, then the offset is \( \ln[1 - 0.14] = -0.1508 \).

The expected pure premium for a policy with a $2500 deductible is lower than that of a similar policy with a $500 deductible. If the mix of deductibles varies by the other classification variables, then we know that completely ignoring deductibles would lead to distorted estimates of the effects of the other classification variables. The use of the offset term takes into account deductible; however, we are assuming the effects of deductibles are known based on the current credits and that there is no (significant) interaction of effects between deductible amount and other classification variables.

In general, an offset factor is a vector of known amounts which adjusts for known effects not otherwise included in the GLM.

“It is recommended that factors for coverage options—deductible factors, ILFs, peril group factors and the like—be estimated outside the GLM, using traditional actuarial loss elimination techniques. The resulting factors should then be included in the GLM as an offset.”

\textsuperscript{55} See Section 2.6 of Goldburd, Khare, and Tevet.
\textsuperscript{56} Either you will update them at some later date, or the deductible credits will be determined by some technique other than by using a GLM.
As another example, one could take the current territories and territory relativities as givens, and include an offset term in a GLM of ln[territory relativity].

“Territories are not a good fit for the GLM framework. ... Since there are usually many complicated relationships between territory and other variables, your GLM should still consider territory. This is accomplished by including territory in your model as an offset.”

Offsets, General Mathematics:

The offset is added to the linear component: \( g(\mu_i) = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \ldots + \beta_px_{ip} + \text{offset} \).

While I have discussed offsets in the context of the log link function, offsets can be used with other link functions.

For example, an actuary is planning to add a predictor to a model that estimates the probability of a policy having a claim. The actuary has decided to offset all of the current model variables before fitting the new variable. Given the following:

- The current model is a logit link binomial GLM (logistic regression).
- The current fitted probability without the new variable is 5% for an individual.

The logit link function is defined as \( g(\mu) = \ln \left( \frac{\mu}{1 - \mu} \right) \).

Thus the offset for this individual is: \( \ln \left( \frac{5\%}{1 - 5\%} \right) = -2.944 \).

In a similar manner one would calculate an offset for each individual.

Where \( x \) is the new variable, one would now fit a model: \( g(\mu_i) = \beta_0 + \beta_1x_i + \text{offset}_i \).57

Assume the fitted model was: \( \beta_0 = -0.5 \) and \( \beta_1 = 0.1 \).

Assume the same individual as before has a value of the new predictor of \( x = 8 \).

Then: \( \ln \left( \frac{\mu_i}{1 - \mu_i} \right) = -0.5 + (0.1)(8) - 2.944 = -2.644 \).

The new fitted probability of a claim for this individual is: \( \mu_i = \frac{\exp[-2.644]}{1 + \exp[-2.644]} = 6.6\% \).

Exercise: Another individual has a current fitted probability of 8%, and \( x = 2 \). Calculate the revised fitted probability of having a claim for this individual.

\[ \text{Solution: Offset} = \ln \left( \frac{8\%}{1 - 8\%} \right) = -2.442. \quad -0.5 + (0.1)(2) - 2.442 = -2.742. \]

\[ \frac{\exp[-2.742]}{1 + \exp[-2.742]} = 6.1\%. \]

Comment: See 8, 11/18, Q.7.]

57 The betas are not the same as shown previously. We would continue to use the same logit link function.
Prior Weights: 58

When observing numbers of claims, the volume of data is numbers of exposures. When observing sizes of claims, the volume of data is numbers of claims. 59 When a given observation is based on more data we give it more weight.

Let us return to the example with two types of drivers, male and female, and two territories, urban and rural. Before we assumed an equal number of claims in each of the four combinations. Instead let us assume that the Urban/Male combination has twice the volume of the others; in other words Urban/Male has twice as many claims as each of other the other combinations.

Let us assume that we have the same observed average severities:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>800</td>
<td>500</td>
</tr>
<tr>
<td>Female</td>
<td>400</td>
<td>200</td>
</tr>
</tbody>
</table>

Let us again assume the following generalized linear model:
Gamma Function
Reciprocal link function
\[ x_1 = 1 \text{ if male.} \]
\[ x_2 = 1 \text{ if female.} \]
\[ x_3 = 1 \text{ if urban and } x_3 = 0 \text{ if rural.} \]

Then \[ 1/\mu = \sum \beta_i x_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \Rightarrow \mu = \frac{1}{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}. \]

Therefore, the modeled means are:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1/(\beta_1 + \beta_3)</td>
<td>1/\beta_1</td>
</tr>
<tr>
<td>Female</td>
<td>1/(\beta_2 + \beta_3)</td>
<td>1/\beta_2</td>
</tr>
</tbody>
</table>

For the Gamma Distribution as per Loss Models, \[ f(y) = (y/\theta)^a \exp[-y/\theta] / (y \Gamma[\alpha]). \]
\[ \ln f(y) = (\alpha-1)\ln(y) - y/\theta - \alpha \ln(\theta) - \ln[\Gamma[\alpha]] = (\alpha-1)\ln(y) - y/\mu/\alpha - \alpha \ln(\mu/\alpha) - \ln[\Gamma[\alpha]] \]
\[ = (\alpha-1)\ln(y) - \alpha y/\mu - \alpha \ln(\mu) + \alpha \ln(a) - \ln[\Gamma[\alpha]] \]
\[ = (\alpha-1)\ln(y) - \alpha y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \alpha \ln(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \alpha \ln(\alpha) - \ln[\Gamma[\alpha]]. \]

---

58 See Section 2.5 of Goldburd, Khare, and Tevet
59 In Buhlmann Credibility, N is number of exposures when estimating frequency or pure premiums, but N is number of claims when estimating severity.
Since it now has twice the number of claims, we multiply the contribution from Urban/Male by two.

The loglikelihood is the sum of the contributions from the four combinations:

\[
(\alpha - 1) \{2 \ln(800) + \ln(400) + \ln(500) + \ln(200)\} \\
- \alpha \{(2)(800)(\beta_1 + \beta_3) + 400(\beta_2 + \beta_3) + 500\beta_1 + 200\beta_2\} \\
+ \alpha \{2\ln(\beta_1 + \beta_3) + \ln(\beta_2 + \beta_3) + \ln(\beta_1) + \ln(\beta_2)\}\} + 5\alpha \ln(\alpha) - 5 \ln[\Gamma(\alpha)].
\]

To maximize the loglikelihood, we set its partial derivatives equal to zero.

Setting the partial derivative with respect to \(\beta_1\) equal to zero:

\[
0 = -\alpha(1600 + 500) + \alpha\{2/(\beta_1 + \beta_3) + 1/\beta_1\} \Rightarrow 2/(\beta_1 + \beta_3) + 1/\beta_1 = 2100.
\]

Setting the partial derivative with respect to \(\beta_2\) equal to zero:

\[
0 = -\alpha(400 + 200) + \alpha\{1/(\beta_2 + \beta_3) + 1/\beta_2\} \Rightarrow 1/(\beta_2 + \beta_3) + 1/\beta_2 = 600.
\]

Setting the partial derivative with respect to \(\beta_3\) equal to zero:

\[
0 = -\alpha(1600 + 400) + \alpha\{2/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3)\} \Rightarrow 2/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3) = 2000.
\]

Solving these three equations in three unknowns:

\[
\beta_1 = 0.00224451, \beta_2 = 0.00392976, \text{ and } \beta_3 = -0.00103566.
\]

\[
\mu = 1 / (0.00224451x_1 + 0.00392976x_2 - 0.00103566x_3).
\]

For Male and Urban: \(x_1 = 1, x_2 = 0, x_3 = 1,\) and \(\mu = 1 / (0.00224451 - 0.00103566) = 827.23.\)

For Female and Urban: \(x_1 = 0, x_2 = 1, x_3 = 1,\) and \(\mu = 1 / (0.00392976 - 0.00103566) = 345.53.\)

For Male and Rural: \(x_1 = 1, x_2 = 0, x_3 = 0,\) and \(\mu = 1/0.00224451 = 445.53.\)

For Female and Rural: \(x_1 = 0, x_2 = 1, x_3 = 0,\) and \(\mu = 1/0.00392976 = 254.47.\)

The fitted severities by cell are:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>827.23</td>
<td>445.30</td>
</tr>
<tr>
<td>Female</td>
<td>345.53</td>
<td>254.47</td>
</tr>
</tbody>
</table>

Which differ from those obtained previously when we had equal weights.

---

60 I used a computer to solve these three equations. There is no need to solve for \(\alpha\) in order to calculate the fitted pure premiums by cell.
Let us examine what I did in a little more detail.

My contribution to the loglikelihood for Male/Urban was:
\[ 2 \ln f(800) = 2\{(\alpha - 1)\ln(800) - \frac{800}{\theta} - \alpha \ln(\theta) - \ln[\Gamma(\alpha)]\} = 2 \{(\alpha - 1)\ln(800) - \frac{800}{\mu} - \alpha \ln(\mu/\alpha) - \ln[\Gamma(\alpha)]\}. \]

This is the same as assuming two claims each of size 800 were observed.

If instead, we had two claims, one of size 600 and one of size 1000, averaging to the same 800, then the contribution to the loglikelihood for Male/Urban would be:
\[
\ln f(600) + \ln f(1000) = (\alpha - 1)\ln(600) - \frac{600}{\mu} - \alpha \ln(\mu/\alpha) - \ln[\Gamma(\alpha)] + (\alpha - 1)\ln(1000) - \frac{1000}{\mu} - \alpha \ln(\mu/\alpha) - \ln[\Gamma(\alpha)] = (\alpha - 1)\{(\ln(600) + \ln(1000)) - 1600\alpha/\mu - 2\alpha \ln(\mu/\alpha) - 2 \ln[\Gamma(\alpha)]\}.
\]

This differs from before by some constant times \(\alpha - 1\). However, this does not affect the fitted maximum likelihood parameters; when we take a partial derivative with respect to \(\beta_i\) these terms will drop out.

If we only use the fact that Urban/Male has two claims summing to 1600, then we can use the fact that the sum of two identically distributed Gammas has twice the alpha.\(^{61}\) The mean will also be twice as big, so that \(\theta = \mu/\alpha\) would remain the same. Thus the contribution to the loglikelihood for Male/Urban would be the log density of this Gamma with 2\(\alpha\) at 1600:
\[
(2\alpha - 1)\ln(1600) - 1600\alpha/\mu - 2\alpha \ln(\mu/\alpha) - \ln[\Gamma(\alpha)].
\]

Again, this differs from before by terms that involve constants and alpha. However, this does not affect the fitted maximum likelihood parameters; when we take a partial derivative with respect to \(\beta_i\) these terms will drop out.

The members of exponential families each have this nice property that the maximum likelihood fit only depends on the average and not the individual values.\(^{62}\)

In general, when modeling severity, let the weights \(w_i\) be the number of claims.

So for example, if an observation is the average size of 10 claims, then the variance will be \(1/10\) of that for an observation of the size of a single claim.

For example, for the Poisson, \(f(x) = \lambda^x e^{-\lambda} / x!\). \(\ln f(x) = x \ln(\lambda) - \lambda - \ln(x!)/x\).

If we have two (independent) exposures each with mean frequency \(x\), then we can multiply the contribution to the loglikelihood by two: \(2x \ln(\lambda) - 2\lambda - 2 \ln(x!)/x\).

If we have two (independent) exposures each with Poissons with mean \(\lambda\), then the number of claims is Poisson with mean \(2\lambda\).

---

\(^{61}\) The other exponential families share the property that when one adds up independent, identical copies one gets another member of the same family.

\(^{62}\) The mean is a sufficient statistic.
Then with a sum of $2x$, and an average frequency of $x$, the log density is:

$$2x \ln(2\lambda) - 2\lambda - \ln(2x!) = 2x \ln(\lambda) - 2\lambda - 2x \ln(2) - \ln(2x!).$$

Except for constants and terms involving $x$, this is the same loglikelihood as before. Thus we would get the same maximum likelihood fit. Thus when modeling claim frequencies, one can weight by the number of exposures.

**When modeling claim frequency or pure premiums, let the weights be exposures.**

When a weight is specified, the assumed variance for (the mean of) observation $i$ is inversely proportional to the weight:

$$\text{Var}[Y_i] = \frac{\phi \text{Var}[\mu_i]}{\omega_i}.$$  

---

63 This is our usual assumption that the variance of an average is inversely proportional to the number of items being averaged.
A Three Dimensional Example of a GLM:

Here is a three dimensional example for private passenger automobile insurance claim frequency, with: age of driver, territory, and vehicle class. It is a multiplicative model, in other words a GLM with a log link function.

There are 9 levels for driver age, 8 territories, and 5 classes of vehicle. An intercept term is used. Therefore, since each of the three factors is a categorical variable, each has one less parameter than its number of levels. In addition to the intercept term, there are 8 driver age parameters, 7 territory parameters, and 4 vehicle class parameters.

Choose age group 40-49, territory C, and vehicle class A, as the base levels.

Let $b_1$ correspond to the intercept term, and assign the other parameters as follows:

<table>
<thead>
<tr>
<th>Age of driver</th>
<th>Territory</th>
<th>Vehicle class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor level</td>
<td>Parameter</td>
<td>Factor level</td>
</tr>
<tr>
<td>17-21</td>
<td>$\beta_2$</td>
<td>A</td>
</tr>
<tr>
<td>22-24</td>
<td>$\beta_3$</td>
<td>B</td>
</tr>
<tr>
<td>25-29</td>
<td>$\beta_4$</td>
<td>C</td>
</tr>
<tr>
<td>30-34</td>
<td>$\beta_5$</td>
<td>D</td>
</tr>
<tr>
<td>35-39</td>
<td>$\beta_6$</td>
<td>E</td>
</tr>
<tr>
<td>40-49</td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>50-59</td>
<td>$\beta_7$</td>
<td>G</td>
</tr>
<tr>
<td>60-69</td>
<td>$\beta_8$</td>
<td>H</td>
</tr>
<tr>
<td>70+</td>
<td>$\beta_9$</td>
<td></td>
</tr>
</tbody>
</table>

The total number of cells is: $(9)(8)(5) = 360.$

So the design matrix would have 360 rows, assuming that there are no cells lacking data.

---

64 See pages 31 to 32 of “A Practitioner’s Guide to Generalized Linear Models,” by Duncan Anderson; Sholom Feldblum; Claudine Modlin; Doris Schirmacher; Ernesto Schirmacher; and Neeza Thandi, (Third Edition), CAS Study Note, February 2007. Not on the syllabus of this exam.

65 Presumably, there would be another GLM fit to severity.

66 We are not told what distributional form is assumed, but it is probably Poisson.

67 We are not given any details of the fitting or any diagnostics.

68 One could make another set of choices and should get the same fitted frequencies.

69 The standard errors of the fitted parameters are smaller if one chooses as the base level the one with the most exposures.
For example, the first row of the design matrix is probably for age 17-21, Territory A, and Class A, with ones in column 1, 2, and 10:
1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0.

The last row of the design matrix is probably for age 70+, Territory H, and Class E, with ones in column 1, 9, 16, and 20:
1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1.

Exercise: For age 35-39, Territory F, and Class C, what does the corresponding row of the design matrix look like?
[Solution: Ones in columns 1, 6, 14, and 18:
1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0.]

The fitted parameters are an intercept term of 0.1412 and:

<table>
<thead>
<tr>
<th>Age of driver</th>
<th>Territory</th>
<th>Vehicle class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor level</td>
<td>Multiplier</td>
<td>Factor level</td>
</tr>
<tr>
<td>17-21</td>
<td>1.6477</td>
<td>A</td>
</tr>
<tr>
<td>22-24</td>
<td>1.5228</td>
<td>B</td>
</tr>
<tr>
<td>25-29</td>
<td>1.5408</td>
<td>C</td>
</tr>
<tr>
<td>30-34</td>
<td>1.2465</td>
<td>D</td>
</tr>
<tr>
<td>35-39</td>
<td>1.2273</td>
<td>E</td>
</tr>
<tr>
<td>40-49</td>
<td>1.0000</td>
<td>F</td>
</tr>
<tr>
<td>50-59</td>
<td>0.8244</td>
<td>G</td>
</tr>
<tr>
<td>60-69</td>
<td>0.9871</td>
<td>H</td>
</tr>
<tr>
<td>70+</td>
<td>0.9466</td>
<td></td>
</tr>
</tbody>
</table>

The estimated frequency for a 40-49 year old driver, from Territory C and Vehicle Class A, is 0.1412; the estimate for the base levels is the intercept term.

For example, a 22-24 year old driver, from Territory G and Vehicle Class D would have an estimated frequency of: \((1.5228)(1.1451)(0.9764)(0.1412) = 0.2404\).

Exercise: What is the estimated frequency for a 30-34 year old driver, from Territory B and Vehicle Class E?
[Solution: \((1.2465)(0.9567)(1.1002)(0.1412) = 0.1853\).]

69 How you arrange the rows of the design matrix does not affect the result, as long as everything is done consistently.

70 For example, the fitted value of \(\beta_2\) is \(\ln(1.6447)\).

The multipliers for the base levels are one by definition.

71 This is presumably illustrative rather than the output of a GLM fit in a practical application.

72 In order to estimate the overall average frequency, one would need the distribution of exposures by cell.
Another (linear) exponential family is the Tweedie Distribution. The Tweedie Distribution has mean $\mu$ and its variance is proportional to $\mu^p$, for $1 < p < 2$.

The Tweedie Distribution is used to model pure premiums (losses divided by exposures) or loss ratios; there is a point mass of probability at zero corresponding to no loss. The Tweedie Distribution is mathematically a special case of a Compound Poisson Distribution.

When the Tweedie is used in GLMs, $p$ and $\phi$ are constant across all observations.

When using the Tweedie distribution, it turns out that an increase in pure premium is made up of both an increase in frequency and an increase in severity. Even if this assumption does not hold in a given application, the Tweedie GLM can still produce very useful and well fitting models of pure premium.

**Details of the Tweedie Distribution:**

It is a Poisson frequency with a Gamma severity, with parameters of the Poisson and Gamma:

$$\lambda = \frac{\mu^{2-p}}{\phi (2 - p)} , \alpha = \frac{2 - p}{p - 1} , \text{and} \theta = \phi (p-1) \mu^{p-1}.$$

Exercise: Verify the mean and variance of the Tweedie as a Compound Poisson.

**Solution:**

Mean = $\lambda \alpha \theta = \frac{\mu^{2-p}}{\phi (2 - p)} \frac{2 - p}{p - 1} \phi (p-1) \mu^{p-1} = \mu$.

Variance = $\lambda$ (2nd moment of Gamma) = $\lambda \alpha(\alpha+1)\theta^2$

$$\quad = \frac{\mu^{2-p}}{\phi (2 - p)} \frac{2 - p}{p - 1} \frac{1}{p - 1} \{\phi (p-1) \mu^{p-1}\}^2 = \phi \mu^p.$$

Exercise: What is the point mass at zero of the Tweedie as a Compound Poisson.

**Solution:** This corresponds to the Poisson in the Compound Poisson being zero. This has probability $e^{-\lambda}$.

---

73 See Section 2.7.3 of Goldburd, Khare, and Tevet.
74 For insurance modeling, $p$ is typically between 1.5 and 1.8. In some software packages, one can specify the Tweedie distribution, which will in turn cause the software package to find the best value for the power parameter, $p$, when solving for the parameters (betas) in the linear equation.
75 This is far from obvious. Why this is the case is discussed subsequently.
76 For use in a GLM, $\phi$ and $\alpha$ (and thus $p$) are fixed for all observations.
\[ \alpha = \frac{2 - p}{p - 1} \Rightarrow p = \frac{\alpha + 2}{\alpha + 1}. \]

As alpha, the shape parameter of the Gamma, approaches infinity, p approaches 1, and the Tweedie approaches a Poisson. For p near one, the CV of the Gamma is small, and most of the randomness is due to the Poisson frequency.\(^{77}\) As alpha approaches zero, p approaches 2, and the Tweedie approaches a Gamma.

Several of the other exponential family distributions are in fact special cases of Tweedie, dependent on the value of p:
- A Tweedie with p = 0 is a Normal distribution.
- A Tweedie with p = 1 is a Poisson distribution.
- A Tweedie with p = 2 is a Gamma distribution.
- A Tweedie with p = 3 is an inverse Gaussian distribution.

The mean of the Tweedie is: \( \mu = \lambda \alpha \theta \). Also it turns out that: \( \phi = \frac{\lambda^{1-p} (\alpha \theta)^{2-p}}{2 - p} \).

For a Compound Poisson with Gamma severity, we have \( \text{Prob}[X = 0] = e^{-\lambda} \), and for \( x > 0 \):\(^{78}\)

\[
f(x) = \sum_{n=1}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} e^{-x/\theta} \frac{x^n\alpha^{-1}}{\Gamma[n\alpha]} \exp[-x/\theta - \lambda] \sum_{n=1}^{\infty} \frac{\left(\frac{\lambda}{\theta\alpha}\right)^n x^n\alpha^{-1}}{n! \Gamma[n\alpha]}.\]

\[ \theta^\alpha = \phi^{(2-p)/(p-1)} (p-1)^{(2-p)/(p-1)} \mu^{2-p}. \] We had: \( \lambda = \frac{\mu^{2-p}}{\phi (2 - p)} \). Thus \( \lambda/\theta^\alpha \) does not depend on \( \mu \).

Thus the above sum does not depend on \( \mu \).

For a given GLM using the Tweedie, \( \phi \) and \( 1 < p < 2 \) are fixed. \( \Rightarrow \alpha = \frac{2 - p}{p - 1} \) is fixed.

If \( \mu \) increases, then \( \lambda = \frac{\mu^{2-p}}{\phi (2 - p)} \) and \( \theta = \phi (p-1) \mu^{p-1} \) each also increase.

Thus if the mean increases, then both mean frequency = \( \lambda \), and mean severity = \( \alpha \theta \) increase.

\(^{77}\) For the Gamma Distribution, the coefficient of variation is \( 1/\sqrt{\alpha} \).

\(^{78}\) Using the fact that the sum of n independent, identically distributed Gammas is another Gamma Distribution with parameters \( n\alpha \) and \( \theta \).
An Example of a Tweedie Distribution:

Exercise: Take $\mu = 10$, $p = 1.5$, and $\phi = 4$.
Determine the parameters of the Poisson and Gamma.

[Solution: $\lambda = \frac{\mu^{2-p}}{\phi (2 - p)} = \frac{10^{0.5}}{(4)(2 - 1.5)} = 1.581$. $\alpha = \frac{2 - p}{p - 1} = (2 - 1.5) / (1.5 - 1) = 1$. $\theta = \phi (p-1) \mu^{p-1} = (4)(1.5 - 1) 10^{(1.5-1)} = 6.325$.]

Comment: The severity piece of the Compound Poisson is an Exponential with mean 6.325.
The mean of the Compound Poisson is: $(1.581) (6.325) = 10 = \mu$.
The variance of the Compound Poisson is:
(mean of Poisson) (second moment of the Exponential) = $(1.581) \{(2)(6.325^2)\} = 126.5 = (4)(10^{1.5}) = \phi \mu^p$.

The density at zero of the Poisson is: $e^{-1.581} = 20.58\%$.
Thus there is a point mass of probability of 20.58\% at zero.

Using a computer, this Tweedie has density at one of 0.1072.
This Tweedie has density at ten of 0.0258.
This Tweedie has density at twenty five of 0.0024.
Here is graph of the density of this Tweedie Distribution, including the point mass of probability 20.58% at zero.\footnote{For example, at 0.01 the density of the Tweedie is 0.1254.} \footnote{See Figure 5 in Goldburd, Khare, and Tevet.}
Standard Errors and Confidence Intervals for Fitted Parameters:

A standard error is the standard deviation of an estimated coefficient. Computer software for fitting GLMs will output the fitted coefficients and the corresponding standard errors.

For GLMs for large samples, the Maximum Likelihood estimator is approximately multivariate Normal and asymptotically unbiased. Thus in GLM output, it is common to graph the fitted parameters and also bands plus or minus two standard errors.

For example we might have fitted coefficients of:

\[ \hat{\beta}_0 = 223, \hat{\beta}_1 = 1.95, \text{ and } \hat{\beta}_2 = -1.07. \]

With corresponding standard errors of: 30.3, 0.607, and 0.632.

An approximate 95% confidence interval for \( \beta_0 \) is:

\[ 223 \pm (1.960)(30.3) = (164, 282). \]

**95% confidence interval for \( \beta_i \) is: \( \hat{\beta}_i \pm 1.96 \text{ (standard error of } \beta_i \).**

Exercise: Determine an approximate 95% confidence interval for \( \beta_1 \),

[Solution: \( 1.95 \pm (1.960)(0.607) = (0.76, 3.14). \)]

Exercise: Determine an approximate 95% confidence interval for \( \beta_2 \),

[Solution: \( -1.07 \pm (1.960)(0.632) = (-2.31, 0.17). \)]

A standard error of 30.3 for \( \hat{\beta}_0 \) can be thought of as follows: if one simulated similar sized data sets many times and fit GLMs, the estimated intercepts would have a variance of 30.3^2. A smaller standard error gives us more confidence in the estimate of the corresponding coefficient.

Larger data sets will produce smaller standard errors than otherwise smaller data sets; the standard errors go down approximately as the square root of the sample size. The larger the estimated dispersion parameter \( \phi \), the more randomness there is in the data, and thus the larger the standard error; the standard error goes up as \( \sqrt{\phi} \).

---

81 This is similar to the standard error of a regression.
82 Most software will also output the covariance matrix. The variances are along the diagonal. The standard errors are the square roots of the variances.
83 See Figure 2 in Goldburd, Khare, and Tevet.
One can perform hypothesis tests. For example, we can test $\beta_1 = 0$ versus $\beta_1 \neq 0$.

The probability value of this two-sided test is:

$$2 \{1 - \Phi[1.95/0.607]\} = 2 \{1 - \Phi[3.21]\} = 0.1\%.$$  

$p$-value = $\text{Prob}[\text{test statistic takes on a value equal to its calculated value or a value less in agreement with } H_0 \text{ (in the direction of } H_1 \text{) } | \ H_0]$. 

For a $p$-value sufficiently small, we can reject the null hypothesis in favor of the alternative hypothesis that the slope is non-zero. In this case, with a $p$-value of 0.1% we reject the hypothesis that $\beta_1 = 0$.

Exercise: Test $\beta_2 = 0$ versus $\beta_2 \neq 0$.

[Solution: $p$-value = $2 \Phi[-1.07/0.632] = 2 \Phi[-1.69] = 9.1\%$. Therefore, we reject the null hypothesis at 10% and do not reject the null hypothesis at 5%.

Comment: Since zero was not in the 95% confidence interval for $b_2$, we reject the null hypothesis at 5%.

Note that “not reject” is the correct statistical language, although actuaries sometimes say “accept”.

At the 10% significance level we can reject the hypothesis that $\beta_2 = 0$. However, at the 5% significance level there is insufficient evidence to reject the hypothesis that $\beta_2 = 0$.

We can perform two-sided tests: $\beta_2 = 0$ versus $\beta_2 \neq 0$.

We can also perform one-sided tests: $\beta_2 = 0$ versus $\beta_2 > 0$, or $\beta_2 = 0$ versus $\beta_2 < 0$.

---

84 A table of the Normal Distribution will not be attached to your exam.
Using a p-value of 5%:

“A common statistical rule of thumb is to reject the null hypothesis where the p-value is 0.05 or lower. However, while this value may seem small, note that it allows for a 1-in-20 chance of a variable being accepted as significant when it is not. Since in a typical insurance modeling project we are testing many variables, this threshold may be too high to protect against the possibility of spurious effects making it into the model.”

For example, if we are testing the potential usefulness of 60 possible predictor variables, then if we use a p-value of 5%, even if none of the variables actually predict the outcome, on average three of these 60 variables will be selected as significant.

I performed a simulation experiment. I simulated 500 random observations from each of 60 independent normally distributed predictor variables. Then I simulated 500 observations from a normally distributed response variable.

However, the response variable was independent of the predictor variables. In other words, none of the 60 predictor variables was actually useful for predicting the response variable.

Then I fit a multiple regression to this data.

The p-values of the 60 fitted slopes, were from smallest to largest:
0.005, 0.009, 0.020, 0.095, 0.109, 0.121, 0.148, 0.159, 0.177, 0.181, 0.196, 0.206, 0.253, 0.275, 0.331, 0.333, 0.387, 0.421, 0.423, 0.455, 0.494, 0.495, 0.495, 0.513, 0.521, 0.522, 0.545, 0.549, 0.562, 0.591, 0.593, 0.610, 0.614, 0.618, 0.629, 0.637, 0.645, 0.649, 0.653, 0.676, 0.684, 0.707, 0.707, 0.758, 0.778, 0.778, 0.790, 0.806, 0.825, 0.861, 0.886, 0.894, 0.894, 0.916, 0.941, 0.952, 0.980, 0.982, 0.987, 0.993.

We note that even though none of the 60 potential predictor variables is useful, three of the slopes are significant at the 5% level. This illustrates the difficulty of relying on p-values when one starts with a large number of potential predictor variables. In such situations, it is very important to test any selected model on a separate holdout data set, as has been discussed.

---

85 Quoted from Section 2.3.2 of Goldburd, Khare, and Tevet.
86 A Normal Distribution was used for simplicity.
87 This is a special case of a GLM, with a Normal response using the identity link function.
88 While we would expect (5%)(60) = 3 significant slopes, the fact that in this simulation it is exactly three is a coincidence.
89 One can instead use k-fold validation, as discussed previously.
Log Link Function and Continuous Variables:\(^{90}\)

As will be discussed, taking the log of continuous variables provides more variety of behaviors. ⇒ One is more likely to find a behavior that fits your data.

Assume we are using the log link function.
For example: \(\mu = \exp[\beta_0 + \beta_1 x_1 + \beta_2 x_2]\).
Then \(\mu = \exp[\beta_0 + \beta_2 x_2] \exp[\beta_1]^{x_1}\).
Thus the multiplicative relativity for \(x_1\) is \(\exp[\beta_1]^{x_1}\).

Assume \(x_1\) is a continuous variable such as amount of insurance.\(^ {91} \) \(^ {92} \)
For example, if \(\beta_1 = 0.5\), then \(\exp[\beta_1]^{x_1} = 1.649\text{AOI}\).
If instead \(\beta_1 = 1.1\), then \(\exp[\beta_1]^{x_1} = 3.004\text{AOI}\).
Both of these curves have the same form, exponential growth: \(c^x\), where \(c\) is some constant.

What if instead of using \(x_1\) as the predictor variable, we used \(\ln[x_1]\)?
\(\mu = \exp[\beta_0 + \beta_1 \ln[x_1] + \beta_2 x_2] = \exp[\beta_0 + \beta_2 x_2] \times x_1^{\beta_1}\).
Now the multiplicative relativity for amount of insurance is \(\text{AOI}^{\beta_1}\).
For example, if \(\beta_1 = 0.5\), then the multiplicative relativity is \(\text{AOI}^{0.5}\).
If instead \(\beta_1 = 1.3\), then the multiplicative relativity is \(\text{AOI}^{1.3}\).
These are significantly different behaviors.\(^ {93} \)

---

\(^ {90} \) See Section 2.4.1 of Goldburd, Khare, and Tevet.
\(^ {91} \) We have not grouped the variable into levels.
\(^ {92} \) Usually the model will be easier to interpret if for example we used \(\text{AOI} / 100,000\).
While this will be easier to interpret, it produces a mathematically equivalent model to using \(\text{AOI}\).
\(^ {93} \) See Figure 1 in Goldburd, Khare, and Tevet.
This variety of behaviors makes it more likely to find a model that fits the data.

⇒ The authors recommend that when using the log link function in a GLM, you log your continuous predictor variables.\footnote{This allows the scale of the predictors to match the scale of the entity they are linearly predicting, which in the case of a log link is the log of the mean of the outcome.}

This is an empirical question. There will be cases where not taking the log of a continuous predictor variable will result in a GLM that better fits the data; for example, this may be the case when the continuous predictor is year.
Logistic Regression

A variable can be categorical; there are a discrete number of categories, but the labels attached to them may have no significance. Variables can be binary; this is a special case of categorical variable with only two categories, which can be thought of as either 0 or 1. Examples include: whether a policyholder renews its policy, whether a newly opened claim will exceed $10,000, whether a newly opened claim will lead to a subrogation opportunity, whether a newly opened claim is fraudulent, etc.

When the response variable is binary we use the Bernoulli Distribution, the Binomial with m = 1. In that case, the probability of the event is $\mu$ and the probability of not having the event is $1-\mu$. The ratio $\mu/(1-\mu)$ is called the odds.

Exercise: If the probability of an event is 80%, what are the odds?

[Solution: $80\% / (1 - 80\%) = 4$.]

[Comment: The event is 4 times as likely to occur as not occur.]

The most common link function to use in this case is the logit, the log of the odds. $g(m) = \ln[\mu/(1-\mu)]$. $\leftrightarrow \mu = \exp[x'b] / \{1 + \exp[x'b]\}$.

One can group similar observations in which case one has a Binomial with parameters $m_i$ and $q_i$, where $m_i$ is the number of observations in the given group.

A GLM with the Bernoulli or Binomial Distribution using the logit link function is called a Logistic Regression.

Example of Logistic Regression

Fit a logistic regressions to data on whether or not a vehicle had a claim. If $x$ is the vehicle value in units of $10,000$, the model is:

$$\ln[\mu/(1-\mu)] = \beta_0 + \beta_1 x + \beta_2 x^2,$$

with $\hat{\beta}_0 = -2.893, \hat{\beta}_1 = 0.220, \hat{\beta}_2 = -0.026$.

For a vehicle worth $30,000$, $x\beta = -2.893 + (0.220)(3) + (-0.026)(3^2) = -2.467$.

Thus the expected probability of a claim for a vehicle worth $30,000$ is:

$$e^{-2.467} / (1 + e^{-2.467}) = 7.8\%.$$  

Exercise: Determine the expected probability of a claim for a vehicle worth $70,000$.

[Solution: $x\beta = -2.893 + (0.220)(7) + (-0.026)(7^2) = -2.627$. $e^{-2.627} / (1 + e^{-2.627}) = 6.7\%$.]

---

95 See Section 2.8 of Goldburd, Khare, and Tevet.
96 The logit is the canonical link function for the Binomial Distribution, including the special case the Bernoulli.
97 See Section 7.3 of Generalized Linear Models for Insurance Data, by de Jong and Heller, not on the syllabus.
One can also fit a model, using instead a categorical version of vehicle value such as 6 groups: less than 25,000, 25K to 50K, 50K to 75K, 75K to 100K, 100K to 125K, more than 125,000. With the first group as the base level, the fitted model had:

\[ \hat{\beta}_0 = -2.648, \hat{\beta}_1 = 0.174, \hat{\beta}_2 = 0.102, \hat{\beta}_3 = -0.571, \hat{\beta}_4 = -0.397, \hat{\beta}_5 = -0.818. \]

Thus a vehicle of value less than $25,000 has an expected probability of a claim of:

\[ \frac{\exp(-2.648)}{1 + \exp(-2.648)} = 6.61\%. \]

A vehicle of value $25,000 to $50,000 has an expected probability of a claim of:

\[ \frac{\exp(-2.648 + 0.174)}{1 + \exp(-2.648 + 0.174)} = 7.77\%. \]

A vehicle of value greater than $125,000 has an expected probability of a claim of:

\[ \frac{\exp(-2.648 - 0.818)}{1 + \exp(-2.648 - 0.818)} = 3.03\%. \]

The odds for a vehicle of value less than $25,000, the base level is:

\[ 6.61\%/(1 - 6.61\%) = 0.0708 = \exp(-2.648) = \exp[\hat{\beta}_0]. \]

The odds for a vehicle of value 25,000 to $50,000 is:

\[ 7.77\%/(1 - 7.77\%) = 0.0842 = \exp(-2.648)\exp[0.174] = \exp[\hat{\beta}_0]\exp[\hat{\beta}_1]. \]

Thus the odds for the second level are those for the first base level times \( \exp[\hat{\beta}_1] \). The odds for the second level are higher than those for the base level by a factor of \( \exp[0.174] = 1.190 \). The odds for the last level are lower than those for the base level by a factor of \( \exp[-0.818] = 0.441 \).

**Grouping Data:**

When one has binary variables, one can group the data into the possible combinations. For example, with vehicle insurance data using driver’s age (6 groups), area (6 territories), vehicle body (13 types), and vehicle value (6 groups), there are (6)(6)(13)(6) = 2808 cells. Only some of these cells contain data.

For example, assume that driver age group 1, Area A, Hatchback, of value less than $25,000 in value has 554 polices with 47 claims. We would take this as a random draw from a Binomial with \( m = 554 \). In general, for a cell with \( n_i \) policies, we would assume the number of claims follows \( B(n_i, q_i) \).

We get the same fitted parameters and standard errors using either individual or grouped data, although the test statistics will differ.
Correlation Among Predictors:  

When the correlation between two predictor variables is large (in absolute value), the GLM will be unstable. The standard errors of the corresponding coefficients can be large and small changes in the data can produce large changes in the coefficients.

For example, years of education of the father and mother are likely to be highly positively correlated. Including both in a model may produce problems.

Software may not catch the presence of highly correlated variables and try to fit the model anyway. Due to the extreme correlation, the model will be highly unstable; the fitting procedure may fail to converge, and even if the model run is successful the estimated coefficients will be nonsensical.

When you start with a very long list of possible predictors to use in a GLM, it is common for some pairs of predictors to be highly correlated. Thus one should check the correlations of pairs of proposed predictor variables with each other.

If potential problems are found, one can:

1. Remove one or more predictors from the model.
2. Use techniques that combine predictors in order to reduce the dimension, such as Principal Component Analysis and Factor Analysis.

“Determining accurate estimates of relativities in the presence of correlated rating variables is a primary strength of GLMs versus univariate analyses; unlike univariate methods, the GLM will be able to sort out each variable’s unique effect on the outcome, as distinct from the effect of any other variable that may correlate with it, thereby ensuring that no information is double-counted.”

---

98 See Section 2.9 of Goldburd, Khare, and Tevet.
99 One could instead include an average of these two variables.
100 While simple, this may lead us to lose valuable information.
101 You are not responsible for any details.

I discuss Principal Component Analysis in my section on the paper by Robertson. When a set of variables are highly correlated, either positively or negatively, the first principal component or the first two principal components capture most of the variation in the original variables. The first principle component is a linear combination of the original variables.
Multicollinearity:

Multicollinearity is a similar situation which also leads to potential problems. **Multicollinearity occurs when two or more predictors in a model are strongly predictive of another one of the predictor variables.**

As discussed, we are concerned when pairs of variables are highly correlated. However, even in situations where pairs of variables are not highly correlated, problems can occur when looking at three or more predictor variables in combination.

For example, an insurer uses among others the following policyholder characteristics: age, years of education, and income. The first two characteristics would help to predict the final characteristic. Depending on how close this relationship was for this insurer’s data, this could create a problem with the output of a GLM due to multicollinearity.

**A high degree of multicollinearity, usually leads to unreliable estimates of the parameters.** The estimation equations are ill-conditioned.

A useful statistic for detecting multicollinearity is the variance inflation factor (VIF). If one or more of the VIFs is large, that is an indication of multicollinearity. **A common statistical rule of thumb is that a VIF greater than 10 is considered high, indicating possible problems from multicollinearity.**

You will not be asked to compute VIF. Most software packages give VIF as an output.

Aliasing:

Where two predictors are perfectly correlated, they are said to be aliased, and the GLM will not have a unique solution. Equivalently, aliasing can be defined as a linear dependency among the columns of the design matrix X.

Intrinsic aliasing is a linear dependency between covariates due to the definition.

For example, if you have only three territories, then knowing an insured is not in territory one or territory two, implies they are in territory three. Such intrinsic aliasing is common with categorical variables; every insured must be in one and only one of the categories.

---

102 Let $X$ be the design matrix and $X'$ be its transpose. In the case of regression, this is often described as $X'X$ being an ill-conditioned matrix; one can also say the data is ill-conditioned.

In this case, the determinant of $X'X$ will be very small.

103 “The VIF for any predictor is a measure of how much the (squared) standard error for the predictor is increased due to the presence of collinearity with other predictors. It is determined by running a linear model for each of the predictors using all the other predictors as inputs, and measuring the predictive power of those models.”

104 In the case of regression, regress the $i^{th}$ independent variable against all of the other independent variables, and let $R_i^2$ be the coefficient of determination of this regression.

Then the Variance Inflation Factor is: $\text{VIF}_i = 1/(1 - R_i^2)$. 
Initially we have three covariates for the three territories and corresponding coefficients: $\beta_1$, $\beta_2$, and $\beta_3$. Ignoring any other factors, the linear predictor is: $\eta = X_1 \beta_1 + X_2 \beta_2 + X_3 \beta_3$. However, $X_1 + X_2 + X_3 = 1$, so we can eliminate any one of three variables from the model. For example, $\eta = X_1 \beta_1 + X_2 \beta_2 + (1 - X_1 - X_2) \beta_3 = X_1 (\beta_1 - \beta_3) + X_2 (\beta_2 - \beta_3) + \beta_3$. Thus one can eliminate $X_3$ from the model, and include an intercept term if it does not already exist.

The fitted values will be the same regardless of which level is eliminated. Selecting as the base level for each factor the one with the most exposure is helpful, since this minimizes the standard errors associated with other parameter estimates.

Exercise: Age of driver has only three levels: Youth, Adult, and Senior. Demonstrate how aliasing can be used to exclude a level from the age variable.

[Solution: We have that $1 = X_{\text{youth}} + X_{\text{adult}} + X_{\text{senior}}$, and thus $X_{\text{adult}} = 1 - X_{\text{youth}} - X_{\text{senior}}$. Therefore, we can eliminate $\beta_{\text{adult}}$ from the model and include an intercept term if it does not already exist.

Comment: One could have eliminated any of the levels. The adult level, which has the most exposures, would be a good choice for a base level. The intercept term would now correspond to the adult base level; there is no separate parameter for adult. We would still have a parameter for Youth and a parameter for Senior.]

In general, when we have a categorical variable with $N$ levels, the model should have $N-1$ parameters in addition to an intercept term. The chosen base level, which is often the one with the most exposures, is associated with the intercept term and will not have a separate associated parameter.

As another example of intrinsic aliasing, age of vehicle would alias with model year, since if you know one you can determine the other.

Extrinsic aliasing is a linear dependency between covariates that arises due to the particular values in the observed data rather than inherent properties of the covariates themselves.\(^{105}\)

For example, if all sports cars in a data base just happen to be red cars and vice-versa.

Most software will detect aliasing and automatically drop one of those predictors from the model.

---

\(^{105}\) Goldburd, Khare, and Tevet, do not distinguish between intrinsic and extrinsic aliasing.
Limitations of GLMs.\textsuperscript{106}

1. GLMs assign full credibility to the data.\textsuperscript{107}
2. GLMs assume that the randomness of outcomes are uncorrelated.\textsuperscript{108}

As has been discussed on the exam on Basic Ratemaking, when estimating classification relativities by older techniques, an actuary uses credibility. The estimated relativities of classes with less data are given less than full weight.

However, using GLMs the estimated relativities are given full weight.

In fact, for a GLM with just one categorical predictor variable, the estimates will just be the observed average for each level. An actuary would not use the observed average for a small class (or the ratio of its observed average to the observed average for the base level) as a reasonable estimate of the future.

It should be noted, that for a class with little data, the standard errors of the fitted coefficient will be large. Thus we may not reject a value of zero for the coefficient of that small class. In a multiplicative model this would imply a relativity of one. Alternately, we could combine the small class with another class. However, neither of these alternatives is as flexible as giving the observed relativity for this small class some positive weight less than one.

In a regression, we assume that the random components, in other words the errors, $\varepsilon_i$, are uncorrelated.\textsuperscript{109} Similarly, in a GLM we assume that the random components are uncorrelated.\textsuperscript{110} \textsuperscript{111}

This assumption can be violated.

For example, the data set may include several years of data from a single policyholder, which appear as separate records. The outcomes of a single policyholder are correlated. Another example, in the case of wind losses, the outcomes for policyholders in the same area will be correlated.\textsuperscript{112}

If there are large correlations of random components, then the GLM would pick up too much random noise, and produce sub-optimal predictions and overoptimistic measures of statistical significance.

\textsuperscript{106} See Section 2.10 of Goldburd, Khare, and Tevet.
\textsuperscript{107} Section 9 of Goldburd, Khare, and Tevet, not on the syllabus, discusses two ways to incorporate something similar to credibility: generalized linear mixed models and elastic net GLMs.
\textsuperscript{108} Goldburd, Khare, and Tevet, mention two methods that account for correlation in the data: generalized linear mixed model, and generalized estimating equations.
\textsuperscript{109} This assumption is often violated when dealing with time series.
\textsuperscript{110} We assume that the systematic components are correlated.
\textsuperscript{111} For example, drivers in the same class and territory are assumed to have similar expected pure premiums.
\textsuperscript{112} I am thinking about wind losses from other than catastrophes; catastrophes would not be modeled using GLMs.
The authors discuss how actuaries build models; much of the material is not specific to GLMs. They give a list of steps or components:

- Setting of objectives and goals
- Communicating with key stakeholders
- Collecting and processing the necessary data for the analysis
- Conducting exploratory data analysis
- Specifying the form of the predictive model
- Evaluating the model output
- Validating the model
- Translating the model results into a product
- Maintaining the model
- Rebuilding the model

Setting Goals and Objectives:

- Determine the goals.
- Determine appropriate data to collect.
- Determine the time frame.
- What are key risks and how can they be mitigated?
- Who will work on the project; do they have the necessary knowledge and expertise?

Communication with Key Stakeholders:

- Legal and regulatory compliance
- Information Technology (IT) Department
- Underwriters
- Agents

Collecting and Processing Data:

- Time-consuming.
- Data is messy.
- Often an iterative process.
- The data should also be split into at least two subsets, so that the model can be tested on data that was not used to build it.
- Formulate a strategy for validating the model.

Any analysis performed by an actuary is no better than the quality of the data that goes into that analysis!

\[113\] See Section 3 of Goldburd, Khare, and Tevet.
\[114\] As always, such lists are somewhat arbitrary. Many actuaries do not require such lists to do their jobs. Another possible step is to read the literature to see what has been done in similar situations in the past.
\[115\] For more detail, see Section 4 of Goldburd, Khare, and Tevet.
\[116\] Garbage in, garbage out.
Conducting Exploratory Data Analysis (EDA):

Spend some time to better understand the nature of the data and the relationships between the target and explanatory variables. Helpful EDA plots include:

- Plotting each response variable versus the target variable to see what (if any) relationship exists. For continuous variables, such plots may help inform decisions on variable transformations.
- Plotting continuous response variables versus each other, to see the correlation between them.\textsuperscript{117}

Specifying Model Form: \textsuperscript{118}

- What type of predictive model works best?
- What is the target variable, and which response variables should be included?
- Should transformations be applied to the target variable or to any of the response variables?
- Which link function should be used?

Evaluating Model Output: \textsuperscript{119}

- Assessing the overall fit of the model.
- Identifying areas in which the model fit can be improved.
- Analyzing the significance of each predictor variable, and removing or transforming variables accordingly.
- Comparing the lift of a newly constructed model over the existing model or rating structure.

Model Validation: \textsuperscript{120}

- Assessing fit with plots of actual vs. predicted on holdout data.
- Measuring lift.
- For Logistic Regression, use Receiver Operating Characteristic (ROC) Curves.

Translating the Model into a Product:

For GLMs, often the desired result is a rating plan.

- The product should be clear and understandable.
- Are there other rating factors included in the rating plan that were not part of the GLM?
  Then it is important to understand the potential relationship between these additional variable(s) and other variables that were included in the model.
  Judgmental adjustments may be needed.

\textsuperscript{117} Recall that a high correlation, either positive or negative, between pairs of predictor variables may lead to problems with the fitted GLM.

\textsuperscript{118} For more detail, see Section 5 of Goldburd, Khare, and Tevet.

\textsuperscript{119} For more detail, see Section 6 and 7 of Goldburd, Khare, and Tevet.

\textsuperscript{120} For more detail, see Section 7 of Goldburd, Khare, and Tevet.
Maintaining and Rebuilding the Model:

Models should be periodically rebuilt in order to maximize their predictive accuracy, but in the interim it may be beneficial to merely refresh the existing model using newer data. In other words, more frequently one would update the classification relativities without updating the rating algorithm or classification definitions. Less frequently, one would do a more complete update, investigating changing the classification definitions, the predictor variables used, and/or the rating algorithm.

In a somewhat different context, perhaps every 2 years one would update ELPPFs using the latest data but the existing grouping of classifications into hazard groups. Perhaps every 10 or 15 years one would update the grouping of classifications into hazard groups.¹²¹

Data Preparation and Considerations:¹²²

Much of this is not unique to GLMs. Data preparation is time consuming.¹²³ Correcting one data error might help you discover another.

• Combining Policy and Claim Data.
• Modifying the Data.
• Splitting the Data.

Ratemaking Data:

Data is used by actuaries for many purposes including ratemaking. For classification and territory ratemaking, more detailed data on exposures, premiums, losses, and ALAE is used, broken down by class and territory. Ratemaking data is usually aggregated into calendar years, accidents years, and/or policy years.

¹²¹ See the syllabus reading by Robertson.
¹²² See Section 4 of Goldburd, Khare, and Tevet.
¹²³ At a large insurer, much of this work would have been routinely done by someone other than the actuary working on a specific GLM project. The actuary is responsible for determining whether it is reasonable to rely on the data supplied by others. See for example, Actuarial Standard of Practice 23 on Data Quality, not on the syllabus.
Combining Policy and Claim Data:

An insurer’s data is often contained in a policy data base with exposures and premiums, and a separate claims data base with losses and alae. These data bases have to be combined in a manner useful to the actuary.

Issues discussed by the authors:

• Are there timing considerations with respect to the way these databases are updated that might render some of the data unusable?
• Is there a unique key that can be used to match the two databases to each other in such a way that each claim record has exactly one matching policy record?
• What level of detail should the data sets be aggregated to before merging?
• Are there fields that can be safely discarded?
• Are there fields that should be in the database but aren’t?

Finding and Correcting Errors in the Data:

Any dataset of sufficient size is likely to have errors.

• Check for duplicate records.
• Check categorical fields against available documentation.
• Check numerical fields for unreasonable values.
• Decide how to handle each error or missing value that is discovered.

---

124 See for example Chapter 3 of “Basic Ratemaking” by Werner and Modlin, on the syllabus of Exam 5.
125 In which case, the actuary may initiate the process to start collecting this additional information. There are many pieces of information currently collected by insurers and rating bureaus that were not collected 50 years ago.
126 When I worked at a rating bureau, a good percentage of my time was spent on this. We developed many systematic ways to detect errors. More than one group of people would be looking at the data from somewhat different points of view. Large errors were easy to find, but smaller errors required more diligence to find. Unfortunately, one can never find all of the errors.
127 For example, an insurer reported to the rating bureau that an employer had as much payroll as the entire state. This error was quickly spotted and when pointed out to the insurer was quickly corrected.
Splitting the Data into Subsets\footnote{See Section 4.3 of Goldburd, Khare, and Tevet.}

For modeling purposes one should split the data into either two or three parts. This can be done either at random or based on time, for example policy year.

The simpler approach is to split the data into a training set and test (holdout) set\footnote{This is done in the syllabus reading by Couret and Venter.}. For example, the training set could be 2/3 of the data while the test set is the remaining 1/3.

One develops the model on the training set. Then once one has come up with a final model or a few candidates for a final model, one would test performance on the test set of data, which was not used in developing the model\footnote{Such testing will be discussed subsequently.}.

The model was developed to fit well to the training set. In doing so, we are concerned that the model may be picking up peculiarities of the training set. If the model does a good job of predicting for the test set, which was not used in developing the model, then it is likely to also work well at predicting the future\footnote{We are interested in how the GLM will perform at predicting the response variable on some future set of data rather than on the set of past data with which we are currently working.}.

Reasons to split the data into a training set and a test set:

\begin{itemize}
\item Attempting to test the performance of any model on the same set of data on which the model was built will produce overoptimistic results. The model-fitting process optimizes the parameters to best fit the data used to train it. Using the training data to compare our model to a model built on different data would give our model an unfair advantage.

\item As we increase the complexity of the model, the fit to the training data will always get better. Thus the performance on the training data can not be used to compare models of different complexity. On the other hand, for data the model fitting process has not seen, eventually increased complexity will worsen the performance of the model\footnote{See Figure 7 in Goldburd, Khare, and Tevet.}.

\end{itemize}

Thus the performance on the test data can be used to compare models of different complexity.

The split of data can be performed either by randomly allocating records between the training and test sets, or by splitting on the basis of a time variable\footnote{One could split by month or by calendar/accident year.}. The latter approach has the advantage in that the model validation is performed “out of time” as well as out of sample, giving us a more accurate view into how the model will perform on unseen years.

\begin{itemize}
\item As in Couret and Venter one could select either the even or odd years of data as the training set and the other as the test set, in order to be neutral with respect to trend and maturity.
\end{itemize}
“Out-of-time validation is especially important when modeling perils driven by common events that affect multiple policyholders at once. An example of this is the wind peril, for which a single storm will cause many incurred losses in the same area. If random sampling is used for the split, losses related to the same event will be present in both sets of data, and so the test set will not be true unseen data, since the model has already seen those events in the training set. This will result in overoptimistic validation results. Choosing a test set that covers different time periods than the training set will minimize such overlap and allow for better measures of how the model will perform on the completely unknown future.”

The actuary should wait as long as possible in the process to use the test set. Once you start comparing to the test set, if you go back and change the form of the model, the usefulness of the test set for further comparisons has been diminished.

Thus sometimes, one uses the more complicated approach of splitting the data in three subsets: a **training set, validation set, and test (holdout) set**. For example, the split might be 40%, 30%, 30%.

As before, one develops the model on the training set. Then once one has come up with a good model or several good models, one would test performance on the validation set of data, which was not used in developing the model(s). If any changes in the form of the model are indicated, one goes back and works again with the training set. This iteration continues until the actuary is satisfied.

Then one would test performance on the test set of data, which was not used so far.

In either the simpler or more complicated case, **once a final form of the model has been decided upon, one should go back and use all of the available data to fit the parameters of the GLM.**

---

134 Quoting from Section 4.3.1 of Goldburd, Khare, and Tevet. See 8 11/17, Q.4c.

135 Hopefully the total amount of data available is big enough to allow this.
Underfitting and Overfitting:

A model may be either overfit or underfit. Think of fitting a polynomial to 20 points. A straight line with no intercept, in other words a model with one parameter, will probably not do a good job of fitting the points. A fitted 19th degree polynomial, in other words a model with 20 parameters, will pass through all of the points.

However, actuaries are using a model to predict the behavior in the future. The one parameter model will probably not do a very good job, since it ignored some of the information in the data. It is underfit. The 20 parameter model will not do a good job of predicting, since it picked up all of the random fluctuation (noise) in the data. It is overfit.

A model should be made as simple as possible, but not simpler.

Underfit. ⇔ Too few Parameters. ⇔ Does not use enough of the useful information.
⇔ Does not capture enough of the signal.

Overfit. ⇔ Too many Parameters. ⇔ Reflects too much of the noise.

We wish to avoid both underfitting and overfitting a model.

Think of fitting loss distributions. We would not use the most complicated model possible.\textsuperscript{136} We would only add parameters to the extent they were statistically significant.\textsuperscript{137} In a particular situation, it might be that an Exponential Distribution (one parameter) is an underfit model, a Transformed Gamma Distribution (3 parameters) is an overfit model, while a Gamma Distribution (2 parameters) is just right.

\textsuperscript{136} Recall that a mixture of two or more distributions can have a lot of parameters.

\textsuperscript{137} Think of the Likelihood Ratio Test, AIC, or BIC (the Schwarz Bayesian Criterion).
In order to produce a sensible model that explains recent historical experience and is likely to be predictive of future experience, one needs to avoid both too little and too much complexity. Each added parameter adds a degree of freedom to the model. This can be due to the addition of a new predictor variable, the addition of a polynomial term, the addition of an interaction term, etc. Each added degree of freedom makes the model more complex.

**Our goal in modeling is to find the right balance where we pick up as much of the signal as possible with minimal noise.** This is illustrated in Figure 7 of the syllabus reading:

As we add more parameters, we get a model that fits the training set better. However, when we compare such a model fitted to the training data to the test data, there is a point past which added parameters reduce the fit to the test data. The right balance is indicated by the vertical dotted line, at about 70 degrees of freedom in this case.  

---

138 Taken from “GLM II, Basic Modeling Strategy”, presented by Lenard Shuichi Llaguno, FCAS, at the 2012 CAS Ratemaking and Product Management Seminar.  
139 Here the authors use degrees of freedom to refer to the number of parameters in the fitted model. In for example the F-test, many authors instead define the degrees of freedom as number of observations minus number of fitted parameters for the fitted model.
Cross Validation

Cross Validation is another technique for data splitting, although it is often of limited usefulness for actuarial work.

Split the data into for example 10 groups. Each group is called a fold. For each fold:
• Train the model using the other folds.
• Test the model using the given fold.

Cross validation has the advantage of using all of the data (at some point) to estimate the mean squared error, rather than only using the portion of the data in the holdout set to do so. Thus cross validation should produce a better estimate of the MSE.

In the case of 10-fold cross validation, fit model form A on the data for the first 9 folds. Then compute the mean squared error (MSE) of this fitted model used to make predictions to the data in the remaining tenth fold.

Now fit the same model form A on the data for the folds other than the ninth. Then compute the mean squared error (MSE) of this fitted model used to make predictions to the data in the remaining ninth fold.

We would continue in this manner and then average these ten mean squared errors. This would be the estimated test MSE for model form A.

We could then determine the MSE of several other model forms, B, C, etc., in a similar manner. The form of model with the lowest test MSE would be best.

For example, we might compare polynomial models with different number of powers of a predictor variable.

---

140 See Section 4.3.4 of Goldburd, Khare, and Tevet.
141 See also, An Introduction to Statistical Learning with Applications in R, by James, Witten, Hastie, and Tibshirani, not on the syllabus of this exam. They also discuss how to apply cross validation to other modeling techniques such as ridge regression and the lasso.
142 According to the authors, this training procedure has to include all of the steps of the model building, including the variable selection and transformation; these steps usually include significant amounts of actuarial judgement.
143 For example, Model A and Model B might use different sets of predictor variables.
One has fit similar GLMs on a set of data, where one of the predictors enters using polynomials of different degrees. The test MSEs were estimated using ten-fold cross-validation:

![Graph showing mean squared error vs. degree]

The model using the third degree model seems to perform best.\(^{144}\)

Limitations of Cross-Validation for Actuarial Work:

Cross-validation can be useful for deciding how many polynomial terms to include.\(^{145}\) However, cross validation is often of limited usefulness for most insurance modeling applications.

The actuary usually applies a great deal of care and judgment in selecting the variables to be included in the model. If using cross validation, this actuarial judgement should be applied separately to each of the data sets created by leaving out one fold. This is not really practical. Thus, using cross validation in place of a holdout set is only really appropriate where a purely automated variable selection process is used.\(^{146}\)

For most actuarial modeling, the use of a holdout set is preferred to the use of cross validation. The final model valuation should always be done using a distinct set of data held out until the end.

\(^{144}\) Due to the data being assigned to the 10 folds at random, if one performed cross validation again, one would get somewhat different estimates of the test MSEs. Therefore, in practical applications one would perform cross validation several times and compare the results.

\(^{145}\) This is an example of evaluating a “tuning parameter” of the model.

\(^{146}\) This is the opinion of the authors of the syllabus reading, who have plenty of experience using GLMs for actuarial modeling.
An Example of k-Fold Cross-Validation

Eight observations of three independent variables and one dependent variable:

<table>
<thead>
<tr>
<th>X₁</th>
<th>X₂</th>
<th>X₃</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>1</td>
<td>-4</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>-4</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>16</td>
<td>118</td>
</tr>
<tr>
<td>17</td>
<td>-8</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>-6</td>
<td>24</td>
<td>61</td>
</tr>
</tbody>
</table>

I will perform 4-fold cross-validation, so that each fold contains 8/4 = 2 observations. We need to divide the original data into 4 random subsets; the estimated test MSE will depend to some extent on this random subdivision. My four folds will be: (1, 7), (2, 4), (3, 5), (6, 8).

If we leave out the first and seventh observations, and fit a regression model to the remaining six observations, the fitted parameters are:

\[ \hat{\beta}_0 = 3.78881, \hat{\beta}_1 = 5.10444, \hat{\beta}_2 = 5.17811, \hat{\beta}_3 = -0.621247. \]

We now plug into this fitted model the values of the predictors for the first observation:

\[ \hat{Y} = (5.10444)(-2) + (5.17811)(1) + (-0.621247)(-4) = 1.24303. \]

We now plug into this fitted model the values of the predictors for the seventh observation:

\[ \hat{Y} = 3.78881 + (5.10444)(17) + (5.17811)(-8) + (-0.621247)(20) = 36.7145. \]

The mean squared difference between the observed values and these predicted values is:

\[ \text{MSE}_1 = \frac{(6 - 1.24303)^2 + (36.7145 - 2)^2}{2} = 613.863. \]

Similarly, we would now instead leave out the 2nd and 4th observations. We continue in this manner, and the four mean squared errors are: 613.863, 231.863, 697.906, 1458.9.

The average of these four values is the 4-fold cross-validation estimate of the test MSE: 750.633.

I used R to perform this same process five separate times and the estimated test MSEs were: 449.2197, 1249.365, 616.1268, 680.8828, 754.928.

With only 8 observations, we see considerable variation in these estimates.

\[ \text{Solely in order to give a simple concrete example; you are not responsible for any details.} \]

\[ \text{Using the R function cv.glm. Each time a different set of random folds is used.} \]
Selection of Model Form:¹⁴⁹

“Selecting the form of a predictive model is an iterative process, and is often more of an art than a science.”

Important decisions on the form of a GLM include:
• Choosing the target variable.
• Choosing a distribution for the target variable.
• Choosing the predictor variables.
• Whether to apply transformations to the predictor variables.
• Grouping categorical variables.
• Whether to include interactions.

¹⁴⁹ See Section 5 of Goldburd, Khare, and Tevet.
An actuary could build two separate models: one for frequency and one for severity. Alternately the actuary could build a single model for pure premium. If there is time, an actuary could do both of these different approaches and compare the results.

Advantages of the frequency/severity approach over pure premium modeling:
• Provides the actuary with more insight.
• Each of frequency and severity is more stable than pure premium.

Disadvantages of pure premium modeling versus the frequency/severity approach:
• Some interesting effects may go unnoticed.
• Pure premium modeling can lead to underfitting or overfitting.
• The Tweedie distribution used to model pure premium contains the implicit assumption that an increase in pure premiums is made up of an increase in both frequency and severity.

For example, urban driving tends to lead to a higher frequency of accidents (per mile driven) than rural driving. However, urban driving tends to lead to a lower severity of accidents than rural driving. These two separate effects could be masked in a pure premium model. In any case, with just a pure premium model, the actuary would not get this interesting and perhaps important insight.

While territory would show up as significant in a frequency model, when testing it in a pure premium model the high variance in severity may overwhelm this effect, rendering the territory statistically insignificant. Thus, a useful predictive variable will be excluded from the model, leading to underfitting.

Assume that a predictor variable has a significant effect on frequency and no effect on severity. If that variable is included in a pure premium model, then the fitted GLM will pick up any effect of severity in the training data even if it is just noise. The corresponding parameter will be overfit.

For frequency and severity, a priori expected patterns help the actuary to produce a better model. To the extent that the historical pattern is erratic, the actuary will be able to use appropriate techniques and knowledge about insurance to build a model that captures the signal in the data.

---

150 See Section 5.1.1 of Goldburd, Khare, and Tevet.
151 If the log link function is used for both, then the pure premium (multiplicative) relativities will be the product of the separate frequency and severity relativities.
152 Recall that the standard for full credibility for pure premium is the sum of those for frequency and severity.
153 The authors assume that one would use the Tweedie Distribution to model pure premiums.
154 While this could happen in general, in the example I have chosen it is unlikely to do so.
For example, when modeling auto collision frequency, the actuary may expect the frequency by age to decrease from youthful to adult and increase again for the most mature drivers. The following figure compares the historical frequencies (triangles) and modeled frequencies (squares) by age.

The modeled frequencies follow the general pattern expected.

---

155 These are frequencies per car year. Most senior citizens have higher expected frequencies per mile driven. However, their average number of miles driven per year is lower.

156 Figure 5, from “GLM Basic Modeling: Avoiding Common Pitfalls,” by Geoff Werner and Serhat Guven, CAS Forum Winter 2007, not on the syllabus.
Modeling Loss Ratios:

If the goal of the project is to identify deficiencies in the existing rating plan, loss ratio may be an appropriate target variable for the GLM.\textsuperscript{157} However, there are disadvantages to modeling loss ratios rather than pure premiums or frequency/severity.

Theoretical and practical disadvantages to loss ratio modeling:\textsuperscript{158}

\begin{itemize}
  \item One needs to put premiums on-level at a granular level; difficult and time consuming. One has to put on the current rate level individual policies; overall on-level factors will not do.
  \item There is no generally accepted error distribution.\textsuperscript{159}
  \item Difficult to distinguish noise from pattern, compared to modeling frequency/severity.
  \item If changes are made to the rates, then models cannot be reused from the last review.
\end{itemize}

\textsuperscript{157} See Section 5.1 of Goldburd, Khare, and Tevet.

\textsuperscript{158} Taken from “GLM II: Basic Modeling Strategy,” by Claudine Modlin, CAS Predictive Modeling Seminar, October 2008.

\textsuperscript{159} However, as discussed previously, one could use the Tweedie Distribution.
Policies with Multiple Coverages and Perils:\textsuperscript{160}

A Businessowners package policy includes building, business personal property, and liability coverage.\textsuperscript{161} Each of those coverages should be modeled separately.

In addition, one may models each peril individually.\textsuperscript{162} For the Businessowners building model, one may wish to create separate models for: fire and lightning, wind and hail, and all other perils.\textsuperscript{163}

One way to combine separate models by peril in order to get a model for all perils:
1. Use the separate models by peril to generate predictions of expected loss due to each peril for some set of exposure data.\textsuperscript{164}
2. Add the peril predictions together to form a combined loss cost for each record.
3. Run a model on that data, using the combined loss cost calculated in Step 2 as the target, and the union of all the individual model predictors as the predictor variables.

Transforming the Target Variable:\textsuperscript{165}

Sometimes it is useful to transform the target variable. Among the possible transformations:

- **Cap large losses** for purposes of modeling pure premium or severity.\textsuperscript{166}
- Remove catastrophe losses.
- Losses may need to be developed.\textsuperscript{167}
- Losses and/or exposures may need to be trended.
- Premium may need to be put on level.\textsuperscript{168}

Year could be included in the model, which should pick up any effects on the target variable related to time, such as trend, loss development, and rate changes.

\textsuperscript{160} See Section 5.1.2 of Goldburd, Khare, and Tevet.
\textsuperscript{161} Similar ideas would apply to Homeowners Insurance.
\textsuperscript{162} Or group of perils.
\textsuperscript{163} Wind and hail should be divided between catastrophe and non-catastrophes; with catastrophes modeled separately as discussed in the syllabus reading by Grossi and Kunreuther.
\textsuperscript{164} The data used for this procedure should reflect the expected mix going forward, and so using only the most recent year of exposures may be ideal. Since the target data fed into this new model is extremely stable, this procedure doesn’t require a whole lot of data.
\textsuperscript{165} See Section 5.1.3 of Goldburd, Khare, and Tevet, which discusses familiar things done in ratemaking.
\textsuperscript{166} Ideally the level chosen for the cap should capture most of the signal and eliminate most of the noise. This is similar in concept to choosing a reasonable accident limit to use in an Experience Rating Plan.
\textsuperscript{167} Either to ultimate or to a common level of maturity.
\textsuperscript{168} For a severity model, the development factor should reflect only expected future development on known claims. Since larger claims take on average longer to report, this may not address the whole issue. For some lines of insurance, one may be better off not using more recent but less mature data in the model. For a pure premium or loss ratio model, the development factor should include the effect of pure IBNR claims as well.
Choosing the Distribution for the Target Variable: If modeling claim frequency, the distribution is likely to be either Poisson or Negative Binomial. If modeling a binary response, then the Bernoulli or Binomial Distributions are used. If modeling claim severity, common choices for the distribution are Gamma and Inverse Gaussian. If modeling pure premiums, the Tweedie Distribution is a common choice.

Selection of Predictor Variables: Sometimes the actuary is just updating the parameters a model using newer data. Other times, the actuary will do a full review of all aspects of a model, including which predictor variables to include.

One would like a predictor variable to have a statistical significant effect on the target variable. Statistical tests can be performed. One would like a small probability value for the null hypothesis that the corresponding parameter is zero. There is no magic cutoff, although a p-value of 5% or less is often used. However, if the p-value is 5%, that means that there is 1/20 chance we are including a predictor variable in the model when we should not. If there is large set of possible predictor variables that are tested for inclusion in the model, this can lead to problems.

In addition to statistical significance, the actuary must take into account practical considerations. For example:

- Will it be cost effective?
- Actuarial standards of practice.
- Regulatory and legal requirements.
- Can the IT (Information Technology) department easily implement the change?

---

169 See Section 5.2 of Goldburd, Khare, and Tevet.
170 Analysis of the deviance residuals, to be discussed subsequently, can help the actuary to choose.
171 Recall that one can also use an overdispersed Poisson.
172 See Section 5.3 of Goldburd, Khare, and Tevet.
173 For a further discussion of p-values see the following subsection, the ASA statement not on the syllabus.
174 There are automated variable selection algorithms, which are not on the syllabus.
175 See ASOP 12: Risk Classification.
ASA Statement on Statistical Significance and P-values

Introduction

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

Underpinning many published scientific conclusions is the concept of “statistical significance,” typically assessed with an index called the p-value. While the p-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of p-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since p-values were first introduced.

In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the p-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in non-technical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

What is a p-value?

Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.
Principles

1. **P-values can indicate how incompatible the data are with a specified statistical model.**

A p-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called “null hypothesis.” Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. **P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**

Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. **Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.**

Practices that reduce data analysis or scientific inference to mechanical “bright-line” rules (such as “p < 0.05”) for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision-making. A conclusion does not immediately become “true” on one side of the divide and “false” on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, “yes-no” decisions, but this does not mean that p-values alone can ensure that a decision is correct or incorrect. The widespread use of “statistical significance” (generally interpreted as “p ≤ 0.05”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.
4. Proper inference requires full reporting and transparency.

P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference and “p-hacking,” leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Researchers should recognize that a p-value without context or other evidence provides limited information. For example, a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large p-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible.
Other approaches

In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct.

Conclusion

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.
Transformation of Predictor Variables:¹⁷⁸

In many cases, a variable will need to be transformed in some way such that the resulting GLM is a better fit to the data. We have already discussed how with a log link function it often make sense to take the log of a continuous variable. Partial Residual Plots are one way for the actuary to detect whether a transforming a predictor variable is indicated.

Partial Residual Plots:¹⁷⁹

Concentrate on one of the explanatory variables $X_j$.

Then the partial residuals are: $r_i = (\text{ordinary residual}) \ g'(\mu_i) + x_{ij} \hat{\beta}_j$.¹⁸⁰

In a Partial Residual Plot, we plot the partial residuals versus the variable of interest.

If there seems to be curvature rather than linearity in the plot, that would indicate a departure from linearity between the explanatory variable of interest and $g(\mu_i)$, adjusting for the effects of the other independent variables.

For a log link, $g'(\mu) = 1/\mu$, so that:

$$r_i = \frac{y_i - \hat{\mu}_i}{\mu_i} + \hat{\beta}_j \ x_{ij}.$$ ¹⁸⁰

¹⁷⁸ See Section 5.4 of Goldburd, Khare, and Tevet.
¹⁷⁹ See Section 5.4.1 of Goldburd, Khare, and Tevet.
¹⁸⁰ For the identity link ratio, this matches the definition of the partial residual for multiple regressions; the second term removes the effect of the $j^{\text{th}}$ variable on the prediction of the $i^{\text{th}}$ response, leaving the partial residual.
For example, assume a GLM where the fitted coefficient on ln[age of building] is -0.314. Assume the following graph of the partial residuals:\(^{181}\)

The linear estimate of the GLM, \(-0.314x\), is superimposed over the plot of the partial residuals. The points are missing the line in a systematic way, indicating that this model can be improved. The model is overpredicting for risks where log building age is less than 2.5, underpredicts between 2.5 and 3.25, and once again overpredicts for older buildings.

As will be discussed subsequently, one can add polynomial terms and examine the resulting graphs of partial residuals.

\(^{181}\) Figure 8 taken from Generalized Linear Models for Insurance Rating, by Goldburd, Khare and Tevet.
Binning Continuous Predictors: If there is nonlinearity, one possible fix for a continuous variable is to group it into intervals. For example, rather than treat age of construction as a continuous variable, one can group it into several categories. We have converted a continuous variable into a categorical variable. For their example, the authors group age of construction into ten bins. Figure 9 in the syllabus reading shows the resulting model: "The model picked up a shape similar to that seen in the points of the partial residual plot. Average severity rises for buildings older than ten years, reaching a peak at the 15-to-17 year range, then gradually declining."

---

182 See Section 5.4.2 of Goldburd, Khare, and Tevet.
183 The bins were chosen so that they each have roughly the same amount of data. While having bins with roughly equal amounts of data has advantages, it is not a necessity.
184 The error bars correspond to plus or minus two standard errors around the estimate.
Disadvantages of binning (grouping) continuous variables:
1. Adds parameters to the model.\(^{185}\)
2. Continuity in the estimates is not guaranteed.
   There is no guarantee that the pattern among intervals makes sense.\(^{186}\) \(^{187}\)
3. Variation within intervals is ignored.
   For example, it may be that the relativity for age of construction less than 5 years may be significantly different than that for 6 to 10 years. However, if we use an interval consisting of less than 10, our model can not pick up any such difference.
4. There may not be enough data in each bin to be credible.
5. There could be non-intuitive results, such as reversals.

Adding Polynomial Terms:\(^{188}\)

Rather than a model that uses \(\beta_0 + \beta_1 x_1 + \ldots\),
one can use \(\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \ldots\), or \(\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \ldots\)
The more polynomial terms that are included, the more flexibility, at the cost of greater complexity.

The authors added the square of the logged building age to their model. Here is the resulting plot of partial residuals with the curve formed by both building age terms superimposed:\(^{189}\)

---

\(^{185}\) By the principle of parsimony, we wish to avoid adding unnecessary parameters to the model.
\(^{186}\) For example, in the previous graph, the estimate for 21-23 years does not follow the general pattern.
\(^{187}\) One may be able to alleviate this problem by applying some smoothing process to the estimates from the model. Alternately, one could group together two or more intervals.
\(^{188}\) See Section 5.4.3 of Goldburd, Khare, and Tevet.
\(^{189}\) See Figure 10 in Goldburd, Khare, and Tevet.
Then the authors added the cube of the logged building age to their model. Here is the resulting plot of partial residuals with the curve formed by both building age terms superimposed:\textsuperscript{190}

“This perhaps yields a better fit, as the points seem to indicate that the declining severity as building age increases does taper off toward the higher end of the scale.”

Unfortunately, it is hard to interpret these models that include powers of the logged building age.

Using Piecewise Linear Functions:\textsuperscript{191}

Let $X_+ = X$ if $X \geq 0$ and 0 if $X < 0$.
Then a hinge function is: $\max\{0, X - c\} = (X - c)_+$, for some constant $c$.
The constant $c$ would be called the breakpoint.
Hinge functions can be used to create piecewise linear functions which can be used in GLMs.

For example, let $X = \ln[AOI]$. Then a usual linear estimator is: $\beta_0 - 0.314 \times + ...$
Using instead a hinge function: $\beta_0 + 1.225 \times - 2.269 (x - 2.75)_+ + .....\textsuperscript{192}$

\textsuperscript{190} See Figure 10 in Goldburd, Khare, and Tevet.
\textsuperscript{191} See Section 5.4.4 of Goldburd, Khare, and Tevet.
\textsuperscript{192} See Table 6 in Goldburd, Khare, and Tevet, “adding a breakpoint at 2.75.”
Here is a graph of the broken line that results from including the hinge function:

![Graph of the broken line](image)

For \( \ln[\text{AOI}] < 2.75 \), we have slope 1.225, while for \( \ln[\text{AOI}] > 2.75 \) we have a slope of:

\[
1.225 - 2.269 = -1.044.
\]

Instead we can use two hinge functions:

\[
\beta_0 + 1.289x - 2.472(x - 2.75)_+ + 1.170(x - 3.60)_+ + \ldots
\]

Here is a graph of the broken line that results from including two hinge functions:

![Graph of the broken line with two hinge functions](image)

For \( \ln[\text{AOI}] < 2.75 \), we have slope 1.289,
for \( 3.60 > \ln[\text{AOI}] > 2.75 \) the slope is: \( 1.289 - 2.472 = -1.183 \),
while for \( 3.60 > \ln[\text{AOI}] > 3.60 \) the slope is: \( 1.289 - 2.472 + 1.170 = -0.013 \).

---

193 See Table 7 in Goldburd, Khare, and Tevet, “adding an additional breakpoint at 3.6.”
Here is a graph of the partial residuals for the straight line:\(^{194}\)

![Graph of the partial residuals for the straight line]

Here is a graph of the partial residuals for the broken line that results from using one hinge function:\(^{195}\)

![Graph of the partial residuals for the broken line]

The model using the broken line does a better job of fitting the authors’ data than the model that uses the straight line.

---

\(^{194}\) See Figure 8 in Goldburd, Khare, and Tevet.

\(^{195}\) See Figure 11 in Goldburd, Khare, and Tevet.
Here is a graph of the partial residuals for the broken line using two hinge functions:

The model using two hinge functions may do a somewhat better job of fitting the authors’ data than the model that uses one hinge function. With limited data it is hard to tell.

Hinge functions provide more flexibility at the cost of greater complexity.

**The breakpoints must be selected by the user.**
“Generally, break points are initially guesstimated by visual inspection of the partial residual plot, and they may be further refined by adjusting them to improve some measure of model fit. However, the GLM provides no mechanism for estimating them automatically.”

---

196 See Figure 11 in Goldburd, Khare, and Tevet.
197 “As this leveling-off effect comports with our intuition, we may decide to keep the third hinge function term in the model.”
198 MARS is a variant of the GLM, which among other things, fits non-linear curves using hinge function, and does it in an automated fashion with no need for tweaking by the user.
A Cherry Tree Example:

We are given the height, diameter, and volume of 31 black cherry trees:


Heights are: 70, 65, 63, 72, 81, 83, 66, 75, 80, 75, 79, 76, 76, 69, 75, 74, 85, 86, 71, 64, 78, 80, 74, 72, 77, 81, 82, 80, 80, 80, 87.


I took $X_1 = \ln[\text{diameter}]$, $X_2 = \ln[\text{height}]$, and $Y = \text{volume}$.

A GLM was fit using a Gamma Distribution and a log link function.

The fitted parameters were: $\hat{\beta}_0 = -8.94859$, $\hat{\beta}_1 = 1.98041$, $\hat{\beta}_2 = 1.13288$.

$\hat{y} = \exp[-8.94859 + 1.9804 \ln[\text{diameter}] + 1.13288 \ln[\text{height}]]
= 0.00012992 \text{diameter}^{1.9804} \text{height}^{1.13288}$.

The covariance matrix is:

\[
\begin{pmatrix}
0.556725 & 0.00760542 & -0.13715 \\
0.00760542 & 0.00545975 & -0.00788943 \\
-0.13715 & -0.00788943 & 0.0405552
\end{pmatrix}
\]

Exercise: Based on geometry, it would make sense for $\beta_1 = 2$. Test whether $\beta_1 = 2$.

[Solution: $(1.98041 - 2) / \sqrt{0.00545975} = -0.265$. p-value is: $2 \Phi[-0.265] = 79.1\%$.]

Exercise: Based on geometry, it would make sense for $\beta_2 = 1$. Test whether $\beta_2 = 1$.

[Solution: $(1.13288 - 1) / \sqrt{0.0405552} = 0.660$. p-value is: $2 \{1 - \Phi[0.660]\} = 50.9\%$.]

---

199 The diameter is measured at 4′6″ above the ground.

Data from a study by Ryan, Joiner, and Ryan.
The first predicted volume is: \( \exp[-8.94859 + 1.9804 \ln[83] + 1.13288 \ln[70]] = 101.04 \).
Thus the first residual is: 103 - 101.04 = 1.96.

The residuals are: 1.96, 3.32, 1.31, -2.19, -9.14, -9.43, -9.12, -8.85, 16.96, 1.22, 28.50, 2.06, 6.06, 16.79, -35.74, -35.71, 36.48, -50.59, -20.02, -0.86, 23.33, -23.46, 38.15, 0.29, -2.43, 43.48, 27.42, 44.45, -29.52, -34.52, -12.21.

For this example, \( g(\mu) = \ln(\mu) \). Thus \( g'(\mu) = 1/\mu \). \( \Rightarrow r_i = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} + \hat{\beta}_j x_{ij} \).

Thus for \( \ln[\text{diameter}] \), the partial residuals are: \( \frac{(y_i - \hat{y}_i)}{\hat{\mu}_i} + \ln[(\text{diameter})] \). 1.98041.
The first partial residual is: \( \frac{(103 - 101.04)}{101.04} + \ln[83](1.98041) = 8.77 \).

Here is a graph of these partial residuals versus \( \ln[\text{diameter}] \):

A departure from linearity is not evident.\(^\text{200}\)

\(^{200}\) If there seems to be curvature rather than linearity in the plot, that would indicate a departure from linearity between the independent variable of interest and \( g(\mu) \), adjusting for the effects of the other independent variables.
Exercise: For ln[height] what is the first partial residual?

[Solution: The partial residuals are: \((y_i - \hat{y}_i) / \hat{y}_i + \ln([\text{height}])\)] 1.13288.

The first partial residual is: \((103 - 101.04)/101.04 + \ln[70] (1.13288) = 4.83\).

The partial residuals for the ln[height] are: 4.83, 4.76, 4.71, 4.83, 4.93, 4.96, 4.69, 4.84, 5.05, 4.90, 5.08, 4.92, 4.94, 4.88, 4.73, 4.74, 5.15, 4.89, 4.76, 4.71, 5.01, 4.90, 4.99, 4.85, 4.92, 5.06, 5.04, 5.05, 4.91, 4.90, 5.04.

Here is a graph of these partial residuals versus ln[height]:

A departure from linearity is not evident.
Grouping Categorical Variables

Some predictor variables are ordinal; they are discrete with several categories with a natural order. Sometimes it is useful for modeling purposes to group such predictor variables into fewer categories. This is particularly useful when there are many categories.

For example, workers compensation claims are categorized as: medical only, temporary total, minor permanent partial, major permanent partial, permanent total, and fatal. For some purposes it might be useful to group the first three categories into nonserious and the last three categories into serious.

One can start with a model without grouping. Statistical tests can determine whether the coefficients of adjacent levels are significantly different. Then one can group adjacent levels with similar fitted coefficients. Now run a new model using these groupings, and iterate the procedure. One needs to balance the competing priorities of: predictive power, parsimony, and avoiding overfitting.

Interactions

If \( x_1 \) and \( x_2 \) are predictor variables, then we can include an interaction term: \( x_1 x_2 \).
Then the model would be: \( g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \ldots \).
This provides more flexibility at the cost of complexity.

For example let \( x_1 \) be gender and \( x_2 \) be age. Then if we include an interaction term the effect of age depends on gender, and the effect of gender depends on age.

The syllabus reading gives an example with building occupancy class and sprinkler status. Models are fit both with and without an interaction term. The model with interactions is:

\[
\mu = \text{(mean for base)} \exp[0.2303 x_1 + 0.4588 x_2 + 0.0701 x_3 - 0.2895 x_4 - 0.2847 x_1 x_4 - 0.0244 x_2 x_4 - 0.2622 x_3 x_4],
\]
where \( x_1 = 1 \) if occupancy class 2, \( x_2 = 1 \) if occupancy class 3, \( x_3 = 1 \) if occupancy class 4, \( x_4 = 1 \) if sprinklered, and occupancy class 1 without sprinklers is the base.

---

201 See Section 5.5 in Goldburd, Khare, and Tevet.
202 This is analogous to Robertson grouping classes into Hazard Groups.
203 The syllabus reading uses the example of driver age, which can be thought of as either continuous or discrete. In the case of age, there may not be any clear breakpoints to use for grouping; actuarial judgement may be needed.
204 See Section 5.6 in Goldburd, Khare, and Tevet.
205 One would only include the interaction term if its coefficient were significantly different from zero.
206 This is a commercial building claims frequency model using a Poisson with a log link function.
207 See Tables 8 and 9 in Goldburd, Khare, and Tevet.
While two of the interaction terms are significantly different from zero, the remaining one is not. They show an intercept which only makes sense if there are other predictor variables in the model.
For a non-sprinklered building in occupancy class 2, the multiplicative relativity to the base is: \( \exp[0.2303] = 1.259 \).

For a sprinklered building in occupancy class 2, the multiplicative relativity to the base is: \( \exp[0.2303 - 0.2895 - 0.2847] = 0.709 \).

Exercise: For a non-sprinklered building in occupancy class 4, determine the multiplicative relativity to the base.
[Solution: \( \exp[0.0701] = 1.073 \).]

Exercise: For a sprinklered building in occupancy class 4, determine the multiplicative relativity to the base.
[Solution: \( \exp[0.0701 - 0.2985 + 0.2622] = 1.044 \).]

Comment: For occupancy class 4, the effect of sprinklers is small, while for occupancy class 2, the effect of sprinklers is large.

The syllabus reading also shows another fitted model, with occupancy class, sprinklered, \( \ln[\text{AOI/200,000}] \), plus an interaction term between sprinklered and \( \ln[\text{AOI/200,000}] \):$^{208}$

\[
\mu = (\text{mean for base}) \exp[0.2919 x_1 + 0.3510 x_2 + 0.0370 x_3 - 0.5153 x_4 + 0.4239 x_5 - 0.1032 x_4 x_5],
\]

where \( x_1 = 1 \) if occupancy class 2, \( x_2 = 1 \) if occupancy class 3, \( x_3 = 1 \) if occupancy class 4, \( x_4 = 1 \) if sprinklered, \( x_5 = \ln[\text{AOI/200,000}] \),

and \( \text{AOI} = 200,000 \) in occupancy class 1 without sprinklers is the base.

For a non-sprinklered building in occupancy class 2 with \( \text{AOI} = 500,000 \), the multiplicative relativity to the base is: \( \exp[0.2919 + 0.4239 \ln[2.5]] = 1.975 \).

Exercise: For a sprinklered building in occupancy class 2 with \( \text{AOI} = 500,000 \), determine the multiplicative relativity to the base.
[Solution: \( \exp[0.2919 - 0.5153 + 0.4239 \ln[2.5] - 0.1032 \ln[2.5]] = 1.073 \).]

For range of sizes of AOI for buildings that are insured, the expected frequency increases at a slower rate with AOI for sprinklered buildings than for non-sprinklered buildings.

---

$^{208}$ See Table 12 in Goldburd, Khare, and Tevet. They show an intercept of -3.771, which implies an expected frequency for the base level of: \( \exp[-3.771] = 2.3\% \).
Loglikelihood:

The loglikelihood is the sum of the contributions of the ln[density] at each of the observations. All other things being equal, a larger loglikelihood indicates a better fit. However, the principle of parsimony means that we should not add additional parameters to a model unless it significantly increases the loglikelihood.

The saturated model has as many parameters as the number of observations. Each fitted value equals the observed value. The saturated model has the largest possible likelihood, of models of a given form.

The minimal model has only one parameter, the intercept. The minimal model has the smallest possible likelihood, of models of a given form.

Deviance:

The deviance is twice the difference between the maximum loglikelihood for the saturated model (with as many parameters as data points) and the maximum loglikelihood for the model of interest.

\[ D = \text{Deviance} = 2 \{ (\loglikelihood \text{ for the saturated model}) - (\loglikelihood \text{ for the fitted model}) \}. \]

The smaller the deviance, the better the fit of the GLM to the data. Maximizing the loglikelihood is equivalent to minimizing the deviance.

By definition, the deviance of the saturated model is zero. Even though the saturated model fits the data perfectly we would not use it to predict the future, since the saturated model is overfit; the saturated model picks up too much of the randomness in the data (called the noise).

The minimal model has the largest possible deviance while the saturated model has the smallest possible deviance of zero. The deviance of a fitted model will lie between those two extremes.

We will be comparing the deviance of models with the same distributional form, same dispersion parameter, and link function, that have been fit to the same data.

---

209 See Section 6.1.1 in Goldburd, Khare, and Tevet. They do not give any details on the form of the deviance for the different distributions. See for example, An Introduction to Generalized Linear Models by Dobson and Barnett.

210 Also sometimes called the null model.

211 In the context of GLMs we would be comparing models with the same distributional form and link function, that have been fit to the same data.

212 See Section 6.1.2 in Goldburd, Khare, and Tevet.

213 Subsequently we will discuss how to test whether an improvement in deviance is statistically significant.

214 If a variable has missing values for some records, the default behavior of most model fitting software is to toss out those records when fitting the model. In that case, the resulting measures of fit are no longer comparable, since the second model was fit with fewer records than the first.
Nested Models and the F-Test

We can use the F-Test to compare two nested models, in other words when one model is a special case of the other. The bigger (more complex) model always has a smaller (better) deviance than the smaller (simpler) model. The question is whether the deviance of the bigger model is significantly better than that of the smaller model (special case).

Assume that we have two nested models.
Then the test statistic (asymptotically) follows an F-Distribution with numbers of degrees of freedom equal to: $\nu_1 = \text{the difference in number of parameters}$, and 
$\nu_2 = \text{number of observations minus number of fitted parameters for the smaller model}$.\(^{217}\)

The test statistic is: \[
\frac{(D_S - D_B)}{\hat{\phi}_S} / \text{(number of added parameters)} \sim F_{\nu_1, \nu_2} \]

$D_S$ = deviance for the smaller (simpler) model.
$D_B$ = deviance for the bigger (more complex) model.
$\hat{\phi}_S$ = estimated dispersion parameter for the smaller (simpler) model.\(^{219}\)\(^{220}\)\(^{221}\)

---

\(^{215}\) See Section 6.2.1 in Goldburd, Khare, and Tevet.
\(^{216}\) This F-Test is analogous to that used to test slopes in multiple regression.
\(^{217}\) A Table of the F-Distribution is not attached to your exam, although they could give some values in a question. An F-Distribution is the ratio of two independent Chi-Square Distributions, with each Chi-Square divided by its number of degrees of freedom. $\nu_1$ = the number of degrees of freedom of the Chi-Square in the numerator.
$\nu_2$ = the number of degrees of freedom of the Chi-Square in the denominator.
If $\nu_1 = 1$, then the F-Distribution is related to the t-distribution.

\[^{218}\] I have seen instead in the denominator the estimated dispersion parameter of the more complex model. In that case, the degrees of freedom associated with the denominator are those of the more complex model. See “A Practitioners Guide to Generalized Linear Models,” by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Dora Schirmacher, Ernesto Schirmacher and Neeza Thandi, in the 2004 CAS Discussion Paper Program.

\[^{219}\] The syllabus reading does not discuss how to estimate the dispersion parameter. One way to estimate the dispersion parameter in a model is as the ratio of the deviance to the number of degrees of freedom of the model.\(^{219}\)

\[^{220}\] There is no requirement that the estimated dispersion parameters of the two models be equal.

\[^{221}\] For cases where the dispersion parameter is one, such as for a Poisson or Negative Binomial Distribution, an actuary would normally use instead the likelihood ratio test, not discussed in the syllabus reading. See “A Practitioners Guide to Generalized Linear Models,” by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Dora Schirmacher, Ernesto Schirmacher and Neeza Thandi, in the 2004 CAS Discussion Paper Program.
If the F-Statistic is sufficiently big, then reject the null hypothesis that the data is from the smaller model in favor of the alternate hypothesis that the data is from the bigger model.\textsuperscript{222}

Exercise: A GLM using a Gamma Distribution has been fit for modeling expenditures upon admission to a hospital. There are 150 observations. It uses 25 variables. It uses 4 categories of self-rated physical health: poor, fair, good, and very good. The deviance is 35.1. An otherwise similar GLM excluding self-rated physical health has a deviance of 38.4. The estimated dispersion parameter for this simpler model is 0.3.

Discuss how you would determine whether physical health is a useful variable for this model. [Solution: The more complex model has 25 variables, and 150 - 25 = 125 degrees of freedom. In order to incorporate physical health, avoiding aliasing, we need 4 - 1 = 3 variables. Thus the simpler model has 22 variables, and 150 - 22 = 128 degrees of freedom. The difference in degrees of freedom is: 128 - 125 = 3 = number of additional variables. Test statistic is: $\frac{D_S - D_B}{\hat{\phi}_S} = \frac{38.4 - 35.1}{3(0.3)} = 3.67$. We compare the test statistic to an F-distribution with 3 and 128 degrees of freedom. The null hypothesis is to use the simpler model, the one without physical health. The alternate hypothesis is to use the more complex model. We reject the null hypothesis if the test statistic is sufficiently big. Comment: The syllabus reading gives a similar example.]

Using a computer, the p-value of this test is 1.4%.\textsuperscript{223} Thus at a 2.5% significance level we would reject the simpler model in favor of the more complex model. At a 1% significance level we would not reject the simpler model.

If we had used a 2.5% significance level, we would have decided to use physical health. We had used four levels of physical health: poor, fair, good, and very good. The next step would be to see how many of these levels are useful. For example, does it significantly improve model performance to separate good from very good?

\textsuperscript{222} The F-Distribution with $\nu_1$ and $\nu_2 > 2$ degrees of freedom has a mean of $\nu_2/(\nu_2 - 1)$. For $\nu_2$ large this mean is approximately 1. We reject the null hypothesis if the F-Statistic is significantly greater than 1.

\textsuperscript{223} The 2.5% critical value is 3.220, while the 1% critical value is 3.938. In other words, for the F-Distribution with 3 and 129 degrees of freedom, the survival function at 3.220 is 2.5%.
AIC and BIC:

AIC and BIC are each methods of comparing models. In each case, a smaller value is better. These penalized measures of fit are particularly useful for comparing models that are not nested.

The Akaike Information Criterion (AIC) is used to compare a bunch of models all fit via maximum likelihood to the same data. The model with the smallest AIC is preferred. For a particular model:

\[
AIC = (-2) \times (\text{maximum loglikelihood}) + (\text{number of parameters})(2) \]

The number of parameters fitted via maximum likelihood are the betas (slopes plus if applicable an intercept).

Since the deviance = (2) (saturated max. loglikelihood - maximum loglikelihood for model), we can compare between the models: Deviance + (number of parameters)(2).

Assume for example, assume we have three Generalized Linear Models fit to the same data:

<table>
<thead>
<tr>
<th>Model #</th>
<th>Number of Parameters</th>
<th>Deviance</th>
<th>Deviance + (number of parameters)(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>888.7</td>
<td>896.7</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>886.2</td>
<td>896.2</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>884.4</td>
<td>896.4</td>
</tr>
</tbody>
</table>

We prefer Model #2, since it has the smallest AIC.

The Bayesian Information Criterion (BIC) can also be used to compare a bunch of models all fit via maximum likelihood to the same data. The model with the smallest BIC is preferred. For a particular model:

\[
BIC = (-2) \times (\text{max. loglikelihood}) + (\text{number of parameters}) \ln(\text{number of data points}) \]

Since the deviance = (2) (saturated max. loglikelihood - maximum likelihood for model), we can compare between the models:

Deviance + (number of parameters) ln(number of data points).

---

224 See Section 6.2.2 in Goldburd, Khare, and Tevet.
225 Thus AIC can be applied to Generalized Linear Models.
226 If a dispersion parameter is fit via maximum likelihood, then the number of parameters in the above formula for AIC is one more. However, if one is using AIC to compare models, it does not matter, as long as one is consistent, since the only difference is to add the same constant to each AIC.
227 The maximum likelihood for the saturated model is the same in each case.
228 In each case, the AIC is: Deviance + (number of parameters)(2) - (2)(loglikelihood for the saturated model).
229 Thus BIC can be applied to Generalized Linear Models.
230 The GLM monograph uses In and log interchangeably to both mean the natural log.
231 The maximum likelihood for the saturated model is the same in each case.
Assume that we have three Generalized Linear Models fit to the same data set of size 20:

<table>
<thead>
<tr>
<th>Model #</th>
<th>Number of Parameters</th>
<th>Deviance</th>
<th>Deviance + (number parameters) ln(20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>888.7</td>
<td>900.7</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>886.2</td>
<td>901.2</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>884.4</td>
<td>902.4</td>
</tr>
</tbody>
</table>

We prefer Model #1, since it has the smallest BIC.\(^{232}\)

We note that in this case, using AIC or BIC would result in different conclusions.

**BIC is mathematically equivalent to the Schwarz Bayesian Criterion.**\(^{233}\) Using the Schwarz Bayesian Criterion, one adjusts the loglikelihoods by subtracting in each case the penalty: 

\[
\text{(number of fitted parameters) ln(number of data points)} / 2.
\]

One then compares these penalized loglikelihoods directly; larger is better.

For a model, when BIC is smaller this penalized loglikelihood is bigger and vice-versa.

“As most insurance models are fit on very large datasets, the penalty for additional parameters imposed by BIC tends to be much larger than the penalty for additional parameters imposed by AIC. In practical terms, the authors have found that AIC tends to produce more reasonable results. Relying too heavily on BIC may result in the exclusion of predictive variables from your model.”

---

\(^{232}\) In each case, the BIC is: Deviance + (number of parameters)ln[20] - (2)(loglikelihood for the saturated model).

\(^{233}\) See for example Loss Models, not on the syllabus of this exam.
A Communicable Disease Example

Assume we have the following reported occurrences of a communicable disease in two areas:

<table>
<thead>
<tr>
<th>Number in Area A</th>
<th>Number in Area B</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>17</td>
<td>19</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>17</td>
<td>25</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td>23</td>
<td>20</td>
</tr>
</tbody>
</table>

Let $X_1 = 0$ if Region A and 1 if Region B.
Let $X_2 = \ln[\text{month}]$.
Fit a GLM with a Poisson using a log link function.
$\mu = \exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2]$.

The fitted parameters are: $\beta_0 = 1.54894$, $\beta_1 = 0.265964$, $\beta_2 = 0.435105$.

The covariance matrix is:

$$
\begin{pmatrix}
0.0618301 & -0.00781226 & -0.0226385 \\
-0.00781226 & 0.0138001 & -6.28837 \times 10^{-18} \\
-0.0226385 & -6.28837 \times 10^{-18} & 0.00948766
\end{pmatrix}
$$

Therefore, approximate 95% confidence intervals for the parameters are:

$1.54894 \pm 1.960 \sqrt{0.0618301} = (1.06, 2.04)$,
$0.265964 \pm 1.960 \sqrt{0.0138001} = (0.04, 0.50)$,
$0.435105 \pm 1.960 \sqrt{0.00948766} = (0.24, 0.63)$.

The loglikelihood is: -47.0892.
The Deviance is: 4.45650.

---

234 Adapted from Section 18.4 of Applied Regression Analysis by Draper and Smith, not on the syllabus.
In order to test whether \( \beta_1 = 0 \), the test statistic is:
\[
\frac{\hat{\beta}_1}{\text{StdDev}[\hat{\beta}_1]} = 0.265964 / \sqrt{0.0138001} = 2.264.
\]
The probability value of a two-sided test is: \( 2\{1 - \Phi[2.264]\} = 2.4\% \).\(^{235}\)

Exercise: Test whether \( \beta_2 = 0 \).

[Solution: \( \frac{\hat{\beta}_2}{\text{StdDev}[\hat{\beta}_2]} = 0.435105 / \sqrt{0.00948766} = 4.467 \).

The probability value of a two-sided test is: \( 2\{1 - \Phi[4.467]\} = 0\% \).

Comment: Using a computer, the p-value is \( 8 \times 10^{-6} \).]

Exercise: Test whether \( \beta_0 = 2 \).

[Solution: \( \frac{(\hat{\beta}_0 - 2)}{\text{StdDev}[\hat{\beta}_0]} = (1.54894 - 2) / \sqrt{0.0618301} = -1.814 \).

The probability value of a two-sided test is: \( 2\Phi[-1.814] = 7.0\% \).]

Now fit an otherwise similar GLM ignoring region, in other words without the dummy variable \( X_1 \).
The fitted parameters are: \( \hat{\beta}_0 = 1.69074, \hat{\beta}_2 = 0.435105 \).

The covariance matrix is:
\[
\begin{pmatrix}
0.0574127 & -0.0226404 \\
-0.0226404 & 0.00948839
\end{pmatrix}
\]

Therefore, approximate 95\% confidence intervals for the parameters are:
\( \hat{\beta}_0: 1.69074 \pm 1.960 \sqrt{0.0574127} = (1.22, 2.16) \),
\( \hat{\beta}_2: 0.435105 \pm 1.960 \sqrt{0.00948839} = (0.24, 0.63) \).

The loglikelihood is: -49.6747.
The Deviance is: 9.62755.

For the model including region, the loglikelihood is -47.0892.
There are 20 data points and this model has 3 fitted betas.
AIC = (-2)(-47.0892) + (3)(2) = 100.178.
BIC = (-2)(-47.0892) + 3 \ln(20) = 103.166.

For the simpler model excluding region, the loglikelihood is -49.6747.
This model has only 2 fitted betas.
AIC = (-2)(-49.6747) + (2)(2) = 103.349.
BIC = (-2)(-49.6747) + 2 \ln(20) = 105.341.

\(^{235}\) There is \textit{not} a Normal Distribution Table attached to your exam.
The first more complicated model has the smaller AIC and thus is preferred on this basis. The more complicated model has the smaller BIC and thus is also preferred on this basis.

The first model has a Deviance of 4.45650, while the second simpler model has a Deviance of 9.62755. Equivalently, we can use these Deviances.

For the first model, Deviance + (number of parameters)(2) = 4.45650 + (3)(2) = 10.45650.
For the second model, Deviance + (number of parameters)(2) = 9.62755 + (2)(2) = 13.62755.
Since 10.45650 < 13.62755, the first more complicated model is preferred on this basis.\[236\]

For the first model, Deviance + (number of parameters) ln(sample size) = 4.45650 + 3 ln(20) = 13.444.
For the second model, Deviance + (number of parameters) ln(sample size) = 9.62755 + 2 ln(20) = 15.619.
Since 113.444 < 15.619, the first more complicated model is also preferred on this basis.\[237\]

Deviance Residuals\[238\]

The (ordinary) residuals are the difference between the observed and fitted values. Other types of residuals are useful when working with GLMs, including Deviance Residuals.\[239\]
Deviance Residuals provide a more general quantification of the conformity of a case to the model specification.

Deviance Residuals are based on the form of the deviance for the particular distribution. Since the syllabus reading does not discuss these forms, you are not responsible for them on this exam.

The square of the deviance residual is the corresponding term in the sum that is the Deviance.\[240\]

We take the sign of the deviance residual as the same as that of the (ordinary) residual \( y_i - \hat{\mu}_i \).

\[236\] This is equivalent to comparing AICs.
\[237\] This is equivalent to comparing BICs.
\[238\] See Section 6.3 in Goldburd, Khare, and Tevet.
\[239\] Pearson Residuals and Anscombe Residuals are also used, but these are not on the syllabus. See for example Generalized Linear Models by McCullagh and Nelder, Generalized Linear Models for Insurance Data by de Jong and Heller, and An Introduction to Generalized Linear Models by Dobson and Barnett.
\[240\] The syllabus reading incorrectly states that the deviance residual itself rather than its square is the corresponding term in the sum that is the deviance.
“We can think of the deviance residual as the residual adjusted for the shape of the assumed GLM distribution, such that its distribution will be approximately Normal if the assumed GLM distribution is correct.”

If the fitted model is appropriate, then we expect:
• The deviance residuals should follow no predictable pattern. 241
• The deviance residuals should be Normally distributed, with constant variance. 242

The syllabus reading shows an example of how to determine whether the deviance residuals are Normal. In the first case, a model was fit with a Gamma Distribution. 243

In the histogram, the deviance residuals do not seem close to the best fit Normal. 244
In the Normal Q-Q plot, the deviance residuals are not near the comparison straight line. 245
We conclude that the deviance residuals are not Normal and therefore the Gamma Distribution is probably not a good choice to model this data.

In the histogram, the deviance residuals are skewed to the right. Thus an Inverse Gaussian Distribution with greater skewness than a Gamma Distribution, might be better for modeling this data.

241 If we discover a pattern in the deviance residuals then we can probably improve our model to pick this pattern up.
242 The property of constant variance is called homoscedasticity. Homoscedasticity is more closely followed for standardized deviance residuals, not on the syllabus.
If the model is correct, standardized residuals should (approximately) follow a Standard Normal Distribution. See an An Introduction to Generalized Linear Models by Dobson and Barnett.
243 See Figure 15 in Goldburd, Khare, and Tevet.
244 See for example Loss Models, not on the syllabus of this exam.
245 See for example Loss Models, not on the syllabus of this exam.
Here is similar graphs for a model that was fit with an Inverse Gaussian Distribution:\(^{246}\)

In the histogram, the deviance residuals are much closer to the best fit Normal than before. In the Normal Q-Q plot, the deviance residuals are much nearer to the comparison straight line than before.

We conclude that the deviance residuals are closer to Normal, and therefore the Inverse Gaussian Distribution is probably a better choice to model this data than the Gamma Distribution.

**Deviance Residuals for Discrete Distributions:**

For discrete distributions such as Poisson or Negative Binomial, or distributions that have a point mass such as the Tweedie, the deviance residuals will likely not follow a Normal Distribution.\(^{247}\) This makes deviance residuals less useful for assessing the appropriateness of such distributions, when each record is for a single risk.\(^{248}\)

Fortunately, for data sets where one record may represent the average frequency for a large number of risks, deviance residuals are more useful than when each record is for a single risk.\(^{249}\)

---

\(^{246}\) See Figure 16 in Goldburd, Khare, and Tevet.

\(^{247}\) This is because the deviance residuals do not adjust for the discreteness; the large numbers of records having the same target values cause the residuals to be clustered together in tight groups.

\(^{248}\) One possible solution is to use randomized quantile residuals, which add random jitter to the discrete points so that they wind up more smoothly spread over the distribution.

\(^{249}\) The target variable will take on a larger number of distinct values, effectively smoothing out the resulting distribution causing it to lose much of its discrete property and approach a continuous distribution.
**Review. Histograms:**

A histogram is an approximate graph of the probability density function. First we need to group the data into intervals.

The height of each rectangle = \( \frac{\text{# values in the interval}}{\text{(total # values) (width of interval)}} \).

For example, let us assume we observe 100 values and group them into four intervals:

- Number that are between -0.15 and -0.05: 10
- Number between -0.05 and 0: 30
- Number between 0 and 0.05: 40
- Number between 0.05 and 0.15: 20

The first interval has width 0.1. The probability in the first interval is: \( \frac{10}{100} \).

We want the area of the first rectangle to be equal to the probability in the first interval.

\[(0.1)(\text{height}) = \frac{10}{100}. \Rightarrow \text{Height} = \frac{10/100}{0.1} = 1.\]

Similarly, the height of the second rectangle is: \( \frac{30}{100} / 0.05 = 6. \)
The height of the third rectangle is: \( \frac{40}{100} / 0.05 = 8. \)
The height of the fourth rectangle is: \( \frac{20}{100} / 0.10 = 2. \)

The histogram of these 100 values:

![Histogram](image)

The sum of the areas of the rectangles is: \( (0.1)(1) + (0.05)(6) + (0.05)(8) + (0.1)(2) = 1. \)

In general the area under a histogram should sum to one, just as for the graph of a probability density function.
Exercise: Draw a histogram of the following grouped data: 0 -10: 6, 10-20: 11, 20-25: 3.

[Solution: The heights are: \( \frac{6}{(20)(10)} = 0.03 \), \( \frac{11}{(20)(10)} = 0.055 \), and \( \frac{3}{(20)(5)} = 0.03 \).

Comment: The sum of the areas of the rectangles is: \( (10)(0.03) + (10)(0.055) + (5)(0.03) = 1 \). With more data, we would get a better idea of the probability density function from which this data was drawn.]

Creating a histogram and comparing the histogram to a graph of a Normal Distribution is one way to determine whether the items of interest appear to be Normally distributed.

First we would want the histogram to look roughly symmetric, since the Normal Distribution is symmetric around its mean.\(^{250}\)

\(^{250}\) If the values are from a Normal Distribution, then one would expect the skewness of the observed values to be close to zero. In addition, since a Normal Distribution has a kurtosis of 3, if the values are from a Normal Distribution, then one would expect the kurtosis of the observed values to be close to 3.
The following histogram is not symmetric, and thus not likely to be a sample from a Normal Distribution:

![Histogram 1](image1)

The following histogram looks approximately symmetric:

![Histogram 2](image2)

---

251 This histogram was based on 1000 data points simulated from a shifted Gamma Distribution.
However, one can superimpose upon it a Normal Distribution with parameters $\mu = \overline{X}$ and $\sigma = \text{sample variance}$:

The histogram of the data seems to be more highly peaked than the Normal and may have heavier tails.\textsuperscript{252} This data has a larger kurtosis than a Normal; the graph displays leptokurtosis.\textsuperscript{253} 

The following histogram, is based on a random sample of size 1000 from a Normal Distribution:

\textsuperscript{252} Heavier tails means more probability in both the lefthand and righthand tails.
\textsuperscript{253} Kurtosis = 4th central moment / square of the variance. All Normal Distributions have a kurtosis of 3, so one would want the kurtosis of the data to also be close to 3. For the data that generated this histogram the kurtosis is 3.85, indicating somewhat heavier tails than a Normal Distribution.
I superimposed upon the above histogram a Normal Distribution, with parameters $\mu = \bar{X}$ and $\sigma = $ sample variance:

As with any finite sample, while the match between the data and a fitted Normal Distribution seems reasonable, it is far from perfect.
Next I simulated 10,000 random draws from a Gamma Distribution (with $\alpha = 4$), and then subtracted a constant.\footnote{The key idea here is that the Gamma is some distribution different than the Normal Distribution.} I then compared a histogram of the data to the probability density function of a Normal Distribution with parameters based on the sample mean and sample variance of the data:

![Histogram and Normal Distribution](image)

The curve of the Normal Distribution is a poor match to the data represented by the histogram.\footnote{Since this Normal Distribution has the same mean and variance as the data, we would expect it to be a good match to the data, provided the data were drawn from a Normal Distribution.}

Even if we did not know the data was simulated from another distribution, we would conclude that this data was not drawn from a Normal Distribution.

**Review, Q-Q Plots:**

A Q-Q plot or quantile-quantile plot is a graphical technique which can be used to either compare a data set and a distribution or compare two data sets. Q-Q plots are most commonly used as a visual test of whether data is appears to be from a Normal Distribution. These are sometimes called Normal Q-Q Plots.

The 95th percentile is also referred to as $Q_{0.95}$, the 95% quantile.

For a distribution, the quantile $Q_\alpha$ is such for $F(Q_\alpha) = \alpha$. In other words, $Q_\alpha = F^{-1}(\alpha)$.

For example, $Q_{0.95}$ for a Standard Normal Distribution is 1.645, since $\Phi[1.645] = 0.95$.
In order to see whether data is drawn from some member of a Distribution Family, which has a scale and/or location parameter, we can create a Q-Q Plot for a standard member of that family $F$.

- Grade the $n$ data points from smallest to largest.
- For $i = 1$ to $n$, plot the points: $(F^{-1}[\frac{i}{n+1}], x_{(i)})$.

If the data is drawn from the given distribution family, then we expect the plotted points to lie close to some straight line.

Take the following 24 data point arranged from smallest to largest:

565, 678, 681, 713, 769, 809, 883, 890, 906, 909, 946, 956, 961, 983, 1046, 1073, 1103, 1171, 1198, 1269, 1286, 1296, 1316, 1643.

For the Standard Normal, $Q_{1/25} = Q_{0.04} = -1.751$.

Thus the first plotted point in a Normal Q-Q Plot is: $(-1.751, 565)$.

Exercise: What is the second plotted point?

[Solution: $Q_{2/25} = Q_{0.08} = -1.405$. Thus the second plotted point is: $(-1.405, 678)$.

Here is the resulting Normal Q-Q Plot:
Other than the final point, the plotted points seem approximately linear, and thus this data could very well be from a single Normal Distribution.\(^{256}\)

One could standardize each data point prior to constructing the Q-Q plot. The data has a sample mean of 1002.08, and a sample variance of 63,387.9. Thus we would subtract 1002.08 from each data point and divide by \(\sqrt{63,387.9}\). For example, \((565 - 1002.08) / \sqrt{63,387.9} = -1.736\).

Here is the Q-Q Plot, using the standardized data, including the comparison line \(x = y\):\(^{257}\)

![Q-Q Plot](image)

Again, other than the final point, the plotted points are close to the 45 degree comparison line, and thus this data could very well be from a single Normal Distribution.

\(^{256}\) With small data sets it is hard to draw a definitive conclusion. There is no specific numerical test we would apply to the Q-Q plot.

\(^{257}\) Having standardized the data, when we compare to the Standard Normal Distribution, we expect the plotted points to be close to the 45 degree comparison line \(x = y\).
The form of the deviance residual depends on the distribution and thus the form of the deviance.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Square of the Deviance Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>(\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2)</td>
</tr>
<tr>
<td>Poisson</td>
<td>(2 \sum_{i=1}^{n} {y_i \ln[y_i / \hat{\lambda}_i] - (y_i - \hat{\lambda}_i)})</td>
</tr>
<tr>
<td>Binomial</td>
<td>(2 \sum_{i=1}^{n} {y_i \ln[\frac{y_i}{\hat{y}_i}] + (m_i - y_i) \ln[\frac{m_i - y_i}{m_i - \hat{y}_i}]})</td>
</tr>
<tr>
<td>Gamma</td>
<td>(2 \alpha \sum_{i=1}^{n} {-\ln[y_i / \hat{y}_i] + (y_i - \hat{y}_i) / \hat{y}_i})</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>(\theta \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i^2 y_i})</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>(2 \sum_{i=1}^{n} {y_i \ln[y_i / \hat{y}_i] - (y_i + r) \ln[\frac{y_i + r}{\hat{y}_i + r}]})</td>
</tr>
</tbody>
</table>

Exercise: For a GLM using a Gamma Distribution, the first observed value is 800 with corresponding fitted value of 853.20. The maximum likelihood fitted parameter \(\alpha = 45.6\). What is the corresponding deviance residual?

[Solution: \(d_1^2 = (2)(45.6) \{-\ln[800/853.20] + (800 - 853.20)/853.20\} = 0.1850.\) Since 800 - 853.20 is negative, we take the deviance residual as negative. \(d_1 = -\sqrt{0.1850} = -0.430.\)]

Comment: This is for the two-dimensional example I discussed previously, using a reciprocal link function.\[258\] Not on the syllabus of this exam.
**Communicable Disease Example Continued:**

For the Poisson Distribution, the deviance is:

\[ D = 2 \sum_{i=1}^{n} \{ y_i \ln\left( \frac{y_i}{\hat{\lambda}_i} \right) - (y_i - \hat{\lambda}_i) \}. \]

Then the square of the deviance residual is the corresponding term in the above sum:

\[ d_i^2 = 2 \{ y_i \ln\left( \frac{y_i}{\hat{\lambda}_i} \right) - (y_i - \hat{\lambda}_i) \}. \]

For example, for the Communicable Disease Example which uses a Poisson Distribution, the first observed count is 8 with corresponding fitted value 6.3632. Thus \( d_1^2 = 2 \{ 8 \ln\left( \frac{8}{6.3632} \right) - (8 - 6.3632) \} = 0.3889. \)

Since the first ordinary residual is positive, \( d_1 = \sqrt{0.3889} = 0.6236. \)

Exercise: For this example, the third observed count is 10 with corresponding fitted value 10.263. Determine the corresponding deviance residual.

[Solution: \( d_3^2 = 2 \{ 10 \ln\left( \frac{10}{10.263} \right) - (10 - 10.263) \} = 0.006798. \)

Since \( 10 - 10.263 < 0, \) we take the deviance residual as negative. \( d_3 = -\sqrt{0.006798} = -0.0824. \)]

For this example, the deviance residuals are: 0.6237, -0.2081, -0.0824, -0.1869, 0.3254, 0.8099, -0.4876, -0.1845, 0.1094, -0.5728, 0.2390, 0.2289, -1.2763, -0.3058, -0.4288, 0.2090, 0.1448, 0.1061, 0.7144, 0.0819.

If the model is correct, then asymptotically the deviance is Chi-Square with \( n - p \) degrees of freedom, where \( n \) is the number of observations and \( p \) is the number of fitted parameters. Thus the expected value of the deviance is \( n - p. \) Therefore, we expect each term of the sum, \( d_i^2, \) to contribute about \( (n - p)/n \equiv 1. \) Thus we expect \( |d_i| \equiv 1. \)

Thus, \( |d_i| \) much bigger than one, indicates that observation \( i \) is contributing to a lack of fit. In this example, the largest absolute value is 1.2763, not much bigger than one.
Here is a graph of the deviance residuals versus the fitted values:

![Graph of deviance residuals versus fitted values]

Here is a graph of the deviance residuals versus month:

![Graph of deviance residuals versus month]

In neither case do I observe an obvious pattern.
Here is a Normal Q-Q plot of the deviance residuals:

![Normal Q-Q plot](image)

Other than the first point, the points seem to be approximately along a straight line, thus this data could be from a Normal. However, there is too little data to make a definite conclusion.\(^{259}\)

\(^{259}\) It turns out the standardized residuals do not seem to follow a Standard Normal. It turns out that for a Gamma Distribution rather than a Poisson Distribution, the standardized residuals seem to follow a Standard Normal. Thus a Gamma Distribution seems to be a better model for this data.
Assessing Model Stability

The actuary would like the GLM to be stable; in other words, the predictions of the model should not be overly sensitive to small changes in the data.

An observation is influential if it has a large effect on the fitted model. An outlier is an observation such that the corresponding fitted value is far from the observation.

**An influential observation is such that its removal from the data set causes a significant change to our modeled results.** An observation is influential when one or more of its predictor values are far from its mean and the observation is an outlier.

A common measure of influence is Cook’s distance. The larger the value of Cook’s distance, the more influential the observation.

The actuary should rerun the model excluding the most influential points to see their impact on the results. If this causes large changes in some of the parameter estimates, the actuary should consider for example whether to give these influential observations less weight.

Cross-validation, as discussed previously, can also be used to assess the stability of a GLM. For example, we can divide the data into ten parts. By combining these parts, we can create ten different subsets each of which contains 90% of the total data. We then fit the model to each of these ten subsets.

The results of the models fit to these different subsets of the data ideally should be similar. The amount by which these results vary is a measure of the stability of the model.

Bootstrapping via simulation can also be used to assess the stability of a GLM. The original data is randomly sampled with replacement to create a new set of data of the same size. One then fits the GLM to this new set of data. By repeating this procedure many times one can estimate the distribution of the parameter estimates of the GLM; we can estimate the mean, variance, confidence intervals, etc. “Many modelers prefer bootstrapped confidence intervals to the estimated confidence intervals produced by statistical software in GLM output.”

---

260 See Section 6.4 of Goldburd, Khare, and Tevet.
261 The syllabus reading gives no details on how Cook’s Distance is calculated. Computer software to fit GLMs will usual include Cook’s Distance as one of the possible outputs.
262 Values of Cook’s Distance greater than unity may require further investigation.
263 See An Introduction to Statistical Learning with Applications in R, by James, Witten, Hastie, and Tibshirani, not on the syllabus of this exam.
Scoring Models:

We have a rating plan or rating plans. We may not know what model if any that the plan(s) came from. We wish to evaluate a rating plan or compare two rating plans.

Methods that are discussed: Plots of Actual vs. Predicted, Simple Quantile Plots, Double Lift Charts, Loss Ratio Charts, the Gini Index, and ROC Curves.

In order for these techniques to be used, one only needs a database of historical observations plus the predictions from each of the competing models. The process of assigning predictions to individual records is called scoring.

Assessing Fit with Plots of Actual versus Predicted:

Create a plot of the actual target variable (on the y-axis) versus the predicted target variable (on the x-axis) for each model. If a model fits well, then the actual and predicted target variables should follow each other closely. Here are two examples:

Model 2 fits the data better than Model 1, as there is a much closer agreement between the actual and predicted target variables for Model 2 than there is for Model 1.

These plots should not use data that was used to fit or train the models. It is common to group the data, for example into percentiles. Often one will plot the graph on a log scale, as in the above examples.

---

264 See Section 7 of Goldburd, Khare, and Tevet.
265 One or more of the rating plans may be proprietary.
266 See Section 7.1 of Goldburd, Khare, and Tevet.
267 See Figure 17 of Goldburd, Khare, and Tevet.
Measuring Model Lift: 268

Lift refers to a model’s ability to prevent adverse selection, measuring the approximate “economic value” of the model. Economic value is produced by comparative advantage in avoidance of adverse selection; thus model lift is a relative concept, comparing two or more competing models, or a model and the current plan. Lift measures a model’s ability to charge each insured an actuarially fair rate, thereby minimizing the potential for adverse selection. Model lift should always be measured on holdout data, in other words not using data used to fit or build the model.

Simple Quantile Plots: 269

To create a quantile plot of a model.
1. Sort the dataset based on the model predicted loss cost from smallest to largest.
2. Group the data into quantiles with equal volumes of exposures.
3. Within each group, calculate the average predicted pure premium based on the model, and the average actual pure premium.
4. Plot for each group, the actual pure premium and the predicted pure premium.

One can create separate quantile plots for two models, for example the current rating plan and a proposed rating plan and compare them: 271

---

268 See Section 7.2 of Goldburd, Khare, and Tevet. Lift differs from goodness of fit measures.
269 See Section 7.2.1 of Goldburd, Khare, and Tevet.
270 For example: quintiles (5 buckets), deciles (10 buckets), or vigintiles (20 buckets).
271 See Figure 18 in Goldburd, Khare, and Tevet.
To compare the models use the following 3 criteria:

1. **Predictive accuracy**.
2. **Monotonicity.** The actual pure premium should increase.
3. **Vertical distance between the actuals in the first and last quantiles.**

“A large difference (also called “lift”) between the actual pure premium in the quantiles with the smallest and largest predicted loss costs indicates that the model is able to maximally distinguish the best and worst risks.”

The previous set of graphs can be used to compare the current and proposed model.

1. Predictive accuracy: the proposed model does a better job of predicting.
2. Monotonicity: the current plan has a reversal in the 6th decile, whereas the proposed model does better with no significant reversals.
3. Vertical distance between the first and last quantiles: The spread of actual loss costs for the current plan is 0.55 to 1.30. The spread of the proposed model is 0.40 to 1.60, which is larger and thus better.

Thus, by all three criteria, the proposed plan outperforms the current one.

---

272 Although small reversals are okay.
Double Lift Charts

A double lift chart directly compares two models A and B.

To create a double lift chart:

1. For each observation, calculate Sort Ratio = \( \frac{\text{Model A Predicted Loss Cost}}{\text{Model B Predicted Loss Cost}} \)
2. Sort the dataset based on the Sort Ratio, from smallest to largest.
3. Group the data.
4. For each group, calculate the pure premiums: predicted by Model A, predicted by Model B, and actual. Then divide the group average by the overall average.
5. For each group, plot the three relativities calculated in the step 4.

The first group contains those risks which Model A thinks are best relative to Model B, while the last group contains those risks which Model B thinks are best relative to Model A. The first and last groups contain those risks on which Models A and B disagree the most in percentage terms.

The “winning” model is the one that more closely matches the actual pure premiums. Here is an example of a double lift chart, comparing a current and proposed plan:

The proposed model more accurately predicts actual pure premium by decile than does the current rating plan. This is particularly clear when looking at the extreme groups on either end.

---

273 See Section 7.2.2 of Goldburd, Khare, and Tevet.
274 Thus a low sort ratio means that model A predicts a lower loss cost than does model B.
275 For example into 5 buckets (quintiles) or 10 buckets (deciles).
276 See Figure 19 in Goldburd, Khare, and Tevet.
“As an alternate representation of a double lift chart, one can plot two curves: the percent error for the model predictions and the percent error for the current loss costs, where percent error is calculated as $\frac{\text{Predicted Loss Cost}}{\text{Actual Loss Cost}} - 1$. In this case, the winning model is the one with the flatter line centered at $y = 0$, indicating that its predictions more closely match actual pure premium.”

**Loss Ratio Charts**

A loss ratio chart is similar to a simple quantile chart, except one works with loss ratios (with respect to the premiums for the current plan) rather than pure premiums.

To create a loss ratio chart:
1. Sort the dataset based on the model prediction.
2. Group the data into quantiles with equal volumes of exposures.
3. Within each group, calculate the actual loss ratio.

Here is an example:

![Loss Ratio Chart](image)

The proposed model is able to segment the data into lower and higher loss ratio buckets, indicating that the proposed model is better than the current model. “The advantage of loss ratio charts over quantile plots and double lift charts is that they are simple to understand and explain.”

---

277 See Section 7.2.3 of Goldburd, Khare, and Tevet.
278 See Figure 20 in Goldburd, Khare, and Tevet.
Lorenz Curves: 279

The Lorenz Curve is used to define the Gini Index, to be discussed subsequently.

Assume that the incomes in a country follow a distribution function $F(x)$. 280 Then $F(x)$ is the percentage of people with incomes less than $x$.

The income earned by such people is: $\int_0^x t \, f(t) \, dt = E[X \wedge x] - x \, S(x) = \int_0^x S(t) \, dt$.

The percentage of total income earned by such people is:

$$\frac{\int_0^x y \, f(y) \, dy}{E[X]} = \frac{E[X \wedge x] - x \, S(x)}{E[X]}.$$ 

Define $G(x) = \frac{\int_0^x y \, f(y) \, dy}{E[X]} = \frac{E[X \wedge x] - x \, S(x)}{E[X]}$. 281

For example, assume an Exponential Distribution. Then $F(x) = 1 - e^{-x/\theta}$.

$$G(x) = \frac{E[X \wedge x] - x \, S(x)}{E[X]} = \frac{\theta \left(1 - e^{-x/\theta}\right) - x \, e^{-x/\theta}}{\theta} = 1 - e^{-x/\theta} - \left(x/\theta\right) e^{-x/\theta}.$$ 

Let $t = F(x) = 1 - e^{-x/\theta}$. Therefore, $x/\theta = -\ln(1 - t)$. 282

Then, $G(t) = t - \{-\ln(1-t)\} \cdot (1-t) = t + (1-t) \ln(1-t)$.

---

279 You should not be responsible for any details of the mathematics of Lorenz curves.

280 Of course, the mathematics applies regardless of what is being modeled. The distribution of incomes is just the most common context.

281 This is not standard notation. I have just used $G$ to have some notation.

282 This is just the VaR (Value at Risk) formula for the Exponential Distribution.
Then we can graph $G$ as a function of $F$:

This curve is referred to as the Lorenz curve or the concentration curve.

Since $F(0) = 0 = G(0)$ and $F(\infty) = 1 = G(\infty)$, the Lorenz curve passes through the points $(0, 0)$ and $(1, 1)$. Usually one would also include in the graph the $45^\circ$ reference line connecting $(0, 0)$ and $(1, 1)$, called the line of equality, as shown below:
\[ G(t) = G[F(x)] = \int_{0}^{x} y f(y) \, dy. \]

\[ \frac{dG}{dt} = \frac{dG}{dx} / \frac{dF}{dx} = \frac{x f(x)}{E[X]} / f(x) = \frac{x}{E[X]} > 0. \]

\[ \frac{d^2G}{dt^2} = \frac{1}{E[X]} \frac{dx}{dx} / \frac{dF}{dx} = \frac{1}{E[X] f(x)} > 0. \]

Thus, in the above graph, as well as in general, the Lorenz curve is increasing and concave up. The Lorenz curve is below the 45° reference line, except at the endpoints when they are equal.

The vertical distance between the Lorenz curve and the 45° comparison line is: \( F - G \).

Thus, this vertical distance is a maximum when: \( 0 = \frac{dF}{dF} - \frac{dG}{dF} \).

\[ \Rightarrow \frac{dG}{dF} = 1. \Rightarrow \frac{x}{E[X]} = 1. \Rightarrow x = E[X]. \]

Thus the vertical distance between the Lorenz curve and the line of equality is a maximum at the mean income.

Exercise: If incomes follow an Exponential Distribution, what is this maximum vertical distance between the Lorenz curve and the line of equality?

[Solution: The maximum occurs when \( x = \theta \).

\( F(x) = 1 - e^{-x/\theta} \). From previously, \( G(x) = 1 - e^{-x/\theta} - (x/\theta) e^{-x/\theta}. \)

\( F - G = (x/\theta) e^{-x/\theta}. \) At \( x = \theta \), this is: \( e^{-1} = 0.3679 \).]
Exercise: Determine the form of the Lorenz Curve, if the distribution of incomes follows a Shifted Pareto Distribution, with \( \alpha > 1 \).

Solution: 

\[
F(x) = 1 - \left( \frac{\theta}{\theta + x} \right)^\alpha, \quad x > 0.
\]

\[
E[X] = \frac{\theta}{\alpha - 1}.
\]

\[
E[X \wedge x] = \frac{\theta}{\alpha - 1} \left[ 1 - \left( \frac{\theta}{\theta + x} \right)^{\alpha - 1} \right].
\]

\[
G(x) = \frac{E[X \wedge x] - x S(x)}{E[X]} = \frac{\theta}{\alpha - 1} \left[ 1 - \left( \frac{\theta}{\theta + x} \right)^{\alpha - 1} \right] - x S(x).
\]

Let \( t = F(x) = 1 - \left( \frac{\theta}{\theta + x} \right)^\alpha \Rightarrow \left( \frac{\theta}{\theta + x} \right)^\alpha = S(x) = 1 - t \). Also, \( x/\theta = (1 - t)^{-1/\alpha} - 1 \).

Therefore, 

\[
G(t) = 1 - (1 - t)^{(\alpha - 1)/\alpha} \left[ (1 - t)^{1/\alpha} - 1 \right] = t + \alpha - t\alpha - (1 - t)^{(1 - 1)/\alpha}, \quad 0 \leq t \leq 1.
\]

Comment: \( G(0) = \alpha - \alpha = 0 \). \( G(1) = 1 + \alpha - \alpha - 0 = 1 \).

Here is graph comparing the Lorenz curves for Shifted Pareto Distributions with \( \alpha = 2 \) and \( \alpha = 5 \):

---

\( ^{283} \) What Bahnemann calls a Shifted Pareto, Loss Models simply calls a Pareto.

\( ^{284} \) This is just the VaR (value at risk) formula for the Shifted Pareto Distribution.
The Shifted Pareto with $\alpha = 2$ has a heavier righthand tail than the Shifted Pareto with $\alpha = 5$. If incomes follow a Shifted Pareto with $\alpha = 2$, then there are more extremely high incomes compared to the mean, than if incomes follow a Shifted Pareto with $\alpha = 5$. In other words, if $\alpha = 2$, then income is more concentrated in the high income individuals than if $\alpha = 5$.\(^{285}\)

The Lorenz curve for $\alpha = 2$ is below that for $\alpha = 5$. In general, the lower curve corresponds to a higher concentration of income. In other words, a higher concentration of income corresponds to a smaller area under the Lorenz curve. Equivalently, a higher concentration of income corresponds to a larger area between the Lorenz curve and the 45° reference line.

Here is a Lorenz Curve for United States 2014 Household Income:\(^{286}\)

![Lorenz Curve](image)

The Gini index is calculated as twice the area between the Lorenz curve and the line of equality. In this case, the Gini index is 48.0%.

---

\(^{285}\) An Exponential Distribution has a lighter righthand tail than either Shifted Pareto. Thus if income followed an Exponential, it would less concentrated than if it followed any Shifted Pareto.

\(^{286}\) See Figure 21 of Goldburd, Khare, and Tevet.
Gini Index: The Gini Index comes up for example in economics, when looking at the distribution of incomes. A subsequent section will discuss how the Gini index can be used to evaluate a rating plan.

The Gini index is a measure of inequality. For example if all of the individuals in a group have the same income, then the Gini index is zero. As incomes of the individuals in a group became more and more unequal, the Gini index would increase towards a value of 1. The Gini index has found application in many different fields of study.

As discussed, for incomes, the Lorenz curve would graph percent of people versus percent of income. This correspondence between areas on the graph of the Lorenz curve the concentration of income is the idea behind the Gini index. Let us label the areas in the graph of a Lorenz Curve:

\[
\text{Gini Index} = \frac{\text{Area A}}{\text{Area A} + \text{Area B}}.
\]

However, Area A + Area B add up to a triangle with area 1/2.

Therefore, Gini Index = \[\frac{\text{Area A}}{\text{Area A} + \text{Area B}} = 2\text{A}\]

= twice the area between the Lorenz Curve and the line of equality = 1 - 2B.

\[287 \text{ See Section 7.2.4 of Goldburd, Khare, and Tevet. Also called the Gini Coefficient or coefficient of concentration.}\]
**Gini Index for Specific Distributions**

For the Exponential Distribution, the Lorenz curve was: \( G(t) = t + (1-t) \ln(1-t) \).

Thus, Area B = area under Lorenz curve = \( \int_0^1 t + (1-t) \ln(1-t) \, dt = 1/2 + \int_0^1 s \ln(s) \, ds \).

Applying integration by parts,
\[
\int_0^1 s \ln(s) \, ds = \left[ s^2/2 \ln(s) \right]_0^1 - \int_0^1 (s^2/2) (1/s) \, ds = 0 - 1/4 = -1/4.
\]

Thus Area B = 1/2 - 1/4 = 1/4.

Therefore, for the Exponential Distribution, the Gini Index is: \( 1 - (2)(1/4) = 1/2 \).

For the Uniform Distribution, the Gini Index is: 1/3.

For the Shifted Pareto Distribution, the Gini Index is: \( 1 / (2\alpha - 1) \), \( \alpha > 1 \).

We note that the Uniform with the lightest right-hand tail of the three has the smallest Gini index, while the Shifted Pareto with the heaviest right-hand tail of the three has the largest Gini index. Among Shifted Pareto Distributions, the smaller alpha, the heavier the right-hand tail, and the larger the Gini index.\(^{289}\)

The more concentrated the income is among the higher earners, the larger the Gini index.

For the Classical (Single Parameter) Pareto Distribution, the Gini Index is: \( 1 / (2\alpha - 1) \), \( \alpha > 1 \).

For the LogNormal Distribution, the Gini Index is: \( 2\Phi[\sigma/\sqrt{2}] - 1 \).

For the Gamma Distribution, the Gini Index is: \( 1 - 2 \beta(\alpha+1, \alpha; 1/2) \).

---

\(^{288}\) Not on the syllabus.

\(^{289}\) As alpha approaches one, the Gini coefficient approaches one.
Gini Index and Rating Plans:  

The Gini index can also be used to measure the lift of an insurance rating plan by quantifying its ability to segment the population into the best and worst risks. Assume we have a rating plan. Ideally we would want the model to identify those insureds with higher expected pure premiums.

The Lorenz curve for the rating plan is determined as follows:

1. Sort the dataset based on the model predicted loss cost.
2. On the x-axis, plot the cumulative percentage of exposures.
3. On the y-axis, plot the cumulative percentage of actual losses.
Draw a 45-degree line connecting (0, 0) and (1, 1), called the line of equality.

Here is an example:

This model identified 60% of exposures which contribute only 20% of the total losses. The Gini index is twice the area between the Lorenz curve and the line of equality, in this case 56.1%. The higher the Gini index, the better the model is at identifying risk differences.

---

290 See Section 7.2.4 of Goldburd, Khare, and Tevet.
291 This should be done on a dataset not used to develop the rating plan.
292 See Figure 21 of Goldburd, Khare, and Tevet.
293 “Note that a Gini index does not quantify the profitability of a particular rating plan, but it does quantify the ability of the rating plan to differentiate the best and worst risks. Assuming that an insurer has pricing and/or underwriting flexibility, this will lead to increased profitability.”
An Example of the Gini Index and an Insurance Rating Plan: 294

We have four classes each with an equal number of exposures, and the result of fitting a GLM. 295

We have already sorted the classes according to the pure premiums predicted by the GLM. 296

<table>
<thead>
<tr>
<th>Class</th>
<th>Predicted Pure Premium</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
</tr>
</tbody>
</table>

Ignoring here any misestimating of the overall rate level, the observed pure premiums would differ from the predicted pure premiums for two reasons: 297 298

1. Imperfection of the GLM, in other words modeling error.
2. Random fluctuation, in other words process variance. 299

Let us assume the following Actual Pure Premiums: 300 301

<table>
<thead>
<tr>
<th>Class</th>
<th>Actual P.P.</th>
<th>Cumulative Losses</th>
<th>% of Losses</th>
<th>% Expos</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>160</td>
<td>160</td>
<td>16%</td>
<td>25%</td>
</tr>
<tr>
<td>2</td>
<td>240</td>
<td>400</td>
<td>40%</td>
<td>50%</td>
</tr>
<tr>
<td>3</td>
<td>260</td>
<td>660</td>
<td>66%</td>
<td>75%</td>
</tr>
<tr>
<td>4</td>
<td>340</td>
<td>1000</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Thus for the Lorenz curve we plot the points: (0, 0), (25, 16), (50, 40), (75, 66), (100, 100).

294 See 8, 11/16, Q.5.
295 I have chosen a one-dimensional example with only four levels solely for illustrative simplicity. Most GLMs would include more than one risk characteristic, and some characteristics would have more than four levels. Also the exposures for each level would usually not all be equal.
296 In a practical application we would have hundreds if not thousands of different cells consisting of risks with all of the same characteristics and thus the same predicted pure premium.
297 We are using the GLM to predict class relativities rather than the overall rate level. In some cases, the GLM output will automatically balance to the observed.
298 Each class is not perfectly homogenous; it may be possible to refine the given classes to produce more homogeneous classes. Of course, if the classes are made too small, we would have issues with credibility.
299 The more data in a class, the less subject to random fluctuation would be the average observed pure premium for that class.
300 These observed pure premiums are from a dataset similar to the one to which the GLM was fit.
301 Assuming solely for simplicity one exposure per class.
It is possible to calculate the area between the above Lorenz Curve and the Line of Equality, by dividing the area in triangles.\textsuperscript{302} \textsuperscript{303} This area turns out to be 0.07.\textsuperscript{304} Thus the Gini Index is twice that or 14\%\textsuperscript{305}.\n
\begin{itemize}
  \item \textsuperscript{302} You will not be asked to do so on your exam!
  \item \textsuperscript{303} The six triangles I used were: \{(0,0), (25,16), (25,25)}, {{(25,16), (25,25), (50,40)}}, ...
  \item \textsuperscript{304} Remembering that for example the value shown as 25 is actually 25\% = 0.25.
  \item \textsuperscript{305} The higher the Gini index, the better the model is at identifying risk differences.

A more complicated model is likely to do better than this very simple class plan.
Solely for illustrative purposes, let us investigate the Gini Index if instead the actual pure premiums exactly matched the predicted pure premiums for each class.\footnote{While this is will not occur in practice, this is the best possible result for this simple plan with only four classes.}\footnote{Assuming solely for simplicity one exposure per class.}

<table>
<thead>
<tr>
<th>Class</th>
<th>Actual P.P.</th>
<th>Cumulative Losses</th>
<th>% of Losses</th>
<th>% Expos</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
<td>10%</td>
<td>25%</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>300</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>600</td>
<td>60%</td>
<td>75%</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>1000</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Thus for the Lorenz curve we plot the points: (0, 0), (25, 10), (50, 30), (75, 60), (100, 100).

The area between the above Lorenz Curve and the Line of Equality, turns out to be 0.125. Thus the Gini Index is twice that or 25%, higher than previously.
Understanding & Validating a Model:\textsuperscript{308}

Model Lift
How well does the model differentiate between best and worst risks?
Does the model help prevent adverse selection?
Is the model better than the current rating plan?

Simple Quantile plots:
Illustrate how well the model helps prevent adverse selection.
Double lift charts:
Compare competing models or compare new model against current rating plan.
Gini Index:
Summarizes model lift into one number.
Loss ratio charts:
Puts lift in a context most people in the insurance industry can understand.

Goodness of Fit
What kind of model statistics are available, and how do you interpret them?
What kind of residual plots should you consider, and how do you interpret them?
What are some considerations regarding actual versus predicted plots?

Internal Stability
How well does the model perform on other data?
How will the model perform over time?
How reliable are the model's parameter estimates?

\textsuperscript{308} “And The Winner Is…? How to Pick a Better Model,” 2015 CAS RPM Seminar, by Hernan L. Medina.
ROC Curves:309

Receiver Operating Characteristic (ROC) Curves can be used to compare models that use the Bernoulli or Binomial Distribution.310

The first step is to pick a threshold. For example, if the discrimination threshold were 8%, then we look at all cells with the fitted probability of an event > 8%, in other words \( q_i > 8\% \). Then we count up the number of times there was an event when an event was predicted. For example, there might be 3740 such true positives. Assume that there 4625 total events. Then the “sensitivity” is the ratio: \( 3740/4625 = 0.81 \).

In general, above a given threshold, the sensitivity is the portion of the time that an event was predicted by the model out of all the times there is an event = \( \frac{\text{true positives}}{\text{total times there is an event}} \). Sensitivity is the rate of true positives.312

All other things being equal, higher sensitivity is good.

Then we look at all cells with the fitted probability of an event \( \leq 8\% \), in other words \( q_i \leq 8\% \).

For example, there might be 54,196 such policies without an event. Assume there are a total of 63,232 policies without an event. Then the “specificity” is the ratio: \( 54,196/63,232 = 0.85 \).

Below a given threshold, the specificity is the portion of the time that an event was not predicted by the model out of all of the times these is not an event = \( \frac{\text{true negatives}}{\text{total times there is not an event}} \).313 All other things being equal, higher specificity is good.

Specificity is the rate of true negatives. 1 - specificity is the rate of false positives.

For this example, for a threshold of 8%, we can display the information in a confusion matrix:314

<table>
<thead>
<tr>
<th>Discrimination Threshold: 8%</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>Event</td>
</tr>
<tr>
<td>Event</td>
<td>3740</td>
</tr>
<tr>
<td>No Event</td>
<td>9036</td>
</tr>
<tr>
<td>Total</td>
<td>55,080</td>
</tr>
</tbody>
</table>

309 See Section 7.3.1 in Goldburd, Khare, and Tevet.
310 ROC analysis was originally developed during World War II for the analysis of radar images.
311 The event could be a claim, a policy renewal, etc.
312 If one has a model to predict the probability of a claim being fraudulent, then for a given threshold:
   Sensitivity = (correct predictions of fraud) / (total number of fraudulent claims).
313 If one has a model to predict the probability of a claim being fraudulent, then for a given threshold:
   Specificity = (correct predictions of no fraud) / (total number of non-fraudulent claims).
314 See Table 13 in Goldburd, Khare, and Tevet.
For a certain threshold, the general form of a confusion matrix:

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>true positive</td>
</tr>
<tr>
<td>No Event</td>
<td>false positive</td>
</tr>
</tbody>
</table>

A confusion matrix is similar to a table from hypothesis testing, where the null hypothesis is no event:  

<table>
<thead>
<tr>
<th>Decision</th>
<th>Reject H$_0$</th>
<th>Do not reject H$_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H$_1$ is True</td>
<td>Correct</td>
<td>Type II Error</td>
</tr>
<tr>
<td>H$_0$ is True</td>
<td>Type I Error</td>
<td>Correct</td>
</tr>
</tbody>
</table>

The false negatives are analogous to making a Type II Error. The false positives are analogous to making a Type I Error.

<table>
<thead>
<tr>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>Event</td>
</tr>
<tr>
<td>Event</td>
<td>3740</td>
</tr>
<tr>
<td>No Event</td>
<td>9036</td>
</tr>
<tr>
<td>Total</td>
<td>12,776</td>
</tr>
</tbody>
</table>

For the 8% threshold, the specificity was:  

\[
\text{specificity} = \frac{\text{true negatives}}{\text{total times there is not an event}} = \frac{54,196}{63,232} = 85\%.
\]

1 - specificity =  

\[
\text{1 - specificity} = \frac{\text{false positives}}{\text{total times there is not an event}} = \frac{9036}{63,232} = 15\%.
\]

1 - specificity is analogous to: 

\[\text{chance of making a Type Error I} = \text{significance level of a statistical test}.\]

For the 8% threshold, the sensitivity was:  

\[
\text{sensitivity} = \frac{\text{true positives}}{\text{total times there is an event}} = \frac{3740}{4625} = 81\%.
\]

Sensitivity is analogous to: \(1 - \text{chance of making a Type Error II} = \text{power of a statistical test}\).

In the ROC Curve we plot the point: \((1 - 0.85, 0.81) = (0.15, 0.81)\).

In general, the ROC curve consists of plotting for various thresholds: 

\((1 - \text{specificity}, \text{sensitivity})\).

In addition, there is a 45% comparison line, the line of equality, from \((0, 0)\) to \((1, 1)\).

While the analogy to hypothesis testing may help your understanding, it should not be tested on your exam.
Here is an example of an ROC curve:

A perfect model would be at (0, 1) in the upper lefthand corner; sensitivity = 1 and specificity = 1. The closer the model curve gets to the upper lefthand corner the better.

The comparison line (line of equality) indicates a model with sensitivity = 1 - specificity, which can be achieved by just flipping a coin to decide your prediction. Thus such models have no predictive value. The closer the model curve gets to the 45 degree comparison line (line of equality), the worse the model.

The comparison line has area 1/2 below it. The larger the area under the model curve, the better it is.

**AUROC is the area under the ROC curve; the larger AUROC the better the model.**

---

316 I just drew my diagram; it does not come from an actual fitted GLM.
317 The AUROC is equal to: (0.5) (normalized Gini) + 0.5, where the normalized Gini is the ratio of the model's Gini index to the Gini index of the hypothetical "perfect" model (where each record's prediction equals its actual value). Note that the prefect model will not have a Gini index of one; it's Gini index depends on the homogeneity of the risks and the randomness of the loss process.
For an example of modeling fraud on claims, the syllabus reading has the following ROC Curve:

This ROC has an area under ROC (AUROC) of 0.857. 

We can see how as one changes the threshold from 0.5 to 0.25, the sensitivity increases, but at the cost a lower specificity. In other words, the rate of true positives increases at the cost of also increasing the rate of false positives.

The selection of the discrimination threshold involves a trade-off: a lower threshold will result in more true positives and fewer false negatives than a higher threshold, but at the cost of more false positives and fewer true negatives.

---

318 See Figure 22 in Goldburd, Khare, and Tevet.
319 The perfect model would have an AUROC of 1.
320 “The ROC curve allows us to select a threshold we are comfortable with after weighing the benefits of true positives against the cost of false positives. Different thresholds may be chosen for different claim conditions; for example, we may choose a lower threshold for a large claim where the cost of undetected fraud is higher. Determination of the optimal threshold is typically a business decision that is out of the scope of the modeling phase.”
A Medical Example of ROC

Let us assume we have a medical test for a disease which results in a numerical score. The lower the score on this test the more likely that the individual has this disease. Assume the following data:

<table>
<thead>
<tr>
<th>Score on Medical Test</th>
<th>Number with Disease</th>
<th>Number without Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 or less</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>5.1 to 7</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>7.1 to 9</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>9 or more</td>
<td>3</td>
<td>39</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>93</td>
</tr>
</tbody>
</table>

We can pick a threshold to use with this test; if the test score is less than or equal to the chosen threshold this indicates that the individual has the disease.

For example, assume a threshold of 5. Then 18 individuals are correctly identified as diseased, and 1 is incorrectly identified as diseased. There are 18 true positives. There is one false positive. 14 individuals who are diseased are incorrectly identified as being without disease. There are 14 false negatives. 92 individuals who are not diseased are correctly identified as being without disease.

We can think of sensitivity as the rate of true positives of a medical test for a disease as a portion of positives. The rate of true positives out of all diseased is: $18/32 = 0.56$.

We can think of specificity as the rate of individuals that the test indicates do not have the disease out of those without the disease. The rate of negatives out of those without the disease: $92/93 = 0.99$. One minus the specificity, 1%, is the rate of false positives out of those without the disease.

The confusion matrix is:

<table>
<thead>
<tr>
<th></th>
<th>Actual Disease</th>
<th>No Disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>18</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>No Disease</td>
<td>14</td>
<td>92</td>
<td>106</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>93</td>
<td>125</td>
</tr>
</tbody>
</table>

321 http://gim.unmc.edu/dxtests/ROC1.htm
322 While a low test score indicates the presence of the disease in this example, it could have been the reverse.
323 Sensitivity is analogous to the probability of rejecting the null hypothesis (healthy) when it is false, which is the power of the test.
324 One minus specificity is analogous to the probability of rejecting the null hypothesis (healthy) when it is true, which is the significance level of the test.
Exercise: What are the sensitivity and specificity if one instead uses a threshold of 7?

[Solution: 25 people have positive tests out of 32 with the disease.⇒ sensitivity is: \( \frac{25}{32} = 0.78 \).
75 people have negative tests out of 93 who are healthy. ⇒ specificity is: \( \frac{75}{93} = 0.81 \).
Comment: With a higher threshold the sensitivity is higher but the specificity is lower.
There is a tradeoff between a high sensitivity and a high specificity.]

Exercise: What are the sensitivity and specificity if one instead uses a threshold of 9?

[Solution: 29 people have positive tests, out of 32 with the disease.⇒ sensitivity is: \( \frac{29}{32} = 0.91 \).
39 people have negative tests out of 93 who are healthy. ⇒ specificity is: \( \frac{39}{93} = 0.42 \).]

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>1 - Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.56</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>0.78</td>
<td>0.81</td>
<td>0.19</td>
</tr>
<tr>
<td>9</td>
<td>0.91</td>
<td>0.42</td>
<td>0.58</td>
</tr>
</tbody>
</table>

The corresponding ROC curve, where I have not connected the dots:\(^{325}\)

---

\(^{325}\) The area under the curve measures discrimination, that is, the ability of the test to correctly classify those with and without the disease.
Other Topics

The syllabus reading discusses three additional topics:
• Why you probably should not model coverage options with GLMs.
• Why territories are not a good fit for the GLM framework.
• Ensembling.

Coverage Options

Insureds can choose coverage options such as deductible amount or limit of liability. There are corresponding deductible credits or increased limits factors. You probably should not model the rating factors for coverage options with GLMs.

For example, a GLM might indicate that one should charge more for a higher deductible. There may be something systematic about insureds with higher deductibles that may make them a worse risk relative to others in their class. In which case, the coefficients estimated by the GLM are reflecting some of this increased risk due to antiselection effects.

To the extent that the factor indicated by the GLM differs from the pure effect on loss potential, it will affect the way insureds choose coverage options in the future. Thus, the selection dynamic will change and the past results would not be expected to be replicated for new policies.

Thus factors for coverage options should be estimated outside the GLM, using traditional actuarial techniques. The resulting factors should then be included in the GLM as an offset, as has been discussed previously.

Examples of GLM Output:

On my webpage I have posted a file in which I discuss some examples of GLM output taken from Chapter 10 and Appendix F of Basic Ratemaking, on the syllabus of the Basic Ratemaking exam. While these examples should not be tested directly on your exam, you may find it helpful to briefly review them.

---

326 See Section 8 of Goldburd, Khare, and Tevet.
327 See Section 8.1 of Goldburd, Khare, and Tevet.
328 These can be distinguished from characteristics of the insured.
329 In general, the insured should pay less for less coverage and more for more coverage.
330 “The choice of high deductible may be the result of a high risk appetite on the part of an insured, which would manifest in other areas as well. Alternately, the underwriter, recognizing an insured as a higher risk, may have required the policy to be written at a higher deductible.”
331 This is the recommendation of Goldburd, Khare, and Tevet.
332 Also included are some related problems from past Basic Ratemaking exams.
Territory Modeling

Territories are not a good fit for the GLM framework.

There may have hundreds of territories, which requires many levels in the GLM. Therefore, the authors recommend the use of other techniques, such as spatial smoothing, to model territories.

One should include the territory relativities produced by the separate model as an offset in the GLM used to determine classification relativities. Similarly, one should include classification relativities produced by the GLM as an offset in the separate model used to determine territory relativities.

Ideally this should be an iterative process.

Ensembling

Two (or more) teams model the same item; they build separate models working independently. The models are evaluated and found to be approximately equal in quality.

Combining the answers from both models is likely to perform better than either individually. A model that combines information from two or more models is called an ensemble model.

A simple means of ensembling is to average the separate model predictions. “Predictive models each have their strengths and weaknesses. Averaged together, they can balance each other out, and the gain in performance can be significant.”

---

333 See Section 8.2 of Goldburd, Khare, and Tevet. I believe the authors are discussing determining territory relativities rather than constructing territories from smaller geographical units such as zipcode. However, they may be discussing doing both together.

334 The authors do not discuss any details of spatial smoothing or other techniques.

335 In determining territories one should adjust the pure premiums for a zipcode by its average class rating factor. Chapter 11 of Basic Ratemaking by Werner and Modlin have a discussion of determining territories.

336 If they being updated at the same time, both models should be run, using the other as an offset, until they reach an acceptable level of convergence.

337 See Section 8.3 of Goldburd, Khare, and Tevet.

338 Of course it is costly to have two teams build two separate models.

“Done properly, though, ensembles can be quite powerful; if resources permit, it may be worth it.”

339 The authors do not discuss more complicated methods of ensembling.
A More Realistic and Complex Example:

Consider the following data on claim severity for personal auto insurance:

<table>
<thead>
<tr>
<th>Observation</th>
<th>Age Group</th>
<th>Vehicle-Use</th>
<th>Severity</th>
<th>Claim Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17–20</td>
<td>Pleasure</td>
<td>250.48</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>17–20</td>
<td>Drive to Work &lt; 10 miles</td>
<td>274.78</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>17–20</td>
<td>Drive to Work &gt; 10 miles</td>
<td>244.52</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>17–20</td>
<td>Business</td>
<td>797.80</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>21–24</td>
<td>Pleasure</td>
<td>213.71</td>
<td>63</td>
</tr>
<tr>
<td>6</td>
<td>21–24</td>
<td>Drive to Work &lt; 10 miles</td>
<td>298.60</td>
<td>171</td>
</tr>
<tr>
<td>7</td>
<td>21–24</td>
<td>Drive to Work &gt; 10 miles</td>
<td>298.13</td>
<td>92</td>
</tr>
<tr>
<td>8</td>
<td>21–24</td>
<td>Business</td>
<td>362.23</td>
<td>44</td>
</tr>
<tr>
<td>9</td>
<td>25–29</td>
<td>Pleasure</td>
<td>250.57</td>
<td>140</td>
</tr>
<tr>
<td>10</td>
<td>25–29</td>
<td>Drive to Work &lt; 10 miles</td>
<td>248.56</td>
<td>343</td>
</tr>
<tr>
<td>11</td>
<td>25–29</td>
<td>Drive to Work &gt; 10 miles</td>
<td>297.90</td>
<td>318</td>
</tr>
<tr>
<td>12</td>
<td>25–29</td>
<td>Business</td>
<td>342.31</td>
<td>129</td>
</tr>
<tr>
<td>13</td>
<td>30–34</td>
<td>Pleasure</td>
<td>229.09</td>
<td>123</td>
</tr>
<tr>
<td>14</td>
<td>30–34</td>
<td>Drive to Work &lt; 10 miles</td>
<td>228.48</td>
<td>448</td>
</tr>
<tr>
<td>15</td>
<td>30–34</td>
<td>Drive to Work &gt; 10 miles</td>
<td>293.87</td>
<td>361</td>
</tr>
<tr>
<td>16</td>
<td>30–34</td>
<td>Business</td>
<td>367.46</td>
<td>169</td>
</tr>
<tr>
<td>17</td>
<td>35–39</td>
<td>Pleasure</td>
<td>153.62</td>
<td>151</td>
</tr>
<tr>
<td>18</td>
<td>35–39</td>
<td>Drive to Work &lt; 10 miles</td>
<td>201.67</td>
<td>479</td>
</tr>
<tr>
<td>19</td>
<td>35–39</td>
<td>Drive to Work &gt; 10 miles</td>
<td>238.21</td>
<td>381</td>
</tr>
<tr>
<td>20</td>
<td>35–39</td>
<td>Business</td>
<td>256.21</td>
<td>166</td>
</tr>
<tr>
<td>21</td>
<td>40–49</td>
<td>Pleasure</td>
<td>208.59</td>
<td>245</td>
</tr>
<tr>
<td>22</td>
<td>40–49</td>
<td>Drive to Work &lt; 10 miles</td>
<td>202.80</td>
<td>970</td>
</tr>
<tr>
<td>23</td>
<td>40–49</td>
<td>Drive to Work &gt; 10 miles</td>
<td>236.06</td>
<td>719</td>
</tr>
<tr>
<td>24</td>
<td>40–49</td>
<td>Business</td>
<td>352.49</td>
<td>304</td>
</tr>
<tr>
<td>25</td>
<td>50–59</td>
<td>Pleasure</td>
<td>207.57</td>
<td>266</td>
</tr>
<tr>
<td>26</td>
<td>50–59</td>
<td>Drive to Work &lt; 10 miles</td>
<td>202.67</td>
<td>859</td>
</tr>
<tr>
<td>27</td>
<td>50–59</td>
<td>Drive to Work &gt; 10 miles</td>
<td>253.63</td>
<td>504</td>
</tr>
<tr>
<td>28</td>
<td>50–59</td>
<td>Business</td>
<td>340.56</td>
<td>162</td>
</tr>
<tr>
<td>29</td>
<td>60+</td>
<td>Pleasure</td>
<td>192.00</td>
<td>260</td>
</tr>
<tr>
<td>30</td>
<td>60+</td>
<td>Drive to Work &lt; 10 miles</td>
<td>196.33</td>
<td>578</td>
</tr>
<tr>
<td>31</td>
<td>60+</td>
<td>Drive to Work &gt; 10 miles</td>
<td>259.79</td>
<td>312</td>
</tr>
<tr>
<td>32</td>
<td>60+</td>
<td>Business</td>
<td>342.58</td>
<td>96</td>
</tr>
</tbody>
</table>

Data taken from Exhibit 1 of “A Systematic Relationship Between Minimum Bias and Generalized Linear Models,” by Stephen J. Mildenhall, PCAS 1999, not on the syllabus.
There are 8 age categories and 4 vehicle use types. Thus there are a large number of ways to set up a GLM. I will make age 40-49 and drive to work less than 10 miles as the base levels.

I will use the following definitions of variables:

- $X_0$ corresponds to the base levels.
- $X_1$ is one if 17-20 years old and zero otherwise.
- $X_2$ is one if 21-24 years old and zero otherwise.
- $X_3$ is one if 25-29 years old and zero otherwise.
- $X_4$ is one if 30-34 years old and zero otherwise.
- $X_5$ is one if 35-39 years old and zero otherwise.
- $X_6$ is one if 50-59 years old and zero otherwise.
- $X_7$ is one if 60+ years old and zero otherwise.
- $X_8$ is one if Pleasure Use and zero otherwise.
- $X_9$ is one if Drive to Work > 10 and zero otherwise.
- $X_{10}$ is one if Business Use and zero otherwise.

A Gamma Distribution with an identity link function was fit to these data:  

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Fitted Value</th>
<th>Standard Error</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>203.522</td>
<td>6.54517</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>62.9056</td>
<td>37.0291</td>
<td>8.9%</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>66.1851</td>
<td>19.4111</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>46.1676</td>
<td>12.5584</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>33.2979</td>
<td>11.3777</td>
<td>0.3%</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-15.289</td>
<td>9.57527</td>
<td>11.0%</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>3.57547</td>
<td>8.79087</td>
<td>68.4%</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-1.84956</td>
<td>9.5907</td>
<td>84.7%</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>-8.63574</td>
<td>8.22596</td>
<td>29.4%</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>45.1086</td>
<td>7.43089</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>122.802</td>
<td>13.4003</td>
<td>0</td>
</tr>
</tbody>
</table>

The fitted severities are: 257.79, 266.43, 311.54, 389.23, 261.07, 269.70, 314.82, 392.51, 241.05, 249.69, 294.80, 372.49, 228.19, 236.82, 281.93, 359.62, 179.60, 188.23, 233.34, 311.04, 194.89, 203.52, 248.63, 326.32, 198.46, 207.10, 252.21, 329.90, 193.04, 201.67, 246.78, 324.47.
Based on their large p-values, $\beta_5$, $\beta_6$, $\beta_7$, and $\beta_8$ are not significantly different than zero. Let us test a model in which we eliminate the corresponding variables. The reduced model will have:
Age 35-39 combined with 40-49.
Age 50-60 combined with 60+.
Pleasure use combined with Drive to Work < 10 miles.

Another GLM with Gamma Distribution with an identity link function was fit to these data. The deviance for the original model with more variables is 31.2438. The deviance for the new model with less variables is 37.0310.

We have two nested models. GLM 1 is a special case of GLM 2. Then the test statistic (asymptotically) follows an F-Distribution with numbers of degrees of freedom equal to: $\nu_1 =$ the difference in number of parameters = 3, and $\nu_2 =$ number of degrees of freedom for the more simpler model $=$ (number of observations) - (number of parameters) = 32 - 7 = 25.

$\hat{\phi}_S =$ estimated dispersion parameter for the smaller (simpler) model
$= \frac{D_S}{\nu_S} = \frac{37.0310}{25} = 1.481$.345

The test statistic is:
$$\frac{D_S - D_B}{\left(\frac{\text{number of added parameters}}{\hat{\phi}_S}\right)} = \frac{(37.0310 - 31.2438)/3}{1.481} = 1.303.$$344

Using a computer, the p-value is 29.5%. Thus we do not reject the null hypothesis of using the simpler model with fewer parameters.346

---

342 The fitted parameters are: 196.36, 67.41, 71.78, 50.88, 38.42, 6.13, 47.01, 125.74.
343 The fitted severities are: 263.77, 310.77, 389.51, 268.14, 315.15, 393.88, 247.24, 294.248, 372.98, 234.78, 281.79, 360.52, 196.36, 243.37, 322.10, 202.49, 202.49, 249.49, 328.23.
344 A computer was used to fit both models and to calculate the deviances.
345 The syllabus reading does not discuss how to estimate $\phi$; this is one way.
346 One could now compare additional models with different subsets of the original variables. One could also fit models using different distributional forms and/or link functions.
Personal lines rates in the United Kingdom have long been based on GLMs. One important aspect to using GLMs is to find relevant variables. Here is a list of some rating variables that are used for Homeowners Insurance.

Postal code (so geodemographic and geophysical factors can be derived)
Amount of insurance
Number of rooms / bedrooms
Wall type
Roof type
State of repair
Extensions
Ownership status (rent/own)
Occupancy in day
Neighborhood watch scheme
Approved locks, alarms, smoke detectors
Deductibles
Endorsements purchased (e.g. riders for jewelry, oriental rugs)
How long held insurance / when last claimed

Policyholder details:
• Age
• Sex
• Marital status
• Number of children
• Occupation
• Residency
• Criminal convictions
• Claims in past 2 or past 5 years

Smokers present in house
Non-family members sharing house
Length of time living at property
Use (principal residence / secondary residence / business / rented)
Coverage selected (buildings/contents/both)
Source of business (e.g. agent, internet, etc.)

---

348 Geodemographics are the average characteristics in an area. Examples are: population density, length of homeownership, average age of residents, and average family income. Geophysical factors can include soil type, and weather data such as the maximum wind speed, the average rainfall, and the average snowfall.
**Homeowners Perils:**

There can be advantages to modeling the different homeowners perils separately.\(^{349}\) One can either model pure premium or separately model frequency and severity.

Some variables may have different effects on different perils. For example, increased population density may be related to an increased frequency for theft claims while being related to a decreased frequency of fire claims.

Some variables may have a significant effect on one peril but not another. For example, more children in the house may be related to an increased frequency of liability while being unrelated to the frequency for wind.

Here is an example of data by peril for the United States.

<table>
<thead>
<tr>
<th>Peril</th>
<th>Frequency (in percent)</th>
<th>Median Claim Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fire</td>
<td>0.310</td>
<td>4,152</td>
</tr>
<tr>
<td>Lightning</td>
<td>0.527</td>
<td>899</td>
</tr>
<tr>
<td>Wind</td>
<td>1.226</td>
<td>1,315</td>
</tr>
<tr>
<td>Hail</td>
<td>0.491</td>
<td>4,484</td>
</tr>
<tr>
<td>Water-Weather Related</td>
<td>0.491</td>
<td>1,481</td>
</tr>
<tr>
<td>Water-NonWeather(^{377})</td>
<td>1.332</td>
<td>2,167</td>
</tr>
<tr>
<td>Liability</td>
<td>0.187</td>
<td>1,000</td>
</tr>
<tr>
<td>Other</td>
<td>0.464</td>
<td>875</td>
</tr>
<tr>
<td>Theft-Vandalism</td>
<td>0.812</td>
<td>1,119</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5.889</strong></td>
<td><strong>1,661</strong></td>
</tr>
</tbody>
</table>

The percent of losses expected by peril varies considerably by geographical location. For example, the expected percent from wind (from hurricanes and other storms) is higher than average on the coast of Florida. For example, the expected percent from theft is higher than average in the center of a large city.

Recently, homeowners insurers have begun to implement rating plans that have separate base rates for each major peril covered and the individual rating variable relativities are applied to the applicable base rate (e.g., burglar alarm discount applies to the theft base rate only).

---

\(^{349}\) See for example, “Predictive Modeling of Multi-Peril Homeowners Insurance,” by Edward W. Frees, Glenn Meyers, and A. David Cummins, in Variance Volume 6 / Issue 1. They show that the perils are not independent.
Problems:

3.1. (1.5 points)
Five Generalized Linear Models have been fit to the same set of 50 observations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Fitted Parameters</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>335.8</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>331.9</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>325.2</td>
</tr>
<tr>
<td>D</td>
<td>12</td>
<td>321.4</td>
</tr>
<tr>
<td>E</td>
<td>14</td>
<td>317.0</td>
</tr>
</tbody>
</table>

Which model has the best AIC (Akaike Information Criterion)?

3.2. (0.5 points) Briefly discuss how to pick the base level of a categorical variable.

*3.3.* (1 point) When a log link is used, it is usually appropriate to take the natural logs of continuous predictors before including them in the model, rather than placing them in the model in their original forms. Discuss why.

3.4. (1.5 points) Fully discuss the use of weights in GLMs.

3.5. (0.5 points) Briefly discuss a primary strength of GLMs versus univariate analyses.

*3.6.* (0.5 points) A continuous predictor $x_1$ has a coefficient of $\beta_1 = 0.4$ in a logistic model. For a unit increase in $x_1$, what is the estimated change in the odds?

3.7. (1 point) Compare and contrast the Poisson and the Negative Binomial Distributions.

3.8. (0.5 points) With respect to GLMs, briefly discuss aliasing.

*3.9.* (0.5 points) List two limitations of GLMs.

3.10. (1 point) One possible fix for nonlinearity in a continuous variable is not to model it as continuous at all; rather, a new categorical variable is created where levels are defined as intervals over the range of the original variable. Briefly discuss two drawbacks to this approach.

3.11. (1.5 points) A GLM has been fit using a Poisson Distribution with $\hat{\beta}_1 = 0.02085$ with standard error 0.00120. Using instead an overdispersed Poisson the estimate of $\phi$ is 7.9435. For this second model, determine a 95% confidence interval for $\beta_1$.

3.12. (1 point) Discuss the Tweedie Distribution.
3.13. (1 point) You are given a double lift chart, sorted by ratio of the model prediction over the current plan prediction. Discuss the lift of the proposed model compared to the current plan.

3.14. (1 point) The flexibility afforded by the ability to use a link function is a good thing because it gives us more options in specifying a model, thereby providing greater opportunity to construct a model that best reflects reality. However, when using GLMs to produce insurance rating plans, an added benefit is obtained when the link function is specified to be the natural log function. Briefly discuss this added benefit.

3.15. (1 point) A logistic regression has been fit to some data. For a certain threshold:

<table>
<thead>
<tr>
<th>Predicted Claims</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Claim</td>
<td>No</td>
<td>6000</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>300</td>
<td>700</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>6300</td>
<td>2700</td>
</tr>
</tbody>
</table>

What point would be plotted in the ROC curve?

3.16. (2 points) List and briefly discuss four components of a predictive modeling project.

3.17. (1.5 points)
(a) (0.5 points) Define the partial residuals.
(b) (1 point) Discuss partial residual plots.
3.18. (0.5 points) Briefly contrast the following two GLMs:
\[ \mu = \exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2]. \]
\[ \mu = \exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2]. \]

3.19. (1 point) Any data set of sufficient size is likely to have errors. Briefly discuss two of the steps that should always be taken to attempt to catch and remedy some of the more common errors that can occur.

3.20. (0.5 points) List two types of Exploratory Data Analysis (EDA) plots and their purposes.

3.21.* (1 point) Discuss some reasons to use frequency and severity models rather than a pure premium model.

3.22. (1.5 points) Fully discuss the use of an offset term in GLMs.

3.23. (0.5 points) Discuss the following graph of Cook’s Distance for 26 observations:

3.24. (1 point) Define the saturated and the null models, and discuss them with respect to deviance.
3.25. (1 point) Briefly comment on the following plot of deviance residuals of a model as a function of a predictor variable $X_2$:

![Residual vs. $X_2$ plot]

3.26. (2 points) A GLM using a Tweedie Distribution and a log link function is being used to model pure premiums of private passenger automobile property damage liability insurance. There are 100,000 observations. 10 parameters including an intercept were fit. The deviance is 233,183.65, and the estimated dispersion parameter is 2.371. Credit score as a categorical variable is added to the model, with a total of 6 categories. The deviance for this more complex model is 233,134.37. Discuss how you would use an F-Test to determine whether credit score should be added to this model.

3.27. (2 points) The following 5 returns on a stock price are observed:
-0.154, 0.239, -0.064, -0.328, 0.195.
Construct the corresponding Normal Q-Q Plot.
3.28. (0.5 points) Areas have been labeled in the following graph of a Lorenz Curve. Determine the Gini index.

3.29. (0.5 points) With respect to GLMs, briefly discuss pricing coverage options such as deductibles or increased limits.

3.30. (0.5 points) Give an example of a hinge function.

*3.31.* (0.5 points) Five logistic regressions have been fit to the same data. ROC curves have been drawn for each model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.58</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0.66</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>0.73</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>0.79</td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Which model is preferred?
3.32. (1 point) For a GLM, the estimated mean for an individual is 35, with variance 5. Determine a 95% confidence interval for the estimated mean.

3.33. (1.5 points) Five different Generalized Linear Models, have been fit to the same set of 400 observations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Fitted Parameters</th>
<th>LogLikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>-730.18</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>-726.24</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>-723.56</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>-721.02</td>
</tr>
<tr>
<td>E</td>
<td>7</td>
<td>-717.50</td>
</tr>
</tbody>
</table>

Which model has the best BIC (Bayesian Information Criterion)?

Use the following information for the following four questions:
- There is data on commercial building insurance claims frequency.
- A Poisson GLM was fit using the log link function.
- A categorical predictor used is building occupancy class, coded 1 through 4, with 1 being the base class.
- A binary predictor used is sprinklered status, with 1 being yes and 0 being no.
- A continuous predictor used is: \( \ln[\text{amount of insurance} / 200,000] = \ln[\text{AOI} / 200,000] \).
- The fitted intercept is \( \beta_0 = -3.8 \).
- The fitted parameters for building occupancy classes 2, 3, and 4 are: \( \beta_1 = 0.3, \beta_2 = 0.5, \beta_3 = 0.1 \).
- The fitted parameter for sprinklers is: \( \beta_4 = -0.5 \).
- The fitted parameter for \( \ln[\text{AOI} / 200,000] \) is: \( \beta_5 = 0.4 \).
- An interaction term between sprinkler status and \( \ln[\text{AOI} / 200,000] \) is included in the model; the fitted parameter is: \( \beta_6 = -0.1 \).

3.34. (1 point) Determine the fitted frequency for a $100,000 building in occupancy class 1 without sprinklers.

3.35. (1 point) Determine the fitted frequency for a $250,000 building in occupancy class 2 with sprinklers.

3.36. (1 point) Determine the fitted frequency for a $300,000 building in occupancy class 3 without sprinklers.

3.37. (1 point) Determine the fitted frequency for a $600,000 building in occupancy class 4 with sprinklers.
3.38. (1 point) The following are histograms of deviance residuals for GLMs. Which of the following histograms represents the best model?

A. ![Histogram A]
B. ![Histogram B]
C. ![Histogram C]
D. ![Histogram D]
E. ![Histogram E]

3.39. (2 points) You are constructing a Generalized Linear Model.
(a) (0.5 point) If the model is additive, what link function would you use?
(b) (0.5 point) If the model is multiplicative, what link function would you use?
(c) (0.5 point) If the variance is proportional to the mean, what distribution would you use?
(d) (0.5 point) If the standard deviation is proportional to the mean, what distribution would you use?
3.40. (1 point) For a GLM, here is a partial residual plot for the predictor variable $X_4$:

Briefly discuss the meaning of this plot. If necessary, what is a possible solution?

3.41. (1.5 points) With respect to GLMs, discuss the training, validation, and test sets.

3.42. (2 points) Exponential families have a relationship between their mean and variance:

$$V(Y_i) = \phi V(\mu_i) / \omega_i,$$

where $V(\mu)$ is the variance function.

List different exponential families and their variance functions.

3.43. (6 points) You are given the following 20 breaking strengths of wires:


With the aid of a computer, construct a Normal Q-Q Plot.
*3.44.* (5 points) You have the following data on reported occurrences of a communicable disease in two areas of the country at 2 month intervals:

<table>
<thead>
<tr>
<th>Months</th>
<th>Area A</th>
<th>Area B</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>12</td>
<td>17</td>
<td>27</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>28</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>29</td>
</tr>
<tr>
<td>18</td>
<td>17</td>
<td>33</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
<td>31</td>
</tr>
</tbody>
</table>

Let $X_1 = \ln(\text{months})$. Let $X_2 = 0$ for Area A and 1 for Area B.

Assume the number of occurrences $Y_i$ are Poisson variables with means $\mu_i$, and

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}.$$  

Set up the equations to be solved in order to fit this model via maximum likelihood.

3.45. (1 point) Which of the following statements are true with respect to Generalized Linear Models?

1. Errors are assumed to be Normally Distributed.
2. The link function defines the relationship between the expected response variable and the linear combination of the predictor variables.
3. The use of a log link function assumes the rating variables relate multiplicatively to one another.

3.46. (1.5 points) Generalized Linear Models with a overdispersed Poisson error structure and a log link function have been fit in order to model claim frequency for Homeowners Insurance. The models use many variables. The homes have been split into four age categories. A model that uses age has a deviance of 3306.9. An otherwise similar model that does not use age has a deviance of 3320.2, and an estimated dispersion parameter of 1.83.

The null hypothesis is to use the model that does not include age.

The alternative hypothesis is to use the model that does include age.

Calculate the F-test statistic.

Discuss how you would perform the test.

*3.47.* (1 point) Discuss model lift.
3.48. (1.5 points) The following graph displays the modeled log of the frequency relativity by age for two different frequency of premium payment: yearly in red pluses, and four times a year in blue dots. Also approximate 95% confidence intervals are shown for each case.

Question continued on the next page.
The following similar graph displays the modeled log of the frequency relativity by age for males in blue dots and females in red pluses. Also approximate 95% confidence intervals are shown for each case.

Briefly compare and contrast the interaction of age of driver and payment frequency with the interaction of age of driver and gender.
3.49. (0.5 points) For two GLMs you are given the following graphs based on holdout data:

Which model do you prefer and why?
3.50. (2 points) There are three age groups of cars: A, B, C. There are also three size categories of cars: small, medium, large. Specify the following structural components of a generalized linear model.
   i. Design matrix
   ii. Vector of model parameters

3.51. (2 points) Briefly discuss, compare and contrast under-fitting and over-fitting a model.

3.52. (0.5 points) Discuss the following graph of Cook’s Distance for 21 observations:

3.53. (2 points) Use the following information on two Generalized Linear Models fit to the same 100 data points:
   Number of Fitted Parameters | Loglikelihood
   -----------------------------|---------------------
   6                           | -321.06             
   7                           | -319.83             
(a) Based on AIC (Akaike Information Criterion), which model is preferred? 
(b) Based on BIC (Bayesian Information Criterion), which model is preferred?
3.54. (1 point) Which of the following Normal Q-Q Plots is most likely to be of data drawn from a Normal Distribution?

A.  

B.  

C.  

D.  

E.  

Sample Quantiles

Normal Quant.
3.55. (2.5 points) For each of the following situations, give the typical generalized linear model form. State the distributional form of the error and link function typically used.
(a) Claim Frequencies.
(b) Claim Counts.
(c) Average Claim Sizes
(d) Probability of Policy Renewal
(e) Pure Premiums

*3.56.* (0.5 points) You are comparing two rating plans. The first has a Gini Index of 0.48, while the second has a Gini Index of 0.55. Which rating plan is preferred?

3.57. (1 point) You are given the following loss ratio chart for a proposed rating plan. Discuss the lift of the proposed plan compared to the current plan.
3.58. (1 point) Below is a graph of a GLM fit to data, showing the natural log of the fitted multiplicative factors for levels of a variable. (8 is the base level.) Also shown are approximate 95% confidence intervals. Briefly discuss what this graph tells the actuary about the fitted model.

3.59. (5 points) The observed claim frequencies for urban vs rural and male vs female drivers are:

<table>
<thead>
<tr>
<th>Claim frequency</th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.200</td>
<td>0.100</td>
</tr>
<tr>
<td>Female</td>
<td>0.125</td>
<td>0.050</td>
</tr>
</tbody>
</table>

There are equal exposures in each of the four cells.
We will fit a GLM using a Poisson Distribution.
(a) (2.5 points) For an additive model, determine the maximum likelihood equations to be solved.
(b) (2.5 points) For an multiplicative model, determine the maximum likelihood equations to be solved.
3.60. (1 point) A logistic regression has been fit to some data. For a certain threshold:

<table>
<thead>
<tr>
<th>Predicted Claims</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Actual No</td>
<td>40,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Claim Yes</td>
<td>1200</td>
<td>1800</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>41,200</td>
<td>11,800</td>
</tr>
</tbody>
</table>

What point would be plotted in the ROC curve?

Use the following information for the next two questions:

X: 1 5 10 25
Y: 5 15 50 100

$Y_1, Y_2, Y_3, Y_4$ are independently Normally distributed with means $\mu_i = \beta X_i$, $i = 1, 2, 3, 4$, and common variance $\sigma^2$.

3.61. (2 points) Determine $\hat{\beta}$ via maximum likelihood.

3.62. (3 points) Estimate the standard deviation of $\hat{\beta}$.

3.63. (1.5 points) A GLM is used to model claim size.
You are given the following information about the model:
• Claim size follows an Inverse Gaussian distribution.
• Log is the selected link function.
• The dispersion parameter is estimated to be 0.00510.
• Territory and gender are used in the model.
• Selected Model Output:
  
<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>8.03</td>
</tr>
<tr>
<td>Territory D</td>
<td>0.18</td>
</tr>
<tr>
<td>Gender - Male</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Calculate the standard deviation of the predicted claim size for a male in Territory D.

3.64. (2 points) List four ways that an actuary can analyze the appropriateness of a Generalized Linear Model.
3.65. (1 point) Briefly comment on the following plot of deviance residuals of a model as a function of the fitted values:

![Residual vs Fitted Value Plot]

3.66. (1 points) You have fit a Generalized Linear Model using an exponential family. What is the deviance?

3.67. (1 point) A GLM has been fit with a log link function. Age is used, grouped into categories. Gender is used. There are categories of Use of Vehicle. Territories are used. The expected pure premium for the base is $207. For the age group 24-26 the coefficient is 0.43. For Male the coefficient is 0.22. For Pleasure Use (No Driving to Work) the coefficient is -0.32. For Territory H the coefficient is 0.36. Determine the expected pure premium for a male, 24-26 years old, Pleasure Use, in Territory H.

3.68. (1 point) Define and briefly discuss ensemble models.
3.69. (2 points) A GLM using a Gamma Distribution and a log link function is being used to model severity of personal injury claims. There are 25,000 observations. 3 parameters were fit: an intercept, time until settlement, and whether there is legal representation. The deviance is 24,359. The estimated dispersion parameter is 1.22. A variable is added to the model, equal to the product of the time until settlement and the legal representation variable. (This is an interaction variable.) The deviance is now 24,352. Determine whether this additional variable should be added to this model. You may use the following: If $X$ follows an F-Distribution with 1 and $n$ degrees of freedom, then $\sqrt{X}$ follows a t-distribution with $n$ degrees of freedom. For $n$ large, a t-distribution is approximately a Standard Normal Distribution. Selected percentiles of the Standard Normal Distribution:

<table>
<thead>
<tr>
<th>Pr(Z &lt; z)</th>
<th>0.800</th>
<th>0.850</th>
<th>0.900</th>
<th>0.950</th>
<th>0.975</th>
<th>0.990</th>
<th>0.995</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>0.842</td>
<td>1.036</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
</tr>
</tbody>
</table>

*3.70.* (1.5 points) Fully discuss model stability and some ways to assess it.

3.71. (8 points) You are given 19 data points: 258, 636, 652, 814, 833, 860, 895, 937, 950, 1009, 1020, 1059, 1103, 1113, 1127, 1139, 1246, 1335, 1770. You wish to compare this data to a Normal Distribution with $\mu = 1000$ and $\sigma = 300$. With the aid of a computer, draw a Q-Q plot.
(4 points) For private passenger automobile liability claim frequency, you use three factors: gender, age of driver, and territory. There are 4 levels for driver age, and 3 territories. A GLM with a log link function is fit. An intercept term is used. Let $\beta_1$ correspond to the intercept term, $\beta_2$ correspond to male, and assign the other parameters as follows:

<table>
<thead>
<tr>
<th>Age of driver</th>
<th>Parameter</th>
<th>Territory</th>
</tr>
</thead>
<tbody>
<tr>
<td>17-21</td>
<td>$\beta_3$</td>
<td>A</td>
</tr>
<tr>
<td>22-29</td>
<td>$\beta_4$</td>
<td>B</td>
</tr>
<tr>
<td>30-59</td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>60+</td>
<td>$\beta_5$</td>
<td></td>
</tr>
</tbody>
</table>

(a) (3 points) What is the design matrix?
(b) (0.5 point) In terms of the fitted parameters, what is the estimated frequency for a 30-59 year old female driver in Territory B?
(c) (0.5 point) In terms of the fitted parameters, what is the estimated frequency for a 22-29 year old male driver in Territory C?

(1.5 points) Five Generalized Linear Models have been fit to the same set of 200 observations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Fitted Parameters</th>
<th>LogLikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>-359.17</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>-357.84</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>-356.42</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>-354.63</td>
</tr>
<tr>
<td>E</td>
<td>7</td>
<td>-353.85</td>
</tr>
</tbody>
</table>

Which model has the best AIC (Akaike Information Criterion)?
3.74. (1.5 points) The following graph displays the modeled log of the relativity by vehicle symbol, for a base level of the other predictor variables in a GLM, for two separate years of data. Approximate 95% confidence intervals are shown.

Here is a second similar graph for a different model, by Territory:

Briefly compare and contrast what the two graphs tell the actuary about each model.
3.75. (1.5 points) Before embarking on a GLM modeling project, it is important to understand the correlation structure among the predictors. Discuss why this is important and what actions may be indicated.

3.76. (1 point) Multiplicative models are the most common type of rating structure used for pricing insurance, due to a number of advantages they have over other structures. Briefly discuss two advantages of a multiplicative rating structure.

3.77. (1.5 points) A GLM using a Gamma Distribution has been fit for modeling severity of medical malpractice claims. There are 1000 observations. 50 parameters were fit, including an intercept. It uses gender and 6 categories of age of claimant. The deviance is 1120.3. An otherwise similar GLM excluding gender and age of claimant has a deviance of 1128.1, and an estimated dispersion parameter of 0.395. Discuss how you would use an F-Test to determine whether age and gender should be used in this model.

3.78. (1.5 point) Briefly discuss limitations on the use of the loglikelihood and deviance to compare the fit of two GLMs.

3.79. (1 point) An insurer sells “Disgrace Insurance” which covers a business against the possibility that their celebrity spokesperson may engage in disgraceful behavior or expressions. You are putting together Generalized Linear Models (GLMs) to try to develop a rating algorithm. Assuming you have plenty of good data, list some variables you would include in your testing of possible GLMs.

3.80. (1 point) Compare and contrast the Gamma and the Inverse Gaussian Distributions.
3.81. (1.5 points) An actuary has historical information relating to personal loan default rates. A logistic model (GLM with a logit link function) was used to estimate the probability of default for a given customer.
The two variables determined to be significant were the size of loan in thousands of dollars and the credit score of the customer.
$\beta_0$ corresponds to the intercept term, $\beta_1$ corresponds to size of loan, and $\beta_2$ corresponds to credit score.
The parameter estimates were determined to be as follows:

- $\beta_0 = 9.5$
- $\beta_1 = 0.01$
- $\beta_2 = -0.02$

a. (0.75 point) Calculate the estimated default rate for a customer who has credit score of 670 and took out a loan for $180,000.
b. (0.75 point) Calculate the estimated default rate for a customer who has credit score of 760 and took out a loan for $100,000.

3.82. (1 point) For a GLM, here is a partial residual plot for the predictor variable $X_1$:

![Partial Residual Plot](image)

Briefly discuss the conclusion from this plot. If necessary, what is a possible solution?
3.83. (6 points) We model average claim severity by type and horsepower of the car:
- Type: Sedan or SUV
- Horsepower: Low, Medium, or High
We observe an equal number of vehicles of each of the six possible types, and the observed average claim severities are:

<table>
<thead>
<tr>
<th>Horsepower</th>
<th>Sedan</th>
<th>SUV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>800</td>
<td>1,500</td>
</tr>
<tr>
<td>Medium</td>
<td>900</td>
<td>1,700</td>
</tr>
<tr>
<td>High</td>
<td>1,100</td>
<td>2,000</td>
</tr>
</tbody>
</table>

We will fit a GLM using a Gamma Distribution.
(a) (3 points) For an additive model, determine the maximum likelihood equations to be solved.
(b) (3 points) For a multiplicative model, determine the maximum likelihood equations to be solved.

3.84. (2 points) A GLM using an Inverse Gaussian Distribution and an inverse link function is being used to model severity of private passenger automobile property damage liability claims. There are 2000 observations.
14 parameters including an intercept were fit. The deviance is 1848.5, and the estimated dispersion parameter is 0.93.
A categorical variable is added to the model based on vehicle type, with a total of 10 categories. The deviance for this more complex model is 1833.0.
Discuss how you would use an F-Test to determine whether vehicle type should be added to this model at the 5% significance level.

3.85. (1.5 points) Five Generalized Linear Models have been fit to the same set of 250 observations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Fitted Parameters</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>1679.1</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>1666.4</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>1655.9</td>
</tr>
<tr>
<td>D</td>
<td>12</td>
<td>1646.2</td>
</tr>
<tr>
<td>E</td>
<td>14</td>
<td>1634.5</td>
</tr>
</tbody>
</table>

Which model has the best BIC (Bayesian Information Criterion)?
3.86. (1 point)
A Generalized Linear Model was fit to data on lapse rates for life insurance policies. Three predictor variables were included in the GLM: calendar year of exposure, policy duration, and product class. The graph below displays logs of the relativities by policy duration. For each band, the black bars at bottom show exposure, quantified on the righthand axis. The GLM results are in green, and are relative to the base level for policy duration. The yellow line (lighter line) is what would have been generated by a ‘one-way’ analysis: i.e., considering just policy duration, without any other factors.

Briefly discuss a likely reason why the green and yellow lines differ.

3.87. (0.5 points) A continuous predictor $x_2$ has a coefficient of $\beta_2 = -0.3$ in a logistic model. For a unit increase in $x_2$, what is the estimated change in the odds?
3.88. (1 point) We are fitting a GLM to private passenger automobile liability pure premiums. Female drivers age 31 to 59 in a rural territory may have observed pure premiums higher or lower than their fitted values. Unmarried male drivers age 17 to 21 in an urban territory may have observed pure premiums higher or lower than their fitted values. Contrast the effect on fitting the GLM of the modeling errors from these two groups.

3.89. (1 point) A logistic regression has been fit to some data. For a certain threshold:

<table>
<thead>
<tr>
<th>Actual Fraud</th>
<th>Predicted Fraud</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>70,000</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>10,000</td>
</tr>
<tr>
<td>Yes</td>
<td>3000</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>5000</td>
</tr>
<tr>
<td>Total</td>
<td>73,000</td>
<td>12,000</td>
</tr>
<tr>
<td></td>
<td>85,000</td>
<td></td>
</tr>
</tbody>
</table>

What point would be plotted in the ROC curve?

3.90. (1 point) How would the standard error help to analyze the results of fitting a Generalized Linear Model (GLM)?

3.91. (1 point) For a rating plan, briefly discuss how to construct a Lorenz Curve and compute the Gini Index.

3.92. (4 points) Assume a set of three observations:
For $z = 1$, we observe 4. For $z = 2$, we observe 7. For $z = 3$, we observe 8.
Fit to these observations a Generalized Linear Model with a Poisson Distribution and a log link function. In other words, assume that each observation is a Poisson random variable, with mean $\lambda$ and $\ln(\lambda) = \beta_0 + \beta_1 z$.

3.93. (1 point) In addition to statistical significance, give other considerations for variable selection.
3.94. (3.5 points) A personal auto class system has three class dimensions:
• Sex: Male vs female
• Age: Youthful vs adult vs retired
• Territory: Urban vs suburban vs rural
An actuary sets rate relativities from the experience of 20,000 cars.
• Urban is the base level in the territory dimension.
• Adult is the base level in the age dimension.
• Male is the base level in the sex dimension.
a. (0.5 point) How many elements does the vector of covariates have in a multiplicative model?
b. (0.5 point) How many elements does the vector of covariates have in an additive model?
c. (1 point) Specify each element of the vector of parameters, with $\beta_0 \leftrightarrow$ the base class.
d. (0.5 point) How many columns does the design matrix have?
e. (0.5 point) How many rows does the design matrix have if each record is analyzed separately?
f. (0.5 point) For grouped data, how many rows does the design matrix have?

3.95. (2 points) Answer the following with respect to deviance residuals of a GLM.
(a) (0.5 points) Define the deviance residual.
(b) (0.5 points) Give an intuitive interpretation of deviance residuals.
(c) (1 point) Discuss how deviance residuals can be used to check the fit of a model.

3.96. (4 points) You have the following data on the renewal of homeowners insurance policies with the ABC Insurance Company:

<table>
<thead>
<tr>
<th>Number of Years Insured</th>
<th>Number of Policies</th>
<th>Number of Policies Renewed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1000</td>
<td>900</td>
</tr>
<tr>
<td>2</td>
<td>900</td>
<td>820</td>
</tr>
<tr>
<td>3</td>
<td>800</td>
<td>740</td>
</tr>
<tr>
<td>4</td>
<td>700</td>
<td>660</td>
</tr>
<tr>
<td>5</td>
<td>600</td>
<td>580</td>
</tr>
</tbody>
</table>

Let $X =$ number of years insured with ABC Insurance Company.
A Generalized Linear Model using a Binomial Distribution with a logit link function will be fit to this data, including an intercept term.
Determine the equations to be solved in order to fit this model via maximum likelihood.

3.97. (0.5 points) The variance of a distribution from the exponential family can be expressed using the following formula: $\text{Var}(y_i) = \frac{\phi \ V(\mu_i)}{\omega_i}$.

Define the parameters $\phi$ and $\omega_i$ in the formula above.
Y₁, Y₂, Y₃, Y₄ are independently Normally distributed with means \( \mu_i = \beta_0 + \beta_1 X_i \), i = 1, 2, 3, 4, and common variance \( \sigma^2 \).

3.98. (2 points) Determine \( \hat{\beta}_1 \) via maximum likelihood.

3.99. (2 points) Determine \( \hat{\beta}_0 \) via maximum likelihood.

3.100. (2 points) Determine \( \hat{\sigma} \) via maximum likelihood.

3.101. (3 points) Estimate the standard deviation of \( \hat{\beta}_1 \).

3.102. (3 points) Estimate the standard deviation of \( \hat{\beta}_0 \).

3.103. (1 point) Five Generalized Linear Models have been fit to the same set of observations. Each model uses the same number of parameters. Which of these models is preferred?

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3609.5</td>
</tr>
<tr>
<td>B</td>
<td>3611.0</td>
</tr>
<tr>
<td>C</td>
<td>3606.3</td>
</tr>
<tr>
<td>D</td>
<td>3602.1</td>
</tr>
<tr>
<td>E</td>
<td>3605.8</td>
</tr>
</tbody>
</table>

3.104. (1 point) Discuss the overdispersed Poisson Distribution.

*3.105.* (1 point) A common statistical rule of thumb is to reject the null hypothesis where the p-value is 0.05 or lower. Is this appropriate for a typical insurance modeling project? Why or why not?
3.106. (1 point) For a GLM, here is a partial residual plot for the predictor variable $X_3$:

Briefly discuss the meaning of this plot. If necessary, what is a possible solution?

*3.107.* (0.5 points) With respect to GLMs, briefly discuss multicollinearity.

3.108. (1 point) An actuary is determining the rates by class and territory. With respect to GLMs, briefly discuss determining territory relativities.

3.109. (1 point) Define a holdout sample of data, and briefly discuss how it can be used in GLM validation.
3.110. (1 point) The following graph shows claim frequency for private passenger automobile insurance by gender and age. (The rectangles represent the number of exposures.)

Briefly discuss the implications for modeling frequency via a Generalized Linear Model.

3.111. (2 points) Using Generalized Linear Models, an actuary Edward Conners has developed a policy renewal model for private passenger automobile insurance written by the Some States Insurance Company. There are two predictor variables:

\( z_1 \) = the number of years the insured has been with Some States.

\( z_2 \) = the age of the principal operator of the vehicle.

The predicted probability of policy renewal is:

\[
\frac{\exp[0.6 + 0.05 z_1 + 0.02 z_2]}{1 + \exp[0.6 + 0.05 z_1 + 0.02 z_2]}.
\]

(a) For a principal operator who is 30 years old, what is the multiplicative relativity of 1 year with Some States compared to 10 years with Some States?

(b) For a principal operator who is 50 years old, what is the multiplicative relativity of 1 year with Some States compared to 10 years with Some States?
3.112. (1 point) Briefly comment on the following plot of deviance residuals of a model as a function of a predictor variable $X_3$:

![Residual plot]

3.113. (6 points) You are given the following information on the labor force participation of 10 married women between the ages of 25 and 35:

<table>
<thead>
<tr>
<th>Child of Age 6 or Less</th>
<th>Years of Education</th>
<th>Participating in the Labor Force</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>12</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>14</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>15</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>16</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>17</td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>11</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>13</td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>15</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>16</td>
<td>Yes</td>
</tr>
</tbody>
</table>

A Generalized Linear Model using a Binomial Distribution with a logit link function will be fit to this data, including an intercept term.

a. (1 point) What are the design matrix and the response vector?

b. (5 points) Determine the equations to be solved in order to fit this model via maximum likelihood.

3.114. (1 point) Les N. DeRisk is an actuary. Les has scrubbed and adjusted the data he will be using for classification ratemaking for a certain line of insurance. Les will run a Generalized Linear Model. List 3 things Les has to specify.
3.115. (1.5 points) You are given two simple quantile plots, one sorted by the current plan and one sorted by a proposed plan. Discuss the lift of the proposed plan compared to the current plan.

![Quantile plots](image)

3.116. (0.5 point) Give an example of a situation where a GLM with a Binomial distribution and logit link function would be used.

Use the following information for the next two questions:

- A GLM using a Gamma Distribution and a log link function has been fit for modeling severity of auto claims.
- The explanatory variables are: \(x_1\) driver age, and \(x_2\) marital status where 1 = married.
- The fitted coefficients are: \(\beta_0 = 8.80\), \(\beta_1 = -0.03\), \(\beta_2 = -0.15\).
- The estimated \(\phi = 0.3\).

*3.117.* (1 point) Determine the estimated mean severity for a 30 year old married driver.

*3.118.* (1 point) Determine the estimated variance of severity for a 40 year old unmarried driver.
3.119. (2 points) The following graph displays the modeled log of the relativity by vehicle symbol, for a base level of the other predictor variables in a GLM. The bold line shows the fitted parameter estimates. Lines indicates two standard errors on either side of the parameter estimate. The dotted line show the relativities implied by a simple one-way analysis. The distribution of exposure for all business considered is also shown as a bar chart at the bottom.

Question continued on the next page.
Here is a second similar graph for a different model.

Briefly compare and contrast what the two graphs tell the actuary about each model.

**3.120.** (1 point) For a line of insurance, an actuary fits separate GLMs to different perils. Discuss one way to combine separate models by peril in order to get a model for all perils.

3.121. (2 points) Claim counts for private passenger automobile insurance are Poisson. The mean frequency, $m$, depends on age and gender.

Briefly discuss and contrast the following two models, where $x$ is age.

(a) (1 point) $\log(\mu) = \alpha_1 + \beta x$, where $\alpha_1$ and $\alpha_2$ depend on gender.

(b) (1 point) $\log(\mu) = \alpha_i + \beta_1 x$, where $\alpha_1$, $\alpha_2$, $\beta_1$, and $\beta_2$ depend on gender.
3.122. (1 point) The following a histograms of deviance residuals for GLMs. Which of the following histograms represents the best model?

A.  
B.  
C.  
D.  
E.  

3.123. (1 point) Geoff Linus Modlin is an actuary using Generalized Linear Models (GLMs) to determine classification rates for private passenger automobile insurance. Geoff notices that the relativity for drivers aged 19 is different between two GLMs based on the same data. Briefly discuss why that can be the case.
3.124. (1 point) You observe 36 monthly returns on a stock. The 9th value from smallest to largest is 0.004. What is the corresponding point in the Normal Q-Q Plot?

3.125. (1.5 points) With respect to GLMs, fully discuss variance inflation factors (VIF).

3.126. (1.5 points) Dollar Bill Bradley, an actuary at the Knickerbocker Insurance Company, has fit a Generalized Linear Model with a overdispersed Poisson error structure and a log link function in order to model claim frequency for automobile liability insurance. His model has a deviance of 2196.1 and estimated dispersion parameter of 2.09. Bill now introduces into the model an additional categorical variable with five categories:
1. Insured has homeowners insurance with Knickerbocker.
2. Insured has homeowners insurance with another insurer.
3. Insured has renters insurance with Knickerbocker.
4. Insured has renters insurance with another insurer.
5. Other
With this additional variable, the model has a deviance of 2179.3. The null hypothesis is to use the simpler model. The alternative hypothesis is to use the more complicated model. Determine the F-test statistic and discuss how you would perform the statistical test.

*3.127.* (1.5 points) Fully discuss cross validation as used with GLMs, including any limitations.

3.128. (1 point) A GLM has been fit in order to predict blood pressure of individuals.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-12.87</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.7033</td>
<td>1.76</td>
</tr>
<tr>
<td>Weight</td>
<td>0.9699</td>
<td>10.42</td>
</tr>
<tr>
<td>Body Surface Area</td>
<td>3.780</td>
<td>6.33</td>
</tr>
<tr>
<td>Duration of Hypertension</td>
<td>0.0684</td>
<td>1.24</td>
</tr>
<tr>
<td>Basal Pulse</td>
<td>-0.0845</td>
<td>4.41</td>
</tr>
<tr>
<td>Stress Index</td>
<td>0.00341</td>
<td>1.83</td>
</tr>
</tbody>
</table>

Briefly discuss this output.
3.129. (2 points) Below are graphs of GLMs fit to Homeowners frequency data, showing the natural log of the fitted multiplicative factors for one or two children in the house relative to none. Also shown are approximate 95% confidence intervals. Briefly compare and contrast what the two graphs tell the actuary about each model.

**HO Liability Frequency**

![Graph of HO Liability Frequency]

**HO Wind Frequency**

![Graph of HO Wind Frequency]
3.130. (2.5 points) An actuary at a private passenger auto insurance company wishes to use a generalized linear model to create an auto severity model using the data below.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Territory A</th>
<th>Territory B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>700,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Female</td>
<td>400,000</td>
<td>300,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Territory A</th>
<th>Territory B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>800</td>
<td>700</td>
</tr>
<tr>
<td>Female</td>
<td>600</td>
<td>500</td>
</tr>
</tbody>
</table>

The model will include three parameters: $\beta_1$, $\beta_2$, $\beta_3$, where $\beta_1$ is the average severity for males, $\beta_2$ is the average severity for Territory A, and $\beta_3$ is an intercept.

Assuming $\beta_3 = 570.356$, solve a generalized linear model with a normal error structure and identity link function for $\beta_1$.

3.131. (1.5 points) Five Generalized Linear Models have been fit to the same set of 60 observations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Fitted Parameters</th>
<th>LogLikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>-220.18</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>-217.40</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>-214.92</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>-213.25</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>-211.03</td>
</tr>
</tbody>
</table>

Which model has the best BIC (Bayesian Information Criterion)?

3.132. (1 point) You fit a GLM using year as one of the predictor variables. The values of year in your data are: 2010, 2011, 2012, 2013, and 2014. You pick 2012 as the base level. Applying statistical tests you determine that the coefficients for 2011 and 2014 are not significant. Discuss what would you do.
### 3.133. (3 points) You are given the following wage distribution table:

<table>
<thead>
<tr>
<th>Ratio to SAWW</th>
<th>Cumulative Portion of Workers</th>
<th>Cumulative Portion of Wages</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.18%</td>
<td>0.01%</td>
</tr>
<tr>
<td>0.20</td>
<td>0.93%</td>
<td>0.13%</td>
</tr>
<tr>
<td>0.30</td>
<td>3.53%</td>
<td>0.79%</td>
</tr>
<tr>
<td>0.40</td>
<td>6.85%</td>
<td>1.96%</td>
</tr>
<tr>
<td>0.50</td>
<td>11.33%</td>
<td>4.00%</td>
</tr>
<tr>
<td>0.60</td>
<td>18.49%</td>
<td>7.98%</td>
</tr>
<tr>
<td>0.70</td>
<td>28.57%</td>
<td>14.56%</td>
</tr>
<tr>
<td>0.80</td>
<td>40.05%</td>
<td>23.13%</td>
</tr>
<tr>
<td>0.90</td>
<td>48.99%</td>
<td>30.75%</td>
</tr>
<tr>
<td>1.00</td>
<td>57.47%</td>
<td>38.80%</td>
</tr>
<tr>
<td>1.10</td>
<td>64.98%</td>
<td>46.69%</td>
</tr>
<tr>
<td>1.20</td>
<td>71.14%</td>
<td>53.76%</td>
</tr>
<tr>
<td>1.30</td>
<td>76.34%</td>
<td>60.25%</td>
</tr>
<tr>
<td>1.40</td>
<td>80.99%</td>
<td>66.51%</td>
</tr>
<tr>
<td>1.50</td>
<td>85.33%</td>
<td>72.80%</td>
</tr>
<tr>
<td>1.75</td>
<td>92.86%</td>
<td>84.92%</td>
</tr>
<tr>
<td>2.00</td>
<td>96.91%</td>
<td>92.48%</td>
</tr>
<tr>
<td>2.25</td>
<td>98.73%</td>
<td>93.41%</td>
</tr>
<tr>
<td>2.50</td>
<td>99.28%</td>
<td>94.41%</td>
</tr>
<tr>
<td>3.00</td>
<td>99.66%</td>
<td>95.79%</td>
</tr>
<tr>
<td>4.00</td>
<td>99.87%</td>
<td>97.28%</td>
</tr>
<tr>
<td>5.00</td>
<td>99.93%</td>
<td>98.05%</td>
</tr>
<tr>
<td>6.00</td>
<td>99.96%</td>
<td>98.52%</td>
</tr>
<tr>
<td>7.00</td>
<td>99.97%</td>
<td>98.84%</td>
</tr>
</tbody>
</table>

With the aid of a computer, draw the corresponding Lorenz curve.

### 3.134. (2 points) Assume there are two models, Model A and Model B, both of which produce an estimate of the expected loss cost (pure premium) for each policyholder. Discuss using Simple Quantile Plots to compare the two models A and B. How are Simple Quantile Plots created? How would one determine the winning model?

### 3.135.* (1.5 points)
A logistic model was built to predict the probability of a claim being fraudulent. (a) Briefly define the discrimination threshold. (b) Briefly discuss the selection of what discrimination threshold to use.
3.136. (4 points) An actuary is considering using a generalized linear model to estimate the expected frequency of a recently introduced insurance product. Given the following assumptions:
- The expected frequency for a risk is assumed to vary by territory and gender.
- A log link function is used.
- A Poisson error structure is used.
- $\beta_0$ is the intercept.
- $\beta_1$ is the effect of gender = Female.
- $\beta_2$ is the effect of Territory = B.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number of Claims</th>
<th>Territory A</th>
<th>Territory B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1200</td>
<td>1100</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>800</td>
<td>900</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number of Exposures</th>
<th>Territory A</th>
<th>Territory B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>24,000</td>
<td>15,000</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>20,000</td>
<td>13,000</td>
<td></td>
</tr>
</tbody>
</table>

Given that $\beta_0 = -3.0300$, determine the expected frequency of a female risk in Territory B.

3.137. (1 point) Briefly discuss interaction in GLMs and give an example of an interaction term.

3.138. (2 points) A GLM has been used to develop an insurance rating plan. There are only two classes A and B, with equal numbers of exposures. The predicted pure premium for Class A is less than that for class B
(a) Determine the Gini Index if the actual losses for the two classes are equal.
(b) Determine the Gini Index if the actual losses for Class A are 0 and for Class B are positive.
3.139. (0.5 points) The following ROC curves are for two medical tests for strep throat:

Which test do you prefer and why?

3.140. (1.5 points)
A GLM has been fit using a Poisson Distribution with $\hat{\beta}_1 = 5.624$ with standard error 0.1978. Using instead an overdispersed Poisson the estimate of $\phi$ is 3.071. For this second model, determine a 95% confidence interval for $\hat{\beta}_1$.

3.141. (1.75 points) An analyst has fit several different variations of a GLM to a large dataset in order to predict pure premiums. For each model variation listed below, draw a simple quintile plot based on the training data. Label the axes and identify each data series.
i. A saturated model
ii. A null model
iii. A model that could be used in practice
3.142. (1 point) Otherwise similar GLMs have been fit, one using a Gamma Distribution and the other using an Inverse Gaussian Distribution. Based on the following histograms of standardized deviance residuals which model do you prefer and why.
3.143. (1 point) The following loss ratio chart for a proposed rating plan was created by:
1. Sorting the dataset based on the model prediction.
2. Bucketing the data into deciles, such that each decile has approximately the same volume of exposures.
3. Within each bucket, calculate the actual loss ratio (under the current plan) for risks within that bucket.
Discuss the lift of the proposed plan compared to the current plan.

![Loss Ratio Chart]

Use the following information for the next two questions:
Three Generalized Linear Models have been fit to the same set of 5000 observations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Fitted Parameters</th>
<th>LogLikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>-9844.16</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>-9822.48</td>
</tr>
<tr>
<td>C</td>
<td>15</td>
<td>-9815.70</td>
</tr>
</tbody>
</table>

3.144. (1 point) Which model has the best AIC (Akaike Information Criterion)?

3.145. (1 point) Which model has the best BIC (Bayesian Information Criterion)?
*3.146.* (1 point) Below are plots of Actual vs. Predicted for two different GLMs.

Which model do you prefer and why.
3.147. (4 points) A GLM has been used to develop an insurance rating plan. The results are given below:

<table>
<thead>
<tr>
<th>Risk</th>
<th>Exposures</th>
<th>Model Predicted Pure Premium</th>
<th>Actual Pure Premium</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>7000</td>
<td>6000</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>1000</td>
<td>4000</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>4000</td>
<td>2000</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>5000</td>
<td>8000</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>3000</td>
<td>1000</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>6000</td>
<td>8000</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>8000</td>
<td>6000</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>2000</td>
<td>4000</td>
</tr>
</tbody>
</table>

Plot the Lorenz curve for this rating plan. Label each axis and the coordinates of each point on the curve.

*3.148.* (2 points) You are given a GLM of collision claim size with the following potential explanatory variables only:
- Vehicle price, which is a continuous variable modeled with a second order polynomial
- Vehicle Age which is a categorical variable with 8 levels
- Average driver age, which is a continuous variable modeled with a first order polynomial
- Number of drivers, which is a categorical variable with three levels
- Gender, which is a categorical variable with two levels
- There is only one interaction in the model, which is between gender and average driver age.

Determine the number of parameters in this model.

3.149. (1.5 points) Discuss how to construct a double lift chart.

3.150. (1.5 points)

Generalized Linear Models have been fit both with and without a certain predictor variable.

<table>
<thead>
<tr>
<th>Model</th>
<th>With</th>
<th>Without</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>8,901.4414</td>
<td>8,905.6226</td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>18,169</td>
<td>18,175</td>
</tr>
<tr>
<td>Scale Parameter</td>
<td>0.4327</td>
<td>0.4523</td>
</tr>
</tbody>
</table>

The null hypothesis is to use the simpler model.
The alternative hypothesis is to use the more complicated model.
Calculate the F-test statistic.
Discuss how you would perform the test.
3.151. (2 points) You are given the following GLM output:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>df</th>
<th>( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>5.07</td>
</tr>
<tr>
<td>Risk Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Group 2</td>
<td>1</td>
<td>0.21</td>
</tr>
<tr>
<td>Group 3</td>
<td>1</td>
<td>0.48</td>
</tr>
<tr>
<td>Vehicle Symbol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symbol 1</td>
<td>1</td>
<td>-0.36</td>
</tr>
<tr>
<td>Symbol 2</td>
<td>0</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Calculate the variance of the pure premium for an insured in Risk Group 3 with Vehicle Symbol 1.

3.152. (1 point) Two GLMs with somewhat different sets of variables have been fit to the same data. Model 1 has a Gini index of 0.16, while Model 2 has a Gini index of 0.12. Briefly discuss which rating plan has better lift.

3.153. (1 point) Discuss how to construct a loss ratio chart.

*3.154.* (1 point) An actuary fits a GLM to a large amount of data on pure premiums for private passenger automobile insurance. The model includes driver age. The actuary wants to test adding a new variable, number of years claims-free: 0, 1, 2, 3, 4 or more. The new variable will only be used for drivers at least 25 years old. The actuary fits an otherwise similar model that includes number of years claims-free to the same data. The effect of driver age in the second model is significantly less than in the first model. Briefly discuss why this may have occurred.
3.155. (1 point) The following loss ratio chart for a proposed rating plan was created by:
1. Sorting the dataset based on the model prediction.
2. Bucketing the data into quintiles, such that each quintile has approximately the same volume of exposures.
3. Within each bucket, calculate the actual loss ratio (under the current plan) for risks within that bucket.
Discuss the lift of the proposed plan compared to the current plan.

![Loss Ratio Chart](image)

3.156. (1 point) An actuary is modeling pure premiums, using a GLM with a log link function. Deductible relativities have been determined separately, and their effect will be included in the GLM via an offset.
The fitted GLM uses two predictors $x_1$ and $x_2$; $\hat{\beta}_0 = 6$, $\hat{\beta}_1 = 0.1$, and $\hat{\beta}_2 = -0.2$.
Calculate the fitted pure premium for a policy with a deductible relativity of 0.8, $x_1 = 13$ and $x_2 = 3$.

*3.157.* (1 point) Laurel and Hardy are each fitting GLMs to the same 100 observations. Laurel proposes using a GLM with 100 parameters. Hardy proposes using a GLM with one parameter, the overall mean. Discuss which if either of their proposals makes sense.
3.158. (1 point) An actuary, Simon Leroy is modeling private passenger automobile liability insurance via a GLM. All of the current rating variables are included in the model; the current territory relativities are included via an offset. Simon adds to the model the number of years a driver has been claims free. The indicated relativities are:

<table>
<thead>
<tr>
<th>Relativity</th>
<th>Percent of Premiums</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 years claims free</td>
<td>1.50</td>
</tr>
<tr>
<td>1 year claims free</td>
<td>1.35</td>
</tr>
<tr>
<td>2 years claims free</td>
<td>1.18</td>
</tr>
<tr>
<td>3 or more years claims free</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>9%</td>
</tr>
<tr>
<td></td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>5%</td>
</tr>
<tr>
<td></td>
<td>82%</td>
</tr>
</tbody>
</table>

What can you infer about the credibility of a single private passenger driver?

*3.159.* (1 point) An actuary is modeling private passenger automobile insurance. Both the number of operators listed on the policy and the age of the youngest operator listed on the policy will be included in the model. Discuss a potential difficulty with this.

*3.160.* (1.5 points) Compare and contrast AIC and BIC. What do the authors of “Generalized Linear Models for Insurance Rating” conclude with respect to the building of GLMs for actuarial work?
3.161. (1 point) You are given the following double lift chart:

Briefly discuss what conclusion you draw and why.
*3.162.* (1 point) Below are shown two simple quantile plots, the first for Plan A and the second for Plan B. In each case, the model plan predictions are shown by dots and the actual by o. Which plan is preferable and why?
3.163. (1 point) Otherwise similar GLMs have been fit, one using a Gamma Distribution and the other using an Inverse Gaussian Distribution. Based on the following histograms of standardized deviance residuals which model do you prefer and why.
3.164. (2 points) You are given the following information about an insurance policy:

- The probability of a policy renewal, \( p(X) \), follows a logistic model with an intercept and one explanatory variable.
- \( \beta_0 = 1 \)
- \( \beta_1 = 0.31 \)

Calculate the odds of renewal at \( x = 8 \).

3.165. (3 points) A logistic model was built to predict the probability of a claim being fraudulent. Consider the predicted probabilities for the 15 claims below to be a representative sample of the total model.

<table>
<thead>
<tr>
<th>Claim Number</th>
<th>Actual Fraud Indicator</th>
<th>Predicted Probability of Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N</td>
<td>37%</td>
</tr>
<tr>
<td>2</td>
<td>N</td>
<td>46%</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>23%</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>13%</td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td>89%</td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>5%</td>
</tr>
<tr>
<td>7</td>
<td>Y</td>
<td>21%</td>
</tr>
<tr>
<td>8</td>
<td>N</td>
<td>74%</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>75%</td>
</tr>
<tr>
<td>10</td>
<td>Y</td>
<td>69%</td>
</tr>
<tr>
<td>11</td>
<td>N</td>
<td>57%</td>
</tr>
<tr>
<td>12</td>
<td>Y</td>
<td>54%</td>
</tr>
<tr>
<td>13</td>
<td>N</td>
<td>53%</td>
</tr>
<tr>
<td>14</td>
<td>N</td>
<td>83%</td>
</tr>
<tr>
<td>15</td>
<td>N</td>
<td>49%</td>
</tr>
</tbody>
</table>

a. (1.5 points) Construct confusion matrices for discrimination thresholds of 0.30 and 0.60.
b. (1.5 points) Plot the Receiver Operating Characteristic (ROC) curve with the discrimination thresholds of 0.30 and 0.60.

Label each axis and the coordinates and discrimination threshold of each point on the curve.

3.166. (1 point) Olaf is an actuary with the Arendelle Insurance Company. Olaf is revising the classification relativities using a GLM with a log link function. Recently another actuary Anna had revised the territory relativities. Olaf with take these territory definitions and relativities as given. Discuss how Olaf should make use of offsets in his modeling of classification relativities.
3.167. (0.75 points) An actuary has split data into training and test groups for a model. The chart below shows the relationship between model performance and model complexity. Model performance is represented by model error and model complexity is represented by degrees of freedom.

![Graph showing model error vs. degrees of freedom for Test and Training sets.]

Briefly discuss the optimal balance of complexity and performance.

*3.168.* (2 points) An actuary is analyzing a partial residual plot of the driver age variable. The plot appears to be non-linear.

a. (1 point) Briefly describe two approaches that can be used to improve the fit of the driver age variable.

b. (1 point) Briefly describe a downside to each of the two approaches discussed in part a. above.
3.169. (2 points) You are given the following information for a GLM of customer retention:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>df</th>
<th>( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>2.182</td>
</tr>
<tr>
<td>Years with Insurer</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>&gt;1</td>
<td>1</td>
<td>1.137</td>
</tr>
<tr>
<td>Last Rate Change</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>&lt;0%</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>0%-10%</td>
<td>1</td>
<td>-0.422</td>
</tr>
<tr>
<td>&gt;10%</td>
<td>1</td>
<td>-0.901</td>
</tr>
</tbody>
</table>

Calculate the probability of retention for a policy with the insurer for 4 years and with a prior rate change of 7%.

3.170. (2 points) An actuary is comparing the output of two generalized linear models to develop a new rating plan for personal auto. Model statistics are shown below:

<table>
<thead>
<tr>
<th></th>
<th>Saturated Model</th>
<th>Model A</th>
<th>Model B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-Likelihood</td>
<td>-100</td>
<td>-130</td>
<td>-123</td>
</tr>
<tr>
<td>Estimated Dispersion Parameter</td>
<td>0.65</td>
<td>0.61</td>
<td>0.63</td>
</tr>
</tbody>
</table>

- Each model is fit to the same set of 50 data points.
- Model A uses 5 parameters (including an intercept).
- Model A is a nested model of Model B, where Model B has an additional variable for driver age.
- Driver age is fit using 5 bins.
- The critical value to be used from the F-distribution is 5.707.

Using two statistical tests, recommend whether or not driver age should be included in the rating plan.
3.171. (2 points) Your coworker, Clifford Clavin, has fit a GLM using the following model form:
\[ g(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \]
The fitted parameters were:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>1</td>
</tr>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>2</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>-5</td>
</tr>
</tbody>
</table>

You know that Cliff used a canonical link function, but do not know which of the following three error distribution Cliff used:
I: Gamma
II: Poisson
III: Normal
Cliff is currently on vacation and cannot be reached; Cliff’s limited notes do not help.
Determine the correct ordering of the three possible models' predicted values at the observed point \((X_1, X_2) = (0.50, 0.29)\).

3.172. (1.25 points) An actuary is planning to add a credit-based insurance score from FIDO to a model that estimates the probability of a policy having a claim. The actuary has decided to offset all of the current model variables before fitting the new variable.

Given the following:
- The current model (without the FIDO insurance score variable) is a logit-link binomial GLM (logistic regression).
- The logit link function is defined as \( g(\mu) = \ln\left(\frac{\mu}{1 - \mu}\right) \)
- The FIDO insurance score is a continuous variable having a value between 1 and 100.
- The current fitted values and FIDO insurance score for three policies as well as regression results from the fit of the FIDO insurance score variable are given below:

<table>
<thead>
<tr>
<th>Policy Number</th>
<th>Fitted Probability Without Insurance Score</th>
<th>FIDO Insurance Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.3%</td>
<td>34</td>
</tr>
<tr>
<td>2</td>
<td>11.2%</td>
<td>66</td>
</tr>
<tr>
<td>3</td>
<td>4.5%</td>
<td>88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.581</td>
</tr>
<tr>
<td>FIDO Insurance Score</td>
<td>-0.032</td>
</tr>
</tbody>
</table>

a. (0.5 point)
Calculate the offset term to be used in the regression for each of the three policies above.
b. (0.75 point)
Calculate the revised fitted probability of having a claim for each of the three policies above.
3.173. (9, 11/03, Q.25) (2 points)
   a. (1 point) Explain why one-way analysis of risk classification relativities can produce indicated relativities that are inaccurate and inconsistent with the data.
   b. (1 point) Describe an approach to calculating risk classification relativities that would reduce the error produced by a one-way analysis.

3.174. (9, 11/06, Q.5) (4 points)
   a. (3 points) Compare the random component, the systematic component, and the link functions of a linear model to those of a generalized linear model.
   b. (1 point) Describe two reasons why the assumptions underlying linear models are difficult to guarantee in application.

3.175. (9, 11/07, Q.4a) (1 point) There are a variety of methods available to a ratemaking actuary when determining classification rates. Compare the Generalized Linear Model to the Classical Linear Model with respect to the following:
   i. The distribution of the response variable.
   ii. The relationship between the mean and variance of the response variable.

3.176. (9, 11/08, Q.3) (2 points) When using a Generalized Linear Model one of the concerns of which the practitioner must be aware is the presence of aliasing within the model.
   a. (1 point) Discuss the two types of aliasing and provide an example of how each can arise in a model.
   b. (1 point) An actuary is using a Generalized Linear Model to determine possible interactions between pure premiums. While reviewing the model, the actuary observes the following pure premiums for liability coverage:

<table>
<thead>
<tr>
<th>Liability Pure Premium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle Size</td>
</tr>
<tr>
<td>Territory</td>
</tr>
<tr>
<td>North</td>
</tr>
<tr>
<td>South</td>
</tr>
<tr>
<td>East</td>
</tr>
<tr>
<td>West</td>
</tr>
</tbody>
</table>

   Assuming equal exposure distribution across all combinations of territory and vehicle size, demonstrate how aliasing can be used to exclude a level from either the territory or the vehicle size variable.

3.177. (2 points) Use the information in the previous question, 9, 11/08, Q.3. Take North and Medium as the base levels.
   Specify the following structural components of a generalized linear model:
   Definition of Variables, Design matrix, Vector of responses, Vector of model parameters.
3.178. (9, 11/09, Q.3) (3 points) Consider a simple private passenger auto classification system that has two rating variables: territory (urban or rural) and gender (male or female). The observed average claim severities are:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>$400</td>
<td>$250</td>
</tr>
<tr>
<td>Female</td>
<td>$200</td>
<td>$100</td>
</tr>
</tbody>
</table>

$Y$, the response variable, is the average claim severity. Male (x1), Female (x2), Urban (x3) and Rural (x4) are the 4 covariates. A uniquely defined model is:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e.$$  

a. (2 points) Using the classical linear model, derive the equations to solve for the parameters $\beta_1$, $\beta_2$ and $\beta_3$ using the sum of squared errors. (Do NOT solve the equations.)

b. (1 point) Briefly describe two underlying assumptions of the classical linear model. Explain why the model may not be able to guarantee these assumptions.

3.179. (9 points) Use the information in the previous question, 9, 11/09, Q.3. As per the exam question, use the following variables: Male ($X_1$), Female ($X_2$), Urban ($X_3$).

a. (2 points) Specify the following structural components of a generalized linear model: Design matrix, Vector of responses, Vector of model parameters.

b. (2 points) Determine the equations that would need to be solved in order to fit the model. Assume a Gamma Distribution and the identity link function. Assume equal exposures for each cell.

c. (2 points) Determine the equations that would need to be solved in order to fit the model. Assume a Gamma Distribution and the inverse link function. Assume equal exposures for each cell.

d. (3 points) Determine the equations that would need to be solved in order to fit the model. Assume a Inverse Gaussian Distribution and the squared inverse link function. Assume equal exposures for each cell.

For the Inverse Gaussian: $f(x) = \frac{\theta}{\sqrt{2\pi}} \frac{\exp\left[ -\frac{\theta^2}{2x} \left( \frac{x}{\mu} - 1 \right)^2 \right]}{x^{1.5}}$, mean = $\mu$, variance = $\mu^3 / \theta$.

3.180. (9, 11/10, Q.3) (3.5 points) The following chart represents claim frequencies for a commercial auto book of business:

<table>
<thead>
<tr>
<th>Claim Frequencies (1,000 Vehicle-Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private Passenger</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Territory A</td>
</tr>
<tr>
<td>Territory B</td>
</tr>
</tbody>
</table>

a. (2 points) Complete the first step in solving a generalized linear model by specifying the design matrix and vector of beta parameters.

b. (0.5 point) For each of the Poisson and gamma error structures, describe the relationship between the variance and the expected value and how these relationships differ.

c. (1 point) Once the link function and error structure have been selected, describe the process to determine the final beta parameters.
3.181. (8, 11/11, Q.3) (1.5 points) An actuary is considering performing a one-way analysis to provide pricing guidance for an insurance company's personal auto book of business.  
   a. (0.5 point) Briefly describe two shortcomings associated with one-way analyses.  
   b. (1 point) Provide an example of how each shortcoming in part a. above may arise.

3.182. (8, 11/12, Q.2) (2.25 points) A private passenger auto insurance company orders a report whenever it writes a policy, showing what other insurance the policyholder has purchased. The following table shows claim frequencies (per 100 earned car-years) for bodily injury liability coverage, split by whether the policyholder has a homeowners policy and whether the policyholder had a prior auto policy:

<table>
<thead>
<tr>
<th>Prior Auto Policy</th>
<th>Homeowners Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>No</td>
<td>8</td>
</tr>
</tbody>
</table>

The table does not include the experience of policyholders with missing data.  
   a. (1.25 points) Specify the following structural components of a generalized linear model that estimates frequencies for this book of business.  
      i. Error distribution  
      ii. Link function  
      iii. Vector of responses  
      iv. Vector of model parameters  
      v. Design matrix  
   b. (1 point) Describe how the missing data may cause problems for the company in developing the model, and suggest a solution.
3.183. (8, 11/12, Q.4) (1.75 points) An actuary has historical information relating to customer retention. A logistic model was used to estimate the probability of renewal for a given customer. The two variables determined to be significant were the size of rate change and number of phone calls the insured made to the company. The parameter estimates were determined to be as follows:

<table>
<thead>
<tr>
<th>Rate Change</th>
<th>Parameter Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decrease to 3.9% increase</td>
<td>0.3323</td>
</tr>
<tr>
<td>4.0% to 6.9% increase</td>
<td>0</td>
</tr>
<tr>
<td>Increase of 7.0% or more</td>
<td>-0.4172</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Phone Calls in Past Year</th>
<th>Parameter Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>-0.2128</td>
</tr>
<tr>
<td>2+</td>
<td>-0.4239</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intercept Term</th>
<th>Parameter Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.793</td>
</tr>
</tbody>
</table>

a. (0.75 point) Calculate the renewal probability for a customer who has a 7% rate increase and called the company twice in the past year.

b. (1 point) The company needs policyholder retention to be above 78% to maintain growth and expense ratio goals. A possible strategy is to add the number of phone calls to the classification plan and use the model results to determine the rate increase.

Construct an argument either in favor of or against the strategy above, describing two reasons for that position.

3.184. (8, 11/13, Q.2) (3.5 points) An actuary at a private passenger auto insurance company wishes to use a generalized linear model to create an auto frequency model using the data below.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Territory A</th>
<th>Territory B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>700</td>
<td>600</td>
</tr>
<tr>
<td>Female</td>
<td>400</td>
<td>420</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Territory A</th>
<th>Territory B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1,400</td>
<td>1,000</td>
</tr>
<tr>
<td>Female</td>
<td>1,000</td>
<td>1,200</td>
</tr>
</tbody>
</table>

The model will include three parameters: \( \beta_1, \beta_2, \beta_3 \), where \( \beta_1 \) is the average frequency for males, \( \beta_2 \) is the average frequency for Territory A, and \( \beta_3 \) is an intercept.

a. (0.5 point) Define the design matrix \([X]\).

b. (0.25 point) Define the vector of responses \([Y]\).

c. (2.25 points) Assuming \( \beta_3 = 0.35 \), solve a generalized linear model with a normal error structure and identity link function for \( \beta_1 \).

d. (0.5 point) The actuary determines that the analysis results would be improved by assuming a Poisson error structure with a log link function.

Identify two reasons this structure may better suit this data.
3.185. (8, 11/14, Q.3) (2 points) The random component of a generalized linear model must come from the exponential family of distributions. The variance of a distribution from the exponential family can be expressed using the following formula: \[ \text{Var}(Y_i) = \frac{\phi V(\mu_i)}{\omega_i} \]

a. (0.5 point) Define the parameters \( \phi \) and \( \omega_i \) in the formula above.

b. (1 point) For each of the data sets below, identify the error distribution that should be used to model the data. Briefly explain why that error distribution is appropriate.

i. Severity

ii. Policy Renewal Retention

c. (0.5 point) For each of the error distributions in part b. above, provide an example of how \( \omega_i \) should be assigned for the type of data being modeled.

3.186. (CAS S Sample Exam 2015, Q.4) (2 points)

An actuary wants to estimate the probability of a home insurance policy having a claim by using a logistic regression model. He has the following pieces of information from 1,000 historical policies:

- Cost of the home, in $000s
- Age of the home, in years
- Whether or not there was a claim on the policy

The actuary is considering a number of different model specifications. Below are the models he is considering along with the calculated deviance of each model:

<table>
<thead>
<tr>
<th>Model #</th>
<th>Included Variables</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept + Cost</td>
<td>1085.0</td>
</tr>
<tr>
<td>2</td>
<td>Intercept + Cost + Age</td>
<td>1084.8</td>
</tr>
<tr>
<td>3</td>
<td>Intercept + Cost + (Cost * Age)</td>
<td>1083.0</td>
</tr>
<tr>
<td>4</td>
<td>Intercept + Cost + Cost^2 + Cost^3</td>
<td>1081.9</td>
</tr>
<tr>
<td>5</td>
<td>Intercept + Cost + Cost^2 + Cost^3 + Cost^4</td>
<td>1081.6</td>
</tr>
</tbody>
</table>

Determine the optimal model using the Bayesian Information Criterion.

3.187. (2 points) In the previous question, determine the optimal model using instead the Akaike Information Criterion.
3.188. (CAS S, 11/15, Q.32) (2 points)
A GLM is used to model claim size. You are given the following information about the model:
- Claim size follows a Gamma distribution.
- Log is the selected link function.
- The dispersion parameter $f$ is estimated to be 2.
- Model Output:

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.32</td>
</tr>
<tr>
<td>Location - Urban</td>
<td>0.00</td>
</tr>
<tr>
<td>Location - Rural</td>
<td>-0.64</td>
</tr>
<tr>
<td>Gender - Female</td>
<td>0.00</td>
</tr>
<tr>
<td>Gender - Male</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Calculate the variance of the predicted claim size for a rural male.

3.189. (CAS S, 11/15, Q.33) (2 points)
You are given the following output from a GLM to estimate the probability of a claim:
- Distribution selected is Binomial.
- Link selected is Logit.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.485</td>
</tr>
<tr>
<td>Vehicle Body</td>
<td></td>
</tr>
<tr>
<td>Coupe</td>
<td>-0.881</td>
</tr>
<tr>
<td>Roadster</td>
<td>-1.047</td>
</tr>
<tr>
<td>Sedan</td>
<td>-1.175</td>
</tr>
<tr>
<td>Station wagon</td>
<td>-1.083</td>
</tr>
<tr>
<td>Truck</td>
<td>-1.118</td>
</tr>
<tr>
<td>Utility</td>
<td>-1.330</td>
</tr>
<tr>
<td>Driver's Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>-0.025</td>
</tr>
<tr>
<td>Area</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.094</td>
</tr>
<tr>
<td>C</td>
<td>0.037</td>
</tr>
<tr>
<td>D</td>
<td>-0.101</td>
</tr>
</tbody>
</table>

Calculate the estimated probability of a claim for:
- Driver Gender: Female
- Vehicle Body: Sedan
- Area: D
3.190. (CAS S, 11/15, Q.34) (1 point)
You are given the following information for a model of vehicle claim counts by policy:
• The response distribution is Poisson and the model has a log link function.
• The model uses two categorical explanatory variables: Number of Youthful Drivers and Number of Adult Drivers.
• The parameters of the model are given:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Degrees of Freedom</th>
<th>(\hat{\beta})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-2.663</td>
</tr>
<tr>
<td>Number of Youthful Drivers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.132</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-0.031</td>
</tr>
<tr>
<td>Number of Adult Drivers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-0.031</td>
</tr>
</tbody>
</table>

Calculate the predicted claim count for a policy with one adult driver and one youthful driver.

3.191. (CAS S, 11/15, Q.35) (2 points)
You are given a GLM of liability claim size with the following potential explanatory variables only:
• Vehicle price, which is a continuous variable modeled with a third order polynomial
• Average driver age, which is a continuous variable modeled with a first order polynomial
• Number of drivers, which is a categorical variable with four levels
• Gender, which is a categorical variable with two levels
• There is only one interaction in the model, which is between gender and average driver age. Determine the maximum number of parameters in this model.

3.192. (CAS S, 11/15, Q.36) (2 points) You are given the following information for two potential logistic models used to predict the occurrence of a claim:
• Model 1: (AIC = 262.68)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>(\hat{\beta})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.264</td>
</tr>
<tr>
<td>Vehicle Value ($000s)</td>
<td>0.212</td>
</tr>
<tr>
<td>Gender-Female</td>
<td>0.000</td>
</tr>
<tr>
<td>Gender-Male</td>
<td>0.727</td>
</tr>
</tbody>
</table>

• Model 2: (AIC = 263.39)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>(\hat{\beta})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.894</td>
</tr>
<tr>
<td>Gender-Female</td>
<td>0.000</td>
</tr>
<tr>
<td>Gender-Male</td>
<td>0.727</td>
</tr>
</tbody>
</table>

• AIC is used to select the most appropriate model.
Calculate the probability of a claim for a male policyholder with a vehicle valued $12,000 by using the selected model.
3.193. (CAS S, 11/15, Q.38) (2 points)
You are testing the addition of a new categorical variable into an existing GLM.
You are given the following information:
• The change in model deviance after adding the new variable is -53.
• The change in AIC after adding the new variable is -47.
• The change in BIC after adding the new variable is -32.
• Prior to adding the new variable, the model had 15 parameters.
Calculate the number of observations in the model.

3.194. (8, 11/15, Q.3) (2.5 points) An actuary is considering using a generalized linear model to estimate the expected frequency of a recently introduced insurance product. Given the following assumptions:
• The expected frequency for a risk is assumed to vary by state and gender.
• A log link function is used.
• A Poisson error structure is used.
• The likelihood function of a Poisson is
  \[ l(y; \mu) = \sum \ln f(y_i; \mu_i) = \sum \{-\mu_i + y_i \ln[\mu_i] - \ln[y_i!]\} \]
• \( \beta_1 \) is the effect of gender = Male.
• \( \beta_2 \) is the effect of gender = Female.
• \( \beta_3 \) is the effect of State = State A.

<table>
<thead>
<tr>
<th>Claim Frequency</th>
<th>State A</th>
<th>State B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.0920</td>
<td>0.0267</td>
</tr>
<tr>
<td>Female</td>
<td>0.1500</td>
<td>0.0500</td>
</tr>
</tbody>
</table>

Given that \( \beta_3 = 1.149 \), determine the expected frequency of a male risk in State A.
### 3.195. (CAS S, 5/16, Q.29) (2 points) You are given the following information for a fitted GLM:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>df</th>
<th>( \hat{\beta} )</th>
<th>( \text{se} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-2.358</td>
<td>0.048</td>
</tr>
<tr>
<td>Driver's Age</td>
<td>2</td>
<td>0.288</td>
<td>0.064</td>
</tr>
<tr>
<td>Area</td>
<td>2</td>
<td>-0.036</td>
<td>0.053</td>
</tr>
<tr>
<td>Vehicle Body</td>
<td>2</td>
<td>1.136</td>
<td>-0.371</td>
</tr>
<tr>
<td>Area</td>
<td>2</td>
<td>0.000</td>
<td>0.151</td>
</tr>
</tbody>
</table>

The probability of a driver in age group 2, from area C and with vehicle body type Other, having an accident is 0.22.

Calculate the odds ratio of the driver in age group 3, from area C and with vehicle body type Sedan having an accident.

### 3.196. (CAS S, 5/16, Q.30) (2 points) You are given the following information for a fitted GLM:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>df</th>
<th>( \hat{\beta} )</th>
<th>( \text{se} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-2.358</td>
<td>0.048</td>
</tr>
<tr>
<td>Area</td>
<td>2</td>
<td>0.000</td>
<td>0.062</td>
</tr>
<tr>
<td>Rural</td>
<td>1</td>
<td>-1.129</td>
<td>0.151</td>
</tr>
</tbody>
</table>

Calculate the modeled probability of an Urban driver having an accident.
3.197. (CAS S, 5/16, Q.31) (2 points) You are given the following information for a fitted GLM:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>df</th>
<th>( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>2.100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Zone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.678</td>
<td>4.227</td>
<td>1.336</td>
<td>0.000</td>
<td>1.734</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vehicle Class</th>
<th>df</th>
<th>( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convertible</td>
<td>1</td>
<td>1.200</td>
</tr>
<tr>
<td>Coupe</td>
<td>1</td>
<td>1.300</td>
</tr>
<tr>
<td>Sedan</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>Truck</td>
<td>1</td>
<td>1.406</td>
</tr>
<tr>
<td>Minivan</td>
<td>1</td>
<td>1.875</td>
</tr>
<tr>
<td>Station wagon</td>
<td>1</td>
<td>2.000</td>
</tr>
<tr>
<td>Utility</td>
<td>1</td>
<td>2.500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Driver Age</th>
<th>df</th>
<th>( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youth</td>
<td>1</td>
<td>2.000</td>
</tr>
<tr>
<td>Middle age</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>Old</td>
<td>1</td>
<td>1.800</td>
</tr>
</tbody>
</table>

Calculate the predicted claim size for an observation from Zone 3, with Vehicle Class Truck and Driver Age Old.
3.198. (CAS S, 5/16, Q.32) (2 points) You are given the following information for a fitted GLM:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>df</th>
<th>( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>2.100</td>
</tr>
<tr>
<td>Zone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>7.678</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4.227</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.336</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1.734</td>
</tr>
<tr>
<td>Vehicle Class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convertible</td>
<td>1</td>
<td>1.200</td>
</tr>
<tr>
<td>Coupe</td>
<td>1</td>
<td>1.300</td>
</tr>
<tr>
<td>Sedan</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>Truck</td>
<td>1</td>
<td>1.406</td>
</tr>
<tr>
<td>Minivan</td>
<td>1</td>
<td>1.875</td>
</tr>
<tr>
<td>Station wagon</td>
<td>1</td>
<td>2.000</td>
</tr>
<tr>
<td>Utility</td>
<td>1</td>
<td>2.500</td>
</tr>
<tr>
<td>Driver Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Youth</td>
<td>1</td>
<td>2.000</td>
</tr>
<tr>
<td>Middle age</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>Old</td>
<td>1</td>
<td>1.800</td>
</tr>
</tbody>
</table>

Calculate the variance of a claim size for an observation from Zone 4, with Vehicle Class Sedan and Driver Age Middle age.
3.199. (CAS S, 5/16, Q.33) (2 points)
You are given the following information for a GLM of customer retention:

- **Response variable**: Retention
- **Response distribution**: Binomial
- **Link**: Logit

<table>
<thead>
<tr>
<th>Parameter</th>
<th>df</th>
<th>( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>1.530</td>
</tr>
<tr>
<td>Number of Drivers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>&gt;1</td>
<td>1</td>
<td>0.735</td>
</tr>
<tr>
<td>Last Rate Change</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>&lt;0%</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>0%-10%</td>
<td>1</td>
<td>-0.031</td>
</tr>
<tr>
<td>&gt;10%</td>
<td>1</td>
<td>-0.372</td>
</tr>
</tbody>
</table>

Calculate the probability of retention for a policy with 3 drivers and a prior rate change of 5%.

3.200. (CAS S, 5/16, Q.35) (2 points) You are given the following information about three candidates for a Poisson frequency GLM on a group of condominium policies:

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables in the Model</th>
<th>DF</th>
<th>Log Likelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Risk Class</td>
<td>5</td>
<td>-47,704</td>
<td>95,418</td>
<td>95,473.61182</td>
</tr>
<tr>
<td>2</td>
<td>Risk Class + Region</td>
<td></td>
<td>-47,495</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Risk Class + Region + Claim Indicator</td>
<td>10</td>
<td>-47,365</td>
<td>94,750</td>
<td></td>
</tr>
</tbody>
</table>

- Insureds are from one of five Risk Class: A, B, C, D, E
- Condominium policies are located in several regions
- Claim Indicator is either Yes or No
- All models are built on the same data

Calculate the absolute difference between the AIC and the BIC for Model 2.
3.201. (CAS S, 5/16, Q.36) (2 points) You are given the following two graphs comparing the fitted values to the residuals of two different linear models:

Determine which of the following statements are true.
I. Graph 1 indicates the data is homoscedastic
II. Graph 1 indicates the data is heteroskedastic (a lack of homoscedasticity)
III. Graph 2 indicates the data is non-normal

3.202. (CAS S, 5/16, Q.37) (2 points) Determine which of the following GLM selection considerations is true.
A. The model with the largest AIC is always the best model in model selection process.
B. The model with the largest BIC is always the best model in model selection process.
C. The model with the largest deviance is always the best model in model selection process.
D. Other things equal, when the number of observations > 1000, AIC penalizes more for the number of parameters used in the model than BIC.
E. Other things equal, when number of observations > 1000, BIC penalizes more for the number of parameters used in the model than AIC.
3.203. (CAS S, 5/16, Q.38) (2 points) You are testing the addition of a new categorical variable into an existing GLM, and are given the following information:
• A is the change in AIC and B is the change in BIC after adding the new variable.
• B > A + 25
• There are 1500 observations in the model.
Calculate the minimum possible number of levels in the new categorical variable.

3.204. (CAS S, 5/16, Q.41) (1 point) A Poisson regression model with log link is used to estimate the number of diabetes deaths. The parameter estimates for the model are:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>df</th>
<th>( \hat{\beta} )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-15.000</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Gender: Female</td>
<td>1</td>
<td>-1.200</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Gender: Male</td>
<td>1</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>0.150</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Age(^2)</td>
<td>1</td>
<td>0.004</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Age \times Gender: Female</td>
<td>1</td>
<td>0.012</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Age \times Gender: Male</td>
<td>0</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

Calculate the expected number of deaths for a population of 100,000 females age 25.

3.205. (8, 11/16, Q.4) (3 points) An actuary is conducting a generalized linear model (GLM) analysis on historical personal automobile data in order to develop a rating plan.

a. (1.5 points)
Argue against the following factors being included as predictors in the actuary's GLM analysis:
   i. Limit of liability.
   ii. Number of coverage changes during the current policy period.
   iii. ZIP code of the garaging location of the automobile.

b. (1 point) The actuary is modeling pure premium with a log-link function and a Tweedie error distribution (1 < p < 2). Provide two arguments against the inclusion of deductible as a predictor in the actuary's GLM analysis.

c. (0.5 point) Other than including deductible as a predictor in the GLM, describe how to determine deductible relativities and how such relativities can be incorporated in a GLM.
3.206. (8, 11/16, Q.5) (2.25 points)
A GLM has been used to develop an insurance rating plan. The results are given below:

<table>
<thead>
<tr>
<th>Risk</th>
<th>Model Predicted Loss</th>
<th>Actual Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,000</td>
<td>2,050</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>220</td>
</tr>
<tr>
<td>3</td>
<td>1,500</td>
<td>1,480</td>
</tr>
<tr>
<td>4</td>
<td>800</td>
<td>850</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
<td>400</td>
</tr>
</tbody>
</table>

a. (1.75 points) Plot the Lorenz curve for this rating plan. Label each axis and the coordinates of each point on the curve.
b. (0.5 point) Briefly describe how the Gini index is calculated and what the Gini index measures when applied to an insurance rating program. Do not calculate the Gini index.

3.207. (8, 11/16, Q.6) (2.5 points) An actuary has constructed a three-variable Tweedie GLM with a log-link function to estimate loss ratios for commercial property new business. The actuary wants to create a second model for renewal business that will include all of the variables from the new business model, plus a variable for the prior year claim count. The actuary requires that the coefficients of the variables: Average Building Age, log(Manual Premium), and Location Count, are consistent between the new and renewal models.

The fitted new business model parameters are as follows:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Name</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>0.910</td>
<td></td>
</tr>
<tr>
<td>Average Building Age (Years)</td>
<td>age</td>
<td>0.013</td>
</tr>
<tr>
<td>log(Manual Premium)</td>
<td>logprem</td>
<td>-0.187</td>
</tr>
<tr>
<td>Location Count</td>
<td>loccnt</td>
<td>0.062</td>
</tr>
</tbody>
</table>

a. (0.75 point) Calculate the modeled loss ratio for a new business policy with a manual premium of $25,000, an average building age of four years, and having eight locations.
b. (0.75 point) Briefly describe how to produce the renewal business model, and specify the resulting equation for the renewal business modeled loss ratio.
c. (1 point) Identify and briefly describe two techniques that the actuary can use to assess the stability of the new variable in the renewal business model.
3.208. (8, 11/16, Q.7) (1.5 points) A company is considering modifying its rating plan to include factors by age group. Below are statistics for the base model and for the new model.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Base Model</th>
<th>New Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loglikelihood</td>
<td>-750</td>
<td>-737.5</td>
</tr>
<tr>
<td>Deviance</td>
<td>500</td>
<td>475</td>
</tr>
<tr>
<td>Parameters</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Data points</td>
<td>1,000,000</td>
<td>1,000,000</td>
</tr>
</tbody>
</table>

a. (1 point) Calculate the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for both models.
b. (0.25 point) Explain whether AIC or BIC is a more reliable test statistic as an indicator of whether to adopt the new model.
c. (0.25 point) Recommend and briefly justify whether to adopt the new model.

3.209. (8, 11/17, Q.4) (1.75 points) An actuary has split data into training and test groups for a model. The chart below shows the relationship between model performance and model complexity. Model performance is represented by model error and model complexity is represented by degrees of freedom.

a. (0.5 point) Briefly describe two reasons for splitting modeling data into training and test groups.
b. (0.75 point) Briefly describe whether each of the following model iterations has an optimal balance of complexity and performance.
   i. Model iteration 1: 10 degrees of freedom
   ii. Model iteration 2: 60 degrees of freedom
   iii. Model iteration 3: 100 degrees of freedom
c. (0.5 points) Identify and briefly describe one situation where it is an advantage to split the data by time rather than by random assignment.
3.210. (8, 11/17, Q.5) (1.75 points) An analyst has fit several different variations of a logistic GLM to a dataset containing 1,000 records of fraudulent claims and 9,000 records of legitimate claims. For each model variation listed below, draw a quintile plot based on the training data. Label the axes and identify each data series.

   i. A saturated model
   ii. A null model
   iii. A model that could be used in practice

3.211. (8, 11/17, Q.6) (3.5 points) A logistic model was built to predict the probability of a claim being fraudulent. Consider the predicted probabilities for the 10 claims below to be a representative sample of the total model.

<table>
<thead>
<tr>
<th>Claim Number</th>
<th>Actual Fraud Indicator</th>
<th>Predicted Probability of Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y</td>
<td>11%</td>
</tr>
<tr>
<td>2</td>
<td>N</td>
<td>23%</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>15%</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>70%</td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td>91%</td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td>30%</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>11%</td>
</tr>
<tr>
<td>8</td>
<td>Y</td>
<td>75%</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>58%</td>
</tr>
<tr>
<td>10</td>
<td>N</td>
<td>27%</td>
</tr>
</tbody>
</table>

a. (1 point) Construct confusion matrices for discrimination thresholds of 0.50 and 0.25.
b. (1.5 points) Plot the Receiver Operating Characteristic (ROC) curve with the discrimination thresholds of 0.50 and 0.25.
   Label each axis and the coordinates and discrimination threshold of each point on the curve.
c. (0.5 point) Describe an advantage and a disadvantage of selecting a discrimination threshold of 0.25 instead of 0.50.
d. (0.5 point) Describe whether a discrimination threshold of 0.25 or 0.50 is more appropriate for a line of business with low frequency and high severity.
Three separate GLMs are fit using the following model form: \( g(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \)

The following error distributions were used for the three GLMs.

Each model also used their canonical link functions:
- Model I: gamma
- Model II: Poisson
- Model III: binomial

When fit to the data, all three models resulted in the same parameter estimates:

<table>
<thead>
<tr>
<th>( \hat{\beta}_0 )</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>1.0</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>-1.0</td>
</tr>
</tbody>
</table>

Determine the correct ordering of the models' predicted values at observed point \((X_1, X_2) = (2, 1)\).

A GLM was used to estimate the expected losses per customer across gender and territory. The following information is provided:
- The link function selected is log
- Q is the base level for Territory
- Male is the base level for Gender
- Interaction terms are included in the model

The GLM produced the following predicted values for expected loss per customer:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Territory</th>
<th>Territory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>148</td>
<td>545</td>
</tr>
<tr>
<td>Female</td>
<td>446</td>
<td>4,024</td>
</tr>
</tbody>
</table>

Calculate the estimated beta for the interaction of Territory R and Female.

You are given the following information about an insurance policy:
- The probability of a policy renewal, \( p(X) \), follows a logistic model with an intercept and one explanatory variable.
- \( \beta_0 = 5 \)
- \( \beta_1 = -0.65 \)

Calculate the odds of renewal at \( x = 5 \).
3.215. (8, 11/18, Q.5) (1.5 points)
An actuary is analyzing a partial residual plot of the driver age variable, which is shown below:

![Partial Residual Plot](image)

a. (1 point)
Adding polynomial terms is one approach to address the non-linearity in the driver age variable. Briefly describe two other alternative approaches and how they can be used to improve the fit of the driver age variable shown above.
b. (0.5 point) Briefly describe a downside to each of the two alternative approaches recommended in part a. above.

3.216. (8, 11/18, Q.6) (2.5 points) An actuary is comparing the output of two generalized linear models to develop a new rating plan for personal auto. Model statistics are shown below:

<table>
<thead>
<tr>
<th></th>
<th>Saturated Model</th>
<th>Model A</th>
<th>Model B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-Likelihood</td>
<td>-1,000</td>
<td>-1,500</td>
<td>-1,465</td>
</tr>
<tr>
<td>Estimated Dispersion Parameter</td>
<td>1.75</td>
<td>1.75</td>
<td>1.75</td>
</tr>
</tbody>
</table>

- Model A is a nested model of Model B, where Model B has an additional variable for driver age.
- Driver age is fit using a second-order polynomial.
- The critical value to be used from the F-distribution is 19.5.

a. (2 points) Using two statistical tests, recommend whether or not driver age should be included in the rating plan.
b. (0.5 point) Describe why the deviance statistic alone should not be used to assess model fit.
3.217. (8, 11/18, Q.7) (2.5 points) An actuary is planning to add a credit-based insurance score to a model that estimates the probability of a policy having a claim. The actuary has decided to offset all of the current model variables before fitting the new variable. Given the following:
- The current model (without the insurance score variable) is a logit-link binomial GLM (logistic regression).
- The logit link function is defined as \( g(\mu) = \ln\left( \frac{\mu}{1 - \mu} \right) \)
- The insurance score is a continuous variable having a value between 1 and 100.
- The current fitted values and insurance score for three policies as well as regression results from the fit of the insurance score variable are given below:

<table>
<thead>
<tr>
<th>Policy Number</th>
<th>Fitted Probability Without Insurance Score</th>
<th>Insurance Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.3%</td>
<td>78</td>
</tr>
<tr>
<td>2</td>
<td>20.3%</td>
<td>92</td>
</tr>
<tr>
<td>3</td>
<td>2.5%</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.250</td>
</tr>
<tr>
<td>Insurance Score</td>
<td>-0.020</td>
</tr>
</tbody>
</table>

a. (0.5 point) Calculate the offset term to be used in the regression for each of the three policies above.
b. (0.75 point) Calculate the revised fitted probability of having a claim for each of the three policies above.
c. (0.5 point) Identify the range of:
   i. the logit function
   ii. the logistic function
d. (0.25 point) Briefly explain why logistic regression is often used to model probabilities.
e. (0.5 point) Identify and briefly describe one situation in which the use of an offset is preferable to (re)fitting all variables.
Solutions:

3.1. Ignoring the loglikelihood of the saturated model, which is a constant, AIC = Deviance + (number of parameters)(2).
For example, AIC = 335.6 + (6)(2) = 347.6.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>335.60</td>
<td>347.60</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>331.90</td>
<td>347.90</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>325.20</td>
<td>345.20</td>
</tr>
<tr>
<td>D</td>
<td>12</td>
<td>321.40</td>
<td>345.40</td>
</tr>
<tr>
<td>E</td>
<td>14</td>
<td>317.00</td>
<td>345.00</td>
</tr>
</tbody>
</table>

Since AIC is smallest for model E, model E is preferred.

3.2. When using categorical variables, it is important to set the base level to be one with populous data, so that our measures of significance will be most accurate.
By choosing the base level to be one with lots of data, the estimates of the coefficients for the non-base levels are more stable.

3.3. This allows the scale of the predictors to match the scale of the entity they are linearly predicting, which in the case of a log link is the log of the mean of the outcome.
When a logged continuous predictor is placed in a log link model, the resulting coefficient becomes a power transform of the original variable. The coefficient $b_1$ becomes an exponent applied to the original variable $x_1$.
Including continuous predictors in their logged form allows a log link GLM flexibility in fitting the appropriate response curve. On the other hand, if the variable $x$ is not logged, the response curve for any positive coefficient will always have the same basic shape: exponential growth, that is, increasing at an increasing rate.

3.4. Frequently, the dataset going into a GLM will include records that represent the averages of the outcomes of groups of similar risks rather than the outcomes of individual risks.
In such instances, it is intuitive that records that represent a greater number of risks should carry more weight in the estimation of the model coefficients, as their outcome values are based on more data. GLMs accommodate that by allowing the user to include a weight variable, which specifies the weight given to each record in the estimation process.
The weight is the number of exposures for frequency or pure premium models.
For severity models, the weight is the number of claims.
The weight variable, usually denoted $\omega$, formally works its way into the math of GLMs as a modification to the assumed variance: $\text{Var}[y_i] = \frac{\phi V(\mu_i)}{\omega_i}$.
3.5. Determining accurate estimates of relativities in the presence of moderately correlated rating variables is a primary strength of GLMs versus univariate analyses. Unlike univariate methods, the GLM will be able to sort out each variable’s unique effect on the outcome, as distinct from the effect of any other variable that may correlate with it, thereby ensuring that no information is double-counted.

3.6. \( \exp(0.4) - 1 = 49.2\% \).
Comment: See page 25 of Goldburd, Khare, and Tevet.
For a logistic model: Odds = \( \mu / (1 - \mu) \).

3.7. Both are discrete distributions used to model frequency.
Both have support from zero to infinity. Both have \( \phi = 1 \).
The Negative Binomial Distribution has an additional parameter \( k > 0 \), called the overdispersion parameter.
The Poisson Distribution has variance function \( V(\mu) = \mu \), while the Negative Binomial Distribution has variance function \( V(\mu) = \mu(1 + k\mu) \). Thus the Negative Binomial Distribution has a variance greater than its mean, while the Poisson has a variance equal to its mean.
The Negative Binomial Distribution has a heavier righthand tail than the Poisson Distribution.
Comment: One way a Negative Binomial Distribution can arise is as a Gamma mixture of Poissons.

3.8. Where two predictors are perfectly correlated, they are said to be aliased, and the GLM will not have a unique solution.

3.9. 1. GLMs assign full credibility to the data.
2. GLMs assume that the randomness of outcomes are uncorrelated.

3.10. "Continuity in the Estimates is Not Guaranteed." Allowing each interval to move freely may not always be a good thing. The ordinal property of the levels of the binned variable have no meaning in the GLM; there is no way to force the GLM to have the estimates behave in any continuous fashion, and each estimate is derived independently of the others. Therefore, there is a risk that some estimates will be inconsistent with others due to random noise."
Variation within Intervals is Ignored. Since each bin is assigned a single estimate, the GLM ignores any variation that may exist within the bins.
Comment: See Section 5.4.2 of Goldburd, Khare, and Tevet.

3.11. The fitted parameter(s) are the same, while the standard errors are multiplied by \( \sqrt{7.9435} \).
The standard error of \( \hat{\beta}_1 \) is: 0.00120 \( \sqrt{7.9435} = 0.00338 \).
95% confidence interval for \( \beta_1 \): 0.02085 \( \pm (1.96) (0.00338) = 0.02085 \pm 0.00662 \).
Comment: One could instead use: 0.02085 \( \pm (2) (0.00338) = 0.02085 \pm 0.00676 \).
3.12. The Tweedie Distribution is an (linear) exponential family, used for modeling pure premiums. Besides the usual parameters $\mu$ and $\phi$, the Tweedie Distribution has a power parameter $p$. The variance function for Tweedie is $V(\mu) = \mu^p$. For use in GLMs we usually take $1 < p < 2$. The Tweedie Distribution can be represented as a compound Poisson with a Gamma severity. One rather interesting characteristic of the Tweedie distribution is that several of the other exponential family distributions are in fact special cases of Tweedie, dependent on the value of $p$.

3.13. It is clear that the proposed model more accurately predicts actual pure premium by decile than does the current rating plan. Specifically, consider the first decile. It contains the risks that the model thinks are best relative to the current plan. As it turns out, the model is correct. Similarly, in the 10th decile, the model more accurately predicts pure premium than does the current plan.

Comment: Graph taken from “Introduction to Predictive Modeling Using GLMs A Practitioner’s Viewpoint,” a presentation by Dan Tevet and Anand Khare.

3.14. The use of a log link results in the linear predictor, which begins as a series of additive terms, transforming into a series of multiplicative factors when deriving the model prediction. Multiplicative models are the most common type of rating structure used for pricing insurance, due to a number of advantages they have over other structures.

3.15. The sensitivity is: $\frac{\text{true positives}}{\text{total times there is an event}} = \frac{700}{1000} = 0.70$.

The specificity is: $\frac{\text{true negatives}}{\text{total times there is not an event}} = \frac{6000}{8000} = 0.75$.

For this threshold, we graph the point: $(1 - \text{specificity}, \text{sensitivity}) = (0.25, 0.70)$.

3.16. 1. Setting of objectives and goals. Determine the goals. Determine appropriate data to collect. Determine the time frame. What are key risks and how can they be mitigated? Who will work on the project; do they have the necessary knowledge and expertise?


3. Collecting and processing the necessary data for the analysis. Time-consuming. Data is messy. Often an iterative process. The data should also be split into at least two subsets, so that the model can be tested on data that was not used to build it. Formulate a strategy for validating the model.
4. Conducting exploratory data analysis (EDA). Spend some time to better understand the nature of the data and the relationships between the target and explanatory variables. Helpful EDA plots include:
Plotting each response variable versus the target variable to see what (if any) relationship exists.
Plotting continuous response variables versus each other, to see the correlation between them.

5. Specifying the form of the predictive model. What type of predictive model works best? What is the target variable, and which response variables should be included? Should transformations be applied to the target variable or to any of the response variables? Which link function should be used?

6. Evaluating the model output. Assessing the overall fit of the model. Identifying areas in which the model fit can be improved. Analyzing the significance of each predictor variable, and removing or transforming variables accordingly. Comparing the lift of a newly constructed model over the existing model or rating structure.


8. Translating the model results into a product. For GLMs, often the desired result is a rating plan. The product should be clear and understandable. Are there other rating factors included in the rating plan that were not part of the GLM?

9. Maintaining the model. Models should be periodically rebuilt in order to maximize their predictive accuracy, but in the interim it may be beneficial to merely refresh the existing model using newer data.

10. Rebuilding the model. More frequently one would update the classification relativities without updating the rating algorithm or classification definitions. Less frequently, one would do a more complete update, investigating changing the classification definitions, the predictor variables used, and/or the rating algorithm.

Comment: See Section 3 of Goldburd, Khare, and Tevet.
3.17. (a) Concentrate on one of the explanatory variables $X_j$.

The partial residuals are: (ordinary residual) $g'(\hat{\mu}_i) + x_{ij} \hat{\beta}_j$.

(b) In a Partial Residual Plot, we plot the partial residuals versus the variable of interest. If there seems to be curvature rather than linearity in the plot, that would indicate a departure from linearity between the explanatory variable of interest and $g(\mu)$, adjusting for the effects of the other independent variables.

3.18. The second model includes an interaction term.

In the second model, the effect of $X_1$ depends on the level of $X_2$ and vice-versa. In contrast, for the first model, the effects of $X_1$ and $X_2$ are independent.

3.19. **“Check for duplicate records.”** If there are any records that are exactly identical, this likely represents an error of some sort. This check should be done prior to aggregation and combination of policy and claim data.”

**“Cross-check categorical fields against available documentation.”** If data base documentation indicates that a roof can be of type A, B, or C, but there are records where the roof type is coded as D, this must be investigated. Are these transcription errors, or is the documentation out of date?”

**“Check numerical fields for unreasonable values.”** For every numerical field, there are ranges of values that can safely be dismissed as unreasonable, and ranges that might require further investigation. A record detailing an auto policy covering a truck with an original cost (new) of $30 can safely be called an error. But if that original cost is $5,000, investigation may be needed.”

**Comment:** Quoted from Section 4.2 of Goldburd, Khare, and Tevet.

**“Decide how to handle each error or missing value that is discovered.”** The solution to duplicate records is easy, delete the duplicates. But fields with unreasonable or impossible values that cannot be corrected may be more difficult to handle.”

3.20. 1. Plot each response variable versus the target variable, to see what if any relationship exists.

2. Plot continuous response variables versus each other, to see the correlation between them.

3.21. Advantages of the frequency/severity approach over pure premium modeling:

- Provides the actuary with more insight.
- Each of frequency and severity is more stable than pure premium.

Disadvantages of pure premium modeling versus the frequency/severity approach:

- Some interesting effects may go unnoticed.
- Pure premium modeling can lead to underfitting or overfitting.
- The Tweedie distribution used to model pure premium contains the implicit assumption that an increase in pure premiums is made up of an increase in both frequency and severity.
3.22. In general, an offset factor is a vector of known amounts which adjusts for known effects not otherwise included in the GLM. For example, one could take the current territories and territory relativities as givens, and include an offset term in a GLM for the current territory relativity. In general: 
$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \text{offset}.$$ 
An offset is used with a Poisson Distribution and a log link function, and there are exposures associated with each observation. In that case, the offset term is \(\ln(\text{exposure}) = \ln(n_i)\).

Then the model is: 
$$\ln(Y_i) = \ln(n_i) + \eta_i \iff Y_i = n_i \exp[\eta_i].$$

3.23. The observation for Slovakia has by far the biggest Cook's Distance, and is thus the most influential. The observations for the Czech Republic and Slovenia are less influential than Slovakia, but more influential than the others.

3.24. The saturated model has an equal number of predictors as there are records in the dataset. Since the saturated model predicts each record perfectly it is the theoretical best a model can possibly do. The null model has only an intercept and no predictors. The null model produces the same prediction for every record: the grand mean.

The deviance for the saturated model is zero, while the deviance of the null model can be thought of as the total deviance inherent in the data. The deviance for your model will lie between those two extremes.

3.25. The deviance residuals seem to be on average positive for small and large values of \(X_2\), while being on average negative for middle values of \(X_2\). Such a pattern is not good. This indicates that one should investigate other possible forms of the model, for example, a model including a term involving \(X_2^2\).

3.26. Adding credit score adds 6 - 1 = 5 parameters to the model. The test statistic is: 
$$F = \frac{D_S - D_B}{(\text{number of added parameters})^{-1}} = \frac{(233,183.65 - 233,134.37) / 5}{2.371} = 4.157.$$ 

The number of degrees of freedom in the numerator is 5. The number of degrees of freedom in the denominator is: number of observations minus the number of parameters in the smaller model 
= 100,000 - 10 = 99,990.

We compare the test statistic to an F-distribution with 5 and 99,990 degrees of freedom. We reject the null hypothesis when the F-Statistic is big.

Comment: Using a computer, the p-value of this test is 0.09%. Thus one would use the more complex model including credit score rather than the simpler model.
3.27. Arranged from smallest to largest: -0.328, -0.154, -0.064, 0.195, 0.239. Plot \((Q_{i/6}, x(i))\).

\[ Q_{1/6} = -0.967, \text{ since } \Phi[-0.967] = 1/6. \quad Q_{2/6} = -0.431. \quad Q_{3/6} = 0. \quad Q_{4/6} = 0.431. \quad Q_{5/6} = 0.967. \]

Thus the five plotted points are:
\((-0.967, -0.328), (-0.431, -0.154), (0, -0.064), (0.431, 0.195), (0.967, 0.239).\)

Here are the 5 points plotted:

There is too little data to decide whether or not these stock price returns are Normally distributed.

3.28. Gini index = \(2A\).

Comment: Gini index = \(\frac{\text{Area A}}{\text{Area A} + \text{Area B}}\).

However, Area A + Area B add up to a triangle with area 1/2.

Therefore, Gini index = \(\frac{\text{Area A}}{\text{Area A} + \text{Area B}} = 2A\)

= twice the area between the Lorenz Curve and the line of equality = 1 - 2B.

3.29. Factors for coverage options should be estimated outside the GLM, using traditional actuarial techniques. The resulting factors should then be included in the GLM as an offset.

3.30. \(\text{Min}[X - c, 0]\), where c is some constant and X is a variable.

For example, \(\text{Min}[X_2 - 13, 0]\) is a hinge function.

3.31. Model D is preferred. Bigger Area Under ROC Curve (AUROC) is better.
3.32. $35 \pm 1.96 \sqrt{5} = (30.62, 39.38)$.

3.33. BIC = $-2 \times \text{(maximum loglikelihood)} + \text{(number of parameters)} \ln[400]$.
For example, BIC = $-2 \times (-730.18) + 3 \ln[400] = 1478.33$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>Loglikelihood</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>-730.18</td>
<td>1478.33</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>-726.24</td>
<td>1476.45</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>-723.56</td>
<td>1477.08</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>-721.02</td>
<td>1477.99</td>
</tr>
<tr>
<td>E</td>
<td>7</td>
<td>-717.50</td>
<td>1476.94</td>
</tr>
</tbody>
</table>

Since BIC is smallest for model B, model B is preferred.

3.34. $\exp[-3.8 + (0.4)\ln[1/2]] = 1.7\%$.
Comment: Loosely based on Table 12 in Generalized Linear Models for Insurance Rating, by Goldburd, Khare and Tevet.

3.35. $\exp[-3.8 + 0.3 - 0.5 + (0.4)\ln[2.5/2] - (0.1)\ln[2.5/2]] = 2.0\%$.

3.36. $\exp[-3.8 + 0.5 + (0.4)\ln[3/2]] = 4.3\%$.

3.37. $\exp[-3.8 + 0.1 - 0.5 + (0.4)\ln[6/2] - (0.1)\ln[6/2]] = 2.1\%$.

3.38. Histogram A most closely matches the Normal Distribution.

3.39. a) Identity link function.
b) Log link function.
c) Poisson Distribution.
d) For the variance proportional to the square of the mean, use the Gamma Distribution.

3.40. The partial residual plot is not linear; thus, we should do something to improve the model. Since the slope seems to change somewhere around 50 or 60, we could use a hinge function: $\text{Min}[0, X - 50]$ or $\text{Min}[0, X - 60]$.
Comment: In general, we could instead group the variable, or add polynomial terms to the model.

3.41. We can divide the original data into three sets.
We fit GLMs to the training data, until we have one or more good candidate models. Then we see how these models perform on the validation set. Based on what we find out, we can go back and fit some other GLMs to the training data. The validation set is used to refine the models during the building process. The test set (holdout data) is held out until the end. We compare the performance of models on the test set to pick a final model to use.
### 3.42.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$V(\mu)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$\mu^0 = 1$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$\mu^1 = \mu$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$\mu^2$</td>
</tr>
<tr>
<td>Binomial (one trial)</td>
<td>$\mu (1-\mu)$</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>$\mu^3$</td>
</tr>
<tr>
<td>Tweedie</td>
<td>$\mu^p$, $p &lt; 0$, $1 &lt; p &lt; 2$, or $p &gt; 2$.</td>
</tr>
</tbody>
</table>

Alternately, for the Binomial Distribution, $V(\mu) = \mu (1 - \mu/m)$.

### 3.43.

$Q_{1/21} = -1.668$, since $\Phi[-1.668] = 1/21$.

Thus the first plotted point is: $(-1.668, 500)$.

The Q-Q Plot:
3.44. For a Poisson, \( f(n) = e^{-\lambda} \frac{\lambda^n}{n!} \).

\[
\ln(f(n)) = -\lambda + n\ln(\lambda) - \ln(n!) = -\exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + \eta_i(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) - \ln(n!).
\]

Loglikelihood:
\[
\log\text{likelihood} = -\sum \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + \sum Y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) + \text{constants}.
\]

Setting the partial derivatives of the loglikelihood with respect to \( \beta_0 \), \( \beta_1 \), and \( \beta_2 \) equal to zero:
\[
0 = -\sum \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + \sum Y_i.
\]
\[
0 = -\sum X_{1i} \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + \sum Y_i X_{1i}.
\]
\[
0 = -\sum X_{2i} \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + \sum Y_i X_{2i}.
\]

\[
\sum Y_i = 8 + 8 + 10 + \ldots + 33 + 31 = 369.
\]
\[
\sum Y_i X_{1i} = 8\ln(2) + 8\ln(4) + 10\ln(6) + \ldots + 33\ln(18) + 31\ln(20) = 872.856.
\]
\[
\sum Y_i X_{2i} = 14 + 19 + \ldots + 33 + 31 = 241.
\]
\[
\exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] = \exp[\beta_0] \exp[\beta_1 X_{1i}] \exp[\beta_2 X_{2i}] = \exp[\beta_0] \exp[X_{1i}]^\beta_1 \exp[X_{2i}]^\beta_2.
\]

The first equation becomes:
\[
\exp[\beta_0] \{2^{\beta_1} + 4^{\beta_1} + \ldots + 20^{\beta_1}\} \exp[\beta_2] + 4^{\beta_1} \exp[\beta_2] + 20^{\beta_1} \exp[\beta_2] = 369. \Rightarrow \exp[\beta_0] (1 + \exp[\beta_2]) \{2^{\beta_1} + 4^{\beta_1} + 6^{\beta_1} + \ldots + 20^{\beta_1}\} = 369.
\]

The second equation becomes:
\[
\exp[\beta_0] (1 + \exp[\beta_2]) \{\ln(2)2^{\beta_1} + \ln(4)4^{\beta_1} + \ln(6)6^{\beta_1} + \ldots + \ln(20)20^{\beta_1}\} = 872.856.
\]

The third equation becomes:
\[
\exp[\beta_0] \exp[\beta_2] \{2^{\beta_1} + 4^{\beta_1} + 6^{\beta_1} + \ldots + 20^{\beta_1}\} = 241.
\]

Comment: Well beyond what you should be asked on your exam!

A Poisson variable with a logarithmic link function.

Dividing the 1st and 3rd equations:
\[
(1 + \exp[\beta_2])/(\exp[\beta_2]) = 369/241. \Rightarrow \beta_2 = \ln(241/148) = 0.6328.
\]

Using a computer, the fitted parameters are: \( \beta_0 = 1.684, \beta_1 = 0.3784, \beta_2 = 0.6328. \)

One can verify that these values satisfy the three equations.

Example taken from *Applied Regression Analysis* by Draper and Smith.

3.45. While one may assume that the errors are Normally Distributed, in a GLM one could assume a different distribution of errors, such as Gamma or Poisson.

Thus Statement #1 is **not** true.

Statements #2 and #3 are true.
3.46. With four age categories, we add 4 - 1 = 3 parameters.

Test statistic is: 
\[ F = \frac{D_S - D_B}{(\text{number of added parameters}) \phi_S} = \frac{3320.2 - 3306.9}{(3) (1.83)} = 2.42. \]

The number of degrees of freedom in the numerator is 3.
The number of degrees of freedom in the denominator is: number of observations minus the number of parameters in the smaller model.
We compare the test statistic to the appropriate F-distribution.
We reject the null hypothesis if the test statistic is sufficiently big.

3.47. “Broadly speaking, model lift is the economic value of a model. The phrase “economic value” doesn’t necessarily mean the profit that an insurer can expect to earn as a result of implementing a model, but rather it refers to a model’s ability to prevent adverse selection. The lift measures … attempt to visually demonstrate or quantify a model’s ability to charge each insured an actuarially fair rate, thereby minimizing the potential for adverse selection.
Model lift is a relative concept, so it requires two or more competing models. That is, it doesn’t generally make sense to talk about the lift of a specific model, but rather the lift of one model over another.
In order to prevent overfitting, model lift should always be measured on holdout data.”

Comment: Quoted from Section 7.2 of Goldburd, Khare, and Tevet.

3.48. The effects of age and gender interact strongly. For example, the relationship between male and female relativities is very different for young drivers than it is for middle-aged drivers. In contrast, the effects of frequency of payment and age do not appear to interact significantly; there seems to be approximately the same relationship for each age group.

Comment: The graphs are adapted from “A Practitioner's Guide to Generalized Linear Models,“ by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi.

3.49. The second model is preferred since the predictions are closer to the actual than in Model 1.

Comment: See Figure 17 in Goldburd, Khare, and Tevet.
Let $X_1 = 1$ if age group A, and 0 otherwise.

$X_2 = 1$ if age group B, and 0 otherwise.

$X_3 = 1$ if small, and 0 otherwise.

$X_4 = 1$ if medium, and 0 otherwise.

$X_5 = 1$ if large, and 0 otherwise.

Then the design matrix is:

\[
\begin{pmatrix}
A/\text{small} \\
A/\text{medium} \\
A/\text{large} \\
B/\text{small} \\
B/\text{medium} \\
B/\text{large} \\
C/\text{small} \\
C/\text{medium} \\
C/\text{large}
\end{pmatrix} 
\begin{pmatrix}
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}
\]

For example, the first row corresponds to age group A and small: $X_1 = 1$, $X_2 = 0$, $X_3 = 1$, $X_4 = 0$, and $X_5 = 0$.

The last row corresponds to age group C and large: $X_1 = 0$, $X_2 = 0$, $X_3 = 0$, $X_4 = 0$, and $X_5 = 1$.

The vector of parameters is:

\[
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5
\end{pmatrix}
\]

Alternately, define medium and age group C as the base level.

Then the constant, $b_0$, would apply to all observations.

Let $X_1 = 1$ if age group A, and 0 otherwise.

$X_2 = 1$ if age group B, and 0 otherwise.

$X_3 = 1$ if small, and 0 otherwise.

$X_4 = 1$ if large, and 0 otherwise.

Then the design matrix is:
The first column of ones corresponds to the constant term which applies to all observations. For example, the first row corresponds to age group A and small: $X_0 = 1$, $X_1 = 1$, $X_2 = 0$, $X_3 = 1$, $X_4 = 0$.

The last row corresponds to age group C and large: $X_0 = 1$, $X_1 = 0$, $X_2 = 0$, $X_3 = 0$, $X_4 = 1$.

The vector of parameters is:

\[ \beta_0 \begin{cases} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{cases} \]

Comment: There is no unique answer. I have given two out of the many possible answers. There are 3 age categories and 3 size categories, so we need to have either $3 + 3 - 1 = 5$ covariates, or 4 covariates and a constant term.

The data would be arranged in a grid such as:

<table>
<thead>
<tr>
<th></th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age A</td>
<td>???</td>
<td>???</td>
<td>???</td>
</tr>
<tr>
<td>Age B</td>
<td>???</td>
<td>???</td>
<td>???</td>
</tr>
<tr>
<td>Age C</td>
<td>???</td>
<td>???</td>
<td>???</td>
</tr>
</tbody>
</table>

The response vector would have 9 rows and one column, containing the observations in the same order as the rows of the design matrix.
3.51. “If the modeler retains variables in the model that reflect a non-systematic effect on the response variable (i.e., noise) or over-specifies the model with high order polynomials, the result is over-fitting. Such a model will replicate the historical data very well (including the noise) but is not going to predict future outcomes reliably (as the future experience will most likely not have the same noise).
Conversely, if the model is missing important statistical effects (the extreme being a model that contains no explanatory variables and fits to the overall mean), the result is under-fitting. This model will predict future outcomes (e.g., in the extreme case mentioned above, the future mean) reliably but hardly help the modeler explain what is driving the result.”
“Considerable disparity between actual and expected results on the hold-out sample may indicate that the model is over or under-fitting.”
Underfit. ⇔ Too few Parameters. ⇔ Does not use enough of the useful information.
Overfit. ⇔ Too many Parameters. ⇔ Reflects too much of the noise.
In general, the actuary wants to avoid both underfitting and overfitting models.
Comment: See page 182 of Basic Ratemaking, on Exam 5.

3.52. The 18th observation has by far the biggest Cook’s Distance, and is thus the most influential.

3.53. (a) AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
For the first model, AIC = (-2)(-321.06) + (6)(2) = 654.12.
For the second model, AIC = (-2)(-319.83) + (7)(2) = 653.66.
Since AIC is smaller for the second model, the second model is preferred.
(b) BIC = (-2) (maximum loglikelihood)) + (number of parameters) ln(number of data points).
For the first model, BIC = (-2)(-321.06) + (6) ln[100] = 669.75.
For the second model, BIC = (-2)(-319.83) + (7) ln[100] = 671.90.
Since BIC is smaller for the first model, the first model is preferred.
Comment: An example where using AIC and BIC lead to different conclusions.

3.54. Graph B is closest to a straight line.
Comment: If the data was drawn from a Normal Distribution with \( m \neq 0 \), then we would expect the plotted points to be close to a straight line, but not a straight line through the origin.

3.55. (a) Poisson with log link function.
(b) Poisson or Negative Binomial with log link function.
(c) Gamma with log link function.
(d) Binomial with logit link function.
(e) Tweedie with log link function.
Comment: Claim frequency is claim count per exposure. If each insured has the same number of exposures, then a model of claim counts and claim frequency are mathematically equivalent.
3.56. Larger Gini index is better, all else being equal. The second rating plan is preferred. 
Comment: The higher the Gini index, the better the model is at identifying risk differences.

3.57. If the current rating plan were perfect, then all risks should have the same loss ratio. The fact that the proposed model is able to segment the data into lower and higher loss ratio buckets is a strong indicator that it is outperforming the current rating plan. 
Comment: Graph taken from “Introduction to Predictive Modeling Using GLMs A Practitioner’s Viewpoint,” a presentation by Dan Tevet and Anand Khare.

3.58. For levels 1 to 7 of the variable, the log of the multiplier is not significantly different than zero; in other words the relativity is not significantly different from one. Also for levels 1 to 7, there is no consistent pattern. Thus perhaps, levels 1 to 8 of this variable should be grouped into one level for purposes of the model; this would be treated as the new base. In contrast, for levels 9 to 15 there is pattern of increasing relativities. For levels 11 to 15 the relativities are significantly different from one. Given the pattern, one could also use the indicated relativities for levels 9 and 10.
Comment: As always, more testing may lead to a different conclusion. For example, it would be interesting to compare the results for different years of data to see if they are consistent. For example, the levels of the variable could be groups of annual income.
3.59. There are many ways to define the variables. Let us define $X_1 = 1$ if male and zero otherwise. $X_2 = 1$ if female and zero otherwise. $X_3 = 1$ if urban and zero otherwise.

For the Poisson, $f(x) = \lambda^x e^{-\lambda} / x!$. \ln f(x) = x \ln(\lambda) - \lambda - \ln(x!) = x \ln(\mu) - \mu - \text{constants}.$

(a) We use an identity link function. The estimated means are:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>$\beta_1 + \beta_3$</td>
<td>$\beta_1$</td>
</tr>
<tr>
<td>Female</td>
<td>$\beta_2 + \beta_3$</td>
<td>$\beta_2$</td>
</tr>
</tbody>
</table>

Ignoring constants, the loglikelihood is:

\[0.2 \ln(\beta_1 + \beta_3) - (\beta_1 + \beta_3) + 0.1 \ln(\beta_1) - (\beta_1) + 0.125 \ln(\beta_2 + \beta_3) - (\beta_2 + \beta_3) + 0.05 \ln(\beta_2) - (\beta_2)\]

Setting the partial derivative with respect to $\beta_1$ equal to zero:

\[0.2/(\beta_1 + \beta_3) + 0.1/\beta_1 = 2.\]

Setting the partial derivative with respect to $\beta_2$ equal to zero:

\[0.125/(\beta_2 + \beta_3) + 0.05/\beta_2 = 2.\]

Setting the partial derivative with respect to $\beta_3$ equal to zero:

\[0.2/(\beta_1 + \beta_3) + 0.125/(\beta_2 + \beta_3) = 2.\]

(b) We use an log link function. The estimated means are:

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>$\exp[\beta_1 + \beta_3]$</td>
<td>$\exp[\beta_1]$</td>
</tr>
<tr>
<td>Female</td>
<td>$\exp[\beta_2 + \beta_3]$</td>
<td>$\exp[\beta_2]$</td>
</tr>
</tbody>
</table>

Ignoring constants, the loglikelihood is:

\[0.2(\beta_1 + \beta_3) - \exp[\beta_1 + \beta_3] + 0.1\beta_1 - \exp[\beta_1] + 0.125(\beta_2 + \beta_3) - \exp[\beta_2 + \beta_3] + 0.05 \beta_2 - \exp[\beta_2].\]

Setting the partial derivative with respect to $\beta_1$ equal to zero:

\[\exp[\beta_1 + \beta_3] + \exp[\beta_1] = 0.3.\]

Setting the partial derivative with respect to $\beta_2$ equal to zero:

\[\exp[\beta_2 + \beta_3] + \exp[\beta_2] = 0.175.\]

Setting the partial derivative with respect to $\beta_3$ equal to zero:

\[\exp[\beta_1 + \beta_3] + \exp[\beta_2 + \beta_3] = 0.325.\]

Comment: Using a computer, the fitted parameters in part (a) are:

$\beta_1 = 0.105556$, $\beta_2 = 0.047500$, $\beta_3 = 0.084444$.

The fitted frequencies are: 0.1900, 0.1056, 0.1319, 0.0475.

Using a computer, the fitted parameters in part (b) are:

$\beta_1 = -2.35665$, $\beta_2 = -2.89565$, $\beta_3 = 0.77319$.

The fitted frequencies are: 0.2053, 0.0947, 0.1197, 0.0553.

3.60. The sensitivity is: $\frac{\text{true positives}}{\text{total times there is an event}} = 1800/3000 = 0.6$.

The specificity is: $\frac{\text{true negatives}}{\text{total times there is not an event}} = 40,000/50,000 = 0.8$.

For this threshold, we graph the point: $(1 - \text{specificity}, \text{sensitivity}) = (0.2, 0.6)$. 
3.61. \( f(y) = \exp[-(y - \mu)^2/(2\sigma^2)]/\sigma\sqrt{2\pi} \). \( \ln f(Y) = -(Y_i - \beta X_i)^2/(2\sigma^2) - \ln(\sigma) - \ln(2\pi)/2 \).

Loglikelihood is: \(-\sum(Y_i - \beta X_i)^2/(2\sigma^2) - n\ln(\sigma) - n\ln(2\pi)/2\).

Set the partial derivative of the loglikelihood with respect to \( b \) equal to zero:

\[
0 = \sum X_i (Y_i - \beta X_i)/\sigma^2.
\]

\[
\Rightarrow \sum X_i Y_i = \beta \sum X_i^2.
\]

\[
\hat{\beta} = \sum X_i Y_i / \sum X_i^2 = 3080/751 = 4.10.
\]

Comment: Matches the linear regression model with no intercept, \( \hat{\beta} = \sum X_i Y_i / \sum X_i^2 \).

3.62. Set the partial derivative of the loglikelihood with respect to \( s \) equal to zero:

\[
0 = \sum (Y_i - \beta X_i)^2/\sigma^3 - n/\sigma.
\]

\[
\Rightarrow \sigma^2 = \sum(Y_i - \beta X_i)^2/n = \frac{(5 - (1)(4.1))^2 + (15 - (5)(4.1))^2 + (50 - (10)(4.1))^2 + (100 - (25)(4.1))^2}{4} = 29.58.
\]

\[
\hat{\beta} = \sum X_i Y_i / \sum X_i^2.
\]

\[
\text{Var}[\hat{\beta}] = \text{Var}[\sum X_i Y_i / \sum X_i^2] = \sum \text{Var}[X_i Y_i / \sum X_i^2] = \sum X_i^2 \text{Var}[Y_i] / (\sum X_i^2)^2 = \sum X_i^2 \sigma^2 / (\sum X_i^2)^2 = 29.58/751 = 0.0394.
\]

\[
\text{StdDev}[\hat{\beta}] = \sqrt{0.0394} = 0.198.
\]

Comment: In the linear regression version of this same example, one would estimate the variance of the regression as:

\[
\sigma^2 = \frac{\sum \hat{\epsilon}^2}{N - 1} = \frac{(5 - (1)(4.1))^2 + (15 - (5)(4.1))^2 + (50 - (10)(4.1))^2 + (100 - (25)(4.1))^2}{4 - 1} = 39.4.
\]

This is an unbiased estimate of \( \sigma^2 \), which is not equal to that from maximum likelihood which is biased.

3.63. Estimated mean severity for a male in Territory D is: \( \exp[8.03 + 0.18 + 0.22] = 4583 \).

For the Inverse Gaussian Distribution, \( \text{Var}[Y] = \phi \mu^3 = (0.00510)(4583^3) = 490,930,199 \).

\( \text{StdDev}[Y] = \sqrt{490,930,199} = 22,157 \).

3.64. 1. Actuarial judgement. Does the model make sense; is the model reasonable.
2. Statistical Tests such as the F-Test.
3. Graph the modeled relativities plus or minus two standard errors.
   We would like the range between plus and minus two standard errors to be relatively narrow.
4. Check the consistency of the model run on different years of data.
5. Check the predictive accuracy of the model on a hold-out data set.

Comment: There are other possible answers.

3.65. The variance of the residuals appears to increasing with the fitted values, indicating heteroscedasticity (a lack of homoscedasticity.) This is not good, and one should try to refine the current model.
3.66. The deviance is equal to twice the difference between the maximum achievable loglikelihood (i.e., the loglikelihood where the fitted value is equal to the observation for every record) and the loglikelihood of the model. Alternately, the deviance is equal to twice the difference between the loglikelihood of the saturated model and the loglikelihood of the fitted model.

3.67. \( \ln[\mu/207] = 0.43 + 0.22 - 0.32 + 0.36 = 0.69. \)
\( \mu = 207 \exp[0.69] = \$413. \)
Comment: This is a multiplicative model with four categorical variables.

3.68. A model that combines information from two or more models is called an ensemble model. Two (or more) teams model the same item; they build separate models working independently. Combining the answers from both models is likely to perform better than either individually. A simple means of ensembling is to average the separate model predictions.

3.69. Test statistic is: \( F = \frac{D_S - D_B}{(\text{number of added parameters}) \phi_S} = \frac{(24,359 - 24,352)}{1} / 1.22 = 5.738. \)

The number of degrees of freedom in the numerator is 1.
The number of degrees of freedom in the denominator is:
number of observations minus the number of parameters in the smaller model
= 20,000 - 4 = 19,996.
This is equivalent to a two sided t-test at \( \sqrt{5.738} = 2.395, \) with 19,996 degrees of freedom.
Using the Normal approximation, the p-value is: (2) (1 - \( \Phi[2.395] \)).
Since 2.326 < 2.395 < 2.576, the two-sided p-value is between 2% and 1%.
Thus at a 2% significance level we should use the more complex model with the added variable, but at a 1% significance level we should use the simpler model without the additional variable.
Comment: Using a computer, the p-value of this test is 1.66%.
The null hypothesis is to use the simpler model. The alternate hypothesis is to use the more complex model. We reject the null hypothesis if the test statistic is sufficiently big.

3.70. The actuary would like the GLM to be stable; in other words, the predictions of the model should not be overly sensitive to small changes in the data.
An observation is influential if it has a large effect on the fitted model.
The larger the value of Cook’s distance, the more influential the observation.
The actuary should rerun the model excluding the most influential points to see their impact on the results. If this causes large changes in some of the parameter estimates, the actuary should consider for example whether to give these influential observations less weight.
Cross-validation can also be used to assess the stability of a GLM. A single model can be run on the set of folds. The results of the models fit to these different subsets of the data ideally should be similar. The amount by which these results vary is a measure of the stability of the model.
Bootstrapping via simulation can also be used to assess the stability of a GLM. The original data is randomly sampled with replacement to create a new set of data of the same size. One then fits the GLM to this new set of data. By repeating this procedure many times one can estimate the distribution of the parameter estimates of the GLM; we can estimate the mean, variance, confidence intervals, etc.
3.71. $\Phi[-1.645] = 1/20$. Thus for the given Normal, $Q_{0.05} = 1000 - (1.645)(300) = 506.5$.
The 19 plotted points are: (506.5, 258), (615.5, 636), (689.1, 652), (747.5, 814),
(797.7, 833), (842.7, 860), (884.4, 895), (924.0, 937), (962.3, 950), (1000.0, 1009),
(1037.7, 1020), (1076.0, 1059), (1115.6, 1103), (1157.3, 1113), (1202.3, 1127),
(1252.5, 1139), (1310.9, 1246), (1384.5, 1335), (1493.5, 1770).
The resulting Q-Q plot:
Alternately, one could standardize the data, by subtracting the sample mean of 987.158 and dividing by the square root of the sample variance of 96,057.8.

For example, \( \frac{(258 - 987.158)}{\sqrt{96,057.8}} = -2.353 \).

Then one compares the standardized data to the quantiles of the Standard Normal Distribution. (-1.645, -2.353), (-1.282, -1.133), (-1.036, -1.081), (-0.842, -0.559), (-0.674, -0.497), (-0.524, -0.410), (-0.385, -0.297), (-0.253, -0.162), (-0.126, -0.120), (0, 0.070), (0.126, 0.106), (0.253, 0.232), (0.385, 0.374), (0.524, 0.406), (0.674, 0.451), (0.842, 0.490), (1.036, 0.835), (1.282, 1.122), (1.645, 2.526).

The resulting Q-Q plot:

Comment: With the exception of the first and last plotted points, the points stay close to the 45 degree comparison line, indicating that this data may be normally distributed.
3.72. a) The total number of cells is: \((2)(4)(3) = 24\). So the design matrix would have 24 rows. Each row has a one in the first column; the intercept term applies to all insureds. For example, the first row has one in columns 3 and 6 corresponding to age 17-21 and Territory A.

\[
\begin{pmatrix}
1 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 \\
\end{pmatrix}
\]

F 17-21 A  
F 22-29 A  
F 30-59 A  
F 60+ A  
M 17-21 A  
M 22-29 A  
M 30-59 A  
M 60+ A  
F 17-21 B  
F 22-29 B  
F 30-59 B  
F 60+ B  
M 17-21 B  
M 22-29 B  
M 30-59 B  
M 60+ B  
F 17-21 C  
F 22-29 C  
F 30-59 C  
F 60+ C  
M 17-21 C  
M 22-29 C  
M 30-59 C  
M 60+ C

b) 30-59 year old female driver in Territory B is the base. Estimated frequency is \(\exp[\hat{\beta}_1]\).

c) For 22-29 year old male driver in Territory C, the estimated frequency is:
\[\exp[\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_4 + \hat{\beta}_7].\]

**Comment:** One can arrange the rows of the design matrix differently, as long as everything is consistent. Since there is an intercept term, and since each of the factors is a categorical variable, each has one less parameter than its number of levels.

We have chosen 30-59 year old female driver in Territory B as the base; some other choice could have been made.
3.73. AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
For example, AIC = (-2)(-359.17) + (3)(2) = 724.34.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>Loglikelihood</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>-359.17</td>
<td>724.34</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>-357.84</td>
<td>723.68</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>-356.42</td>
<td>722.84</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>-354.63</td>
<td>721.26</td>
</tr>
<tr>
<td>E</td>
<td>7</td>
<td>-353.85</td>
<td>721.70</td>
</tr>
</tbody>
</table>

Since AIC is smallest for model D, model D is preferred.

3.74. In the first graph, the relativities indicated by the separate years are similar to each other. Also the relativities for each year display a similar pattern of increase with vehicle symbol, which makes sense. Vehicle symbol appears to be a significant factor for the first model; it is likely to be a good predictor of future experience.
In the second graph, the relativities indicated by separate years are not consistent. Territory does not appear to be a significant factor for the second model.
Comment: The graphs are adapted from ones showing more information in Sections 2.40-2.41 of “A Practitioner’s Guide to Generalized Linear Models,” by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi, not on the syllabus.

3.75. If two predictors are highly correlated (have a correlation coefficient close to plus or minus one) coefficients may behave erratically. Furthermore, the standard errors associated with those coefficients will be large, and small perturbations in the data may swing the coefficient estimates wildly. Such instability in a model should be avoided. As such it is important to look out for instances of high correlation prior to modeling, by examining two-way correlation tables. Where high correlation is detected, means of dealing with this include the following:
• For any group of correlated predictors, remove all but one from the model.
• Preprocess the data using dimensionality reduction techniques such as principal component analysis.
Multicollinearity: A more subtle potential problem may exist where two or more predictors in a model may be strongly predictive of a third, a situation known as multicollinearity. The same instability problems as above may result. A useful statistic for detecting multicollinearity is the variance inflation factor (VIF), which can be output by most statistical packages. A common statistical rule of thumb is that a VIF greater than 10 is considered high.
Aliasing: Where two predictors are perfectly correlated, they are said to be aliased, and the GLM will not have a unique solution. Where they are nearly perfectly correlated, the model will be highly unstable; the fitting procedure may fail to converge, and even if the model run is successful the estimated coefficients will be nonsensical. Such problems can be avoided by looking out for and properly handling correlations among predictors, as discussed above.
Comment: See Section 2.9 of Goldburd, Khare, and Tevet.
Not necessary to say all of the above rather than some of the above.
3.76. 1. They are simple and practical to implement.
2. Having additive terms in a model can result in negative premiums, which doesn’t make sense. With a multiplicative plan you guarantee positive premium without having to implement clunky patches like minimum premium rules.
3. A multiplicative model has more intuitive appeal. It doesn’t make much sense to say that having a violation should increase your auto premium by $500, regardless of whether your base premium is $1,000 or $10,000. Rather it makes more sense to say that the surcharge for having a violation is 10%.
Comment: For these and other reasons, log link models, which produce multiplicative structures, are usually the most natural model for insurance risk.
“As for the link function, it is usually the case that the desirability of a multiplicative rating plan trumps all other considerations, so the log link is almost always used. One notable exception is where the target variable is binary (i.e., occurrence or non-occurrence of an event), for which a special link function (logistic) must be used.”

3.77. In order to incorporate age, avoiding aliasing, we need 6 - 1 = 5 variables. In order to incorporate gender, we would need one more variable for a total of 6. So getting rid of age and gender would produce a model with 6 fewer parameters.
Test statistic is: $F = \frac{D_S - D_B}{\text{(number of added parameters)}} = \frac{1128.1 - 1120.3}{0.395} / 6 = 3.291$.
The number of degrees of freedom in the numerator is 6.
The number of degrees of freedom in the denominator is: number of observations minus the number of parameters in the smaller model = 1000 - 44 = 956.
We compare the test statistic to an F-distribution with 6 and 956 degrees of freedom. The null hypothesis is to use the simpler model, the one without age and gender. The alternate hypothesis is to use the more complex model. We reject the null hypothesis if the test statistic is sufficiently big.
Comment: Using a computer, the p-value of this test is 3.3%.

3.78. Firstly, when comparing two models using log-likelihood or deviance, the comparison is valid only if the data sets used to fit the two models are exactly identical. If a new variable has missing values for some records, the default behavior of most model fitting software is to toss out those records when fitting the model. In that case, the resulting measures of fit are no longer comparable, since the second model was fit with fewer records than the first.
For any comparisons of models that use deviance it is also necessary that the assumed distribution and the dispersion parameter (basically, everything other than the coefficients) must be identical as well.
Comment: See Section 6.1.3 of Goldburd, Khare, and Tevet.
3.79. Age of spokesperson, gender of spokesperson, marital status of the spokesperson, time he has been a spokesperson, type of celebrity (actor, singer, athlete, etc.), criminal record of the spokesperson, past drug/alcohol abuse of the spokesperson, etc. 
**Comment:** There are other reasonable answers. This is often sold as death, disability, and disgrace insurance.

3.80. Both are continuous distributions used to model severity. Both are right-skewed, with a sharp peak and a long tail to the right, and a lower bound at zero. 
The Gamma Distribution has variance function \( V(\mu) = \mu^2 \), while the Inverse Gaussian Distribution has variance function \( V(\mu) = \mu^3 \). 
The Inverse Gaussian Distribution has a sharper peak and a wider tail than the Gamma Distribution. 
Therefore, the Inverse Gaussian Distribution is appropriate for situations where the skewness of the severity curve is more extreme. 
**Comment:** The skewness for the Gamma distribution is always twice times the coefficient of variation, while the skewness for the Inverse Gaussian distribution is always three times the coefficient of variation.

3.81. a) \( 9.5 + (0.01)(180) + (-0.02)(670) = -2.1 \). Using the inverse of the logit link function, the probability of default is: 
\[
\frac{\exp(-2.1)}{1 + \exp(-2.1)} = 10.9\%.
\]
b) \( 9.5 + (0.01)(100) + (-0.02)(760) = -4.7 \). Probability of default is: 
\[
\frac{\exp(-4.7)}{1 + \exp(-4.7)} = 0.9\%.
\]
**Comment:** Similar to 8, 11/12, Q.4a. Not intended as a realistic model.

3.82. The partial residual plot is not linear; thus, we should do something to improve the model. We could group the variable \( X_1 \), converting it into a categorical variable. 
We could add polynomial terms such \( X_1^2 \) to the model. 
We could use piecewise linear functions such as: \( \text{Min}[0, X_1 + 1] \) and \( \text{Min}[0, X_1 - 1] \).
3.83. There are many ways to define the variables. Let us define $X_1 = 1$ if low horsepower and zero otherwise. $X_2 = 1$ if medium horsepower and zero otherwise. $X_3 = 1$ if high horsepower and zero otherwise. $X_4 = 1$ if sedan and zero otherwise.

For the Gamma Distribution, $f(y) = \theta^{-\alpha} y^{\alpha-1} e^{-y/\theta} / \Gamma(\alpha)$. 

$$
\ln f(y) = (\alpha-1) \ln(y) - y/\theta - \alpha \ln(\theta) - \ln[\Gamma(\alpha)] = (\alpha-1) \ln(y) - y/\theta - \alpha \ln(\theta) - \ln[\Gamma(\alpha)]
$$

a) With the identity link function: $\mu = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$.

Ignoring terms that do not involve the betas, the loglikelihood is:

$$
-\alpha \frac{800}{(\beta_1 + \beta_4)} - \alpha \ln(\beta_1 + \beta_4) - \alpha \frac{900}{(\beta_2 + \beta_4)} - \alpha \ln(\beta_2 + \beta_4) - \alpha \frac{1100}{(\beta_3 + \beta_4)} - \alpha \ln(\beta_3 + \beta_4) - \alpha \frac{1500}{(\beta_1)} - \alpha \ln(\beta_1) - \alpha \frac{1700}{(\beta_2)} - \alpha \ln(\beta_2) - \alpha \frac{2000}{(\beta_3)} - \alpha \ln(\beta_3).
$$

Setting the partial derivative with respect to $\beta_1$ equal to zero:

$$
800/(\beta_1 + \beta_4)^2 + 1500/\beta_1^2 = 1/(\beta_1 + \beta_4) + 1/\beta_1.
$$

Setting the partial derivative with respect to $\beta_2$ equal to zero:

$$
900/(\beta_2 + \beta_4)^2 + 1700/\beta_2^2 = 1/(\beta_2 + \beta_4) + 1/\beta_2.
$$

Setting the partial derivative with respect to $\beta_3$ equal to zero:

$$
1100/(\beta_3 + \beta_4)^2 + 2000/\beta_3^2 = 1/(\beta_3 + \beta_4) + 1/\beta_3.
$$

Setting the partial derivative with respect to $\beta_4$ equal to zero:

$$
800/(\beta_1 + \beta_4)^2 + 900/(\beta_2 + \beta_4)^2 + 1100/(\beta_3 + \beta_4)^2 = 1/(\beta_1 + \beta_4) + 1/(\beta_2 + \beta_4) + 1/(\beta_3 + \beta_4).
$$
(b) We use a log link function. \( \mu = \exp[\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4] \).

Ignoring terms that do not involve the betas, the loglikelihood is:

\[
-\alpha 800 \exp[-\beta_1 - \beta_4] - \alpha (\beta_1 + \beta_4) - \alpha 900 \exp[-\beta_2 - \beta_4] - \alpha (\beta_2 + \beta_4) - \alpha 1100 \exp[-\beta_3 + \beta_4] - \alpha (\beta_3 + \beta_4) \\
- \alpha 1500 \exp[-\beta_1] - \alpha (\beta_1) - \alpha 1700 \exp[-\beta_2] - \alpha (\beta_2) - \alpha 2000 \exp[-\beta_3] - \alpha (\beta_3).
\]

Setting the partial derivative with respect to \( \beta_1 \) equal to zero:

\[800 \exp[-\beta_1 - \beta_4] + 1500 \exp[-\beta_1] = 2.\]

Setting the partial derivative with respect to \( \beta_2 \) equal to zero:

\[900 \exp[-\beta_2 - \beta_4] + 1700 \exp[-\beta_2] = 2.\]

Setting the partial derivative with respect to \( \beta_3 \) equal to zero:

\[1100 \exp[-\beta_3 - \beta_4] + 2000 \exp[-\beta_3] = 2.\]

Setting the partial derivative with respect to \( \beta_4 \) equal to zero:

\[800 \exp[-\beta_1 - \beta_4] + 900 \exp[-\beta_2 - \beta_4] + 1100 \exp[-\beta_3 - \beta_4] = 3.\]

**Comment:** Using a computer, the fitted parameters in part (a) are:

\( \beta_1 = 1567.71, \beta_2 = 1688.03, \beta_3 = 1914.41, \beta_4 = -784.60. \)

The fitted severities are: 783.11, 903.43, 1129.81, 1567.71, 1688.03, 1914.41.

Using a computer, the fitted parameters in part (b) are:

\( \beta_1 = 7.30933, \beta_2 = 7.43082, \beta_3 = 7.61246, \beta_4 = -0.620811.\)

The fitted severities are: 803.13, 906.88, 1087.51, 1494.17, 1687.20, 2023.24.

### 3.84. Adding vehicle type adds 10 - 1 = 9 parameters to the model.

Test statistic is: \( F = \frac{D_S - D_B}{(\text{number of added parameters}) \phi_S} = \frac{(1848.5 - 1833.0)}{9} = 1.852. \)

The number of degrees of freedom in the numerator is 9.
The number of degrees of freedom in the denominator is:

number of observations minus the number of parameters in the smaller model = 2000 - 14 = 1986.

We compare the test statistic to an F-distribution with 9 and 1986 degrees of freedom.
The null hypothesis is to use the simpler model, the one without vehicle type.
The alternate hypothesis is to use the more complex model.
We reject the null hypothesis at 5% if the test statistic is bigger than the 5% critical value, which is where the F-distribution is 95%.

**Comment:** Using a computer, the p-value of this test is 5.5%.
Thus we would not reject the null hypothesis at 5%.
If we reduced the number of vehicle type categories by combining some of the 10 categories we used, it might turn out that now we should use vehicle type at the 5% significance level.
3.85. Ignoring the loglikelihood of the saturated model, which is a constant, \( \text{BIC} = \text{Deviance} + (\text{number of parameters}) \ln[250] \).
For example, \( \text{BIC} = 1679.1 + 6 \ln[250] = 1712.23 \).

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>Deviance</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>1679.10</td>
<td>1712.23</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>1666.40</td>
<td>1710.57</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>1655.90</td>
<td>1711.11</td>
</tr>
<tr>
<td>D</td>
<td>12</td>
<td>1646.20</td>
<td>1712.46</td>
</tr>
<tr>
<td>E</td>
<td>14</td>
<td>1634.50</td>
<td>1711.80</td>
</tr>
</tbody>
</table>

Since BIC is smallest for model B, model B is preferred.

3.86. The difference between the yellow univariate line and the green GLM line, which better represents the underlying reality, arises from correlation between policy duration shown in the graph and the two other factors in the model.
Comment: One does not have to understand the life insurance details in order to answer the question asked.

3.87. \( \exp[-0.3] - 1 = -25.9\% \).
Comment: See page 25 of Goldburd, Khare, and Tevet.
For a logistic model: Odds = \( \mu / (1 - \mu) \).

3.88. Female drivers age 31 to 59 in a rural territory have lower (process) variances than unmarried male drivers age 17 to 21 in an urban territory.
Therefore, the fitted model shifts to agree more closely with the observed values for the first group compared to the second group.
A GLM is more concerned with differences between observed and fitted where the (process) variances in observations are smaller. A GLM is less concerned with differences between observed and fitted where the variances in observations are larger.

3.89. The sensitivity is: \( 2000/5000 = 0.40 \).
The specificity is: \( 70,000 / 80,000 = 0.875 \).
For this threshold, we graph the point: \( (1 - \text{specificity}, \text{sensitivity}) = (0.125, 0.40) \).

3.90. “Two standard errors from the parameter estimates are akin to a 95% confidence interval. This means the GLM parameter estimate is a point estimate, and the standard errors show the range in which the modeler can be 95% confident the true answer lies within.”
Comment: See page 179 of Basic Ratemaking, on Exam 5.
3.91. The Lorenz curve for the rating plan is determined as follows:
1. Sort the dataset based on the model predicted loss cost.
2. On the x-axis, plot the cumulative percentage of exposures.
3. On the y-axis, plot the cumulative percentage of losses.
Draw a 45-degree line connecting (0, 0) and (1, 1), called the line of equality.
The Gini index is twice the area between the Lorenz curve and the line of equality.

3.92. \( \ln(\lambda) = \beta_0 + \beta_1 z. \implies \lambda = \exp[\beta_0 + \beta_1 z]. \)

For the Poisson Distribution: \( f(y) = e^{-\lambda} \frac{\lambda^y}{y!}. \)
\[ \ln f(y) = -\lambda + y \ln(\lambda) - \ln(y!) = -\exp[\beta_0 + \beta_1 z] + y(\beta_0 + \beta_1 z) - \ln(y!). \]
The loglikelihood is the sum of the contributions from the three observations:
\[ -\exp[\beta_0 + \beta_1] - \exp[\beta_0 + 2\beta_1] - \exp[\beta_0 + 3\beta_1] + 4(\beta_0 + \beta_1) + 7(\beta_0 + 2\beta_1) + 8(\beta_0 + 3\beta_1) - \ln(4!) - \ln(7!) - \ln(8!). \]
To maximize the loglikelihood, we set its partial derivatives equal to zero.
Setting the partial derivative with respect to \( \beta_0 \) equal to zero:
\[ 0 = -\exp[\beta_0 + \beta_1] - \exp[\beta_0 + 2\beta_1] - \exp[\beta_0 + 3\beta_1] + 19. \]
Setting the partial derivative with respect to \( \beta_1 \) equal to zero:
\[ 0 = -\exp[\beta_0 + \beta_1] - 2\exp[\beta_0 + 2\beta_1] - 3\exp[\beta_0 + 3\beta_1] + 42. \]
Thus we have two equations in two unknowns:
\[ \exp[\beta_0 + \beta_1](1 + \exp[\beta_1] + \exp[2\beta_1]) = 19. \]
\[ \exp[\beta_0 + \beta_1](1 + 2\exp[\beta_1] + 3\exp[2\beta_1]) = 42. \]
Dividing the second equation by the first equation:
\[ \frac{1 + 2\exp[\beta_1] + 3\exp[2\beta_1]}{1 + \exp[\beta_1] + \exp[2\beta_1]} = \frac{42}{19}. \]
\[ \Rightarrow 19 + 38\exp[\beta_1] + 57\exp[2\beta_1] = 42 + 42\exp[\beta_1] + 42\exp[2\beta_1]. \]
\[ \Rightarrow 15\exp[2\beta_1] - 4\exp[\beta_1] - 23 = 0. \]
Letting \( v = \exp[\beta_1] \), this equation is: \( 15v^2 - 4v - 23 = 0 \), with positive solution:
\[ v = \frac{4 + \sqrt{1396}}{30} = 1.3788. \]
\[ \exp[\beta_1] = 1.3788. \Rightarrow \beta_1 = 0.3212. \]
\[ \Rightarrow \exp[\beta_0] = 19/(\exp[\beta_1] + \exp[2\beta_1] + \exp[3\beta_1]) = 19/(1.3788 + 1.3788^2 + 1.3788^3) = 3.2197. \]
\[ \Rightarrow \beta_0 = 1.1693. \]
\[ \lambda = \exp[\beta_0 + \beta_1 z] = \exp[\beta_0] \exp[\beta_1]^z = (3.2197)(1.3788^z). \]
For \( z = 1 \), \( \lambda = 4.439 \). For \( z = 2 \), \( \lambda = 6.121 \). For \( z = 3 \), \( \lambda = 8.440 \).
Comment: An ordinary linear regression fit to these same observations turns out to be:
\[ y = 2.333 + 2x, \] with fitted values: 4.333, 6.333, and 8.333.
3.93. Examples include:
- Will it be cost-effective to collect the value of this variable when writing new and renewal business?
- Does inclusion of this variable in a rating plan conform to actuarial standards of practice and regulatory requirements?
- Can the electronic quotation system be easily modified to handle the inclusion of this variable in the rating formula?

3.94. a. We would have one parameter for gender, two parameters for age, and two parameters for territory. In addition we would have a parameter related to the base level.
A total of 6 parameters.
\[(2-1) + (3-1) + (3-1) + 1 = 6.\]
Sex   Age   Terr.   Base
b. A total of 6 parameters. The link function does not affect the number of parameters.
c. \(\beta_0\) is the intercept term that applies to all insureds.
\(\beta_1\) corresponds to Female.
\(\beta_2\) corresponds to Youthful.
\(\beta_3\) corresponds to Retired.
\(\beta_4\) corresponds to Suburban.
\(\beta_5\) corresponds to Rural.
(There are many other possible orders for the parameters.)
d. With 6 parameters, the design matrix has 6 columns.
e. With 20,000 cars, the design matrix has 20,000 rows.
f. The number combinations are: \((2)(3)(3) = 18\). Thus the design matrix has 18 rows.
(I have assumed that none of these cells is empty.
I have assumed that there are no records with missing classification information.)

3.95. (a) The deviance residual for any given record is defined as that record’s contribution to the deviance, adjusted for the sign of actual minus predicted; the deviance residual is taken to be negative where actual is less expected, and positive where actual is more than expected. (Actually the deviance residual is the square root of the contribution to the deviance.)
(b) Intuitively, we can think of the deviance residual as the residual adjusted for the shape of the assumed GLM distribution, such that its distribution will be approximately normal if the assumed GLM distribution is correct.
(c) In a well-fit model, we expect deviance residuals to follow no predictable pattern, and be normally distributed, with constant variance.
One could plot the deviance residuals versus the fitted values or versus an important predictor variable, in order to see whether there is a pattern.
We can check for the normality of the deviance residuals via either a histogram or q-q plot.
\( p/(1-p) = \exp[\beta_0 + \beta_1 X]. \Rightarrow 1/p - 1 = \exp[-\beta_0 - \beta_1 X]. \Rightarrow p = 1/ (1 + \exp[-\beta_0 - \beta_1 X]). \nabla \)

\( 1 - p = \exp[\beta_0 - \beta_1 X] / (1 + \exp[-\beta_0 - \beta_1 X]) = 1 / (1 + \exp[\beta_0 + \beta_1 X]). \nabla \)

For a Binomial with parameters \( m \) and \( p \),
\[ f(n) = p^n(1-p)^{m-n} m! / {(n!)(m-n)!}. \]

\[ \ln f(n) = n \ln p + (m-n)\ln(1-p) + \ln(m!) - \ln(n!) - \ln[(m-n)!] = n \ln[p/(1-p)] + m \ln(1-p) + \text{constants} = n(\beta_0 + \beta_1 X) - m \ln[1 + \exp[\beta_0 + \beta_1 X]] + \text{constants}. \]

Loglikelihood = \( \sum n_i(\beta_0 + \beta_1 X_i) - \sum m_i \ln[(1 + \exp[\beta_0 + \beta_1 X_i])] + \text{constants}. \)

Setting the partial derivatives of the loglikelihood with respect to \( \beta_0 \) and \( \beta_1 \) equal to zero:
\[ \begin{align*}
 0 &= \sum n_i - \sum m_i \exp[\beta_0 + \beta_1 X_i]/(1 + \exp[\beta_0 + \beta_1 X_i]). \\
 0 &= \sum n_i X_i - \sum m_i X_i \exp[\beta_0 + \beta_1 X_i]/(1 + \exp[\beta_0 + \beta_1 X_i]).
\end{align*} \]

\[ \begin{align*}
 0 &= \sum n_i = 900 + 820 + 740 + 660 + 580 = 3700. \\
 10300 &= \sum n_i X_i = (1)(900) + (2)(820) + (3)(740) + (4)(660) + (5)(580) = 10,300.
\end{align*} \]

The first equation becomes:
\[ 3700 = 1000/(1 + \exp[-\beta_0 - \beta_1]) + 900/(1 + \exp[-\beta_0 - 2\beta_1]) + 800/(1 + \exp[-\beta_0 - 3\beta_1]) + 700/(1 + \exp[-\beta_0 - 4\beta_1]) + 600/(1 + \exp[-\beta_0 - 5\beta_1]). \]

The second equation becomes:
\[ 10300 = 1000/(1 + \exp[-\beta_0 - \beta_1]) + 1800/(1 + \exp[-\beta_0 - 2\beta_1]) + 2400/(1 + \exp[-\beta_0 - 3\beta_1]) + 2800/(1 + \exp[-\beta_0 - 4\beta_1]) + 3000/(1 + \exp[-\beta_0 - 5\beta_1]). \]

Comment: An example of a Logistic Regression.
Using a computer, the maximum likelihood fit is: \( \beta_0 = 1.88543 \) and \( \beta_1 = 0.245509 \).

The covariance matrix of the fitted parameters is:
\[ \begin{pmatrix}
  \beta_0 & 0.0154836 & -0.00501396 \\
 0.0154836 & -0.00501396 & 0.00212092
\end{pmatrix}. \]

Thus the standard error of \( \beta_0 \) is: \( \sqrt{0.0154836} = 0.1244 \),
and the standard error of \( \beta_1 \) is: \( \sqrt{0.00212092} = 0.04605 \).
Here is a graph of the data, the fitted curve, and 95% confidence intervals:

3.97. $\phi$ is the dispersion parameter, which scales the variance.
$\omega_i$ is a (prior) weight, representing the amount of data we have for observation $i$; the variance is inversely proportional to the volume of data.

3.98. & 3.99. $f(y) = \exp[-(y - \mu)^2/(2\sigma^2)] / \{\sigma \sqrt{2\pi}\}$.

$\ln f(Y_i) = -(Y_i - \beta_0 - \beta_1 X_i)^2/(2\sigma^2) - n \ln(\sigma) - n \ln(2\pi)/2.$

Loglikelihood is: $-\sum(Y_i - \beta_0 - \beta_1 X_i)^2/(2\sigma^2) - n \ln(\sigma) - n \ln(2\pi)/2$.

Set the partial derivative of the loglikelihood with respect to $\beta_0$ equal to zero:

$0 = \sum(Y_i - \beta_0 - \beta_1 X_i)/\sigma^2. \Rightarrow \sum Y_i = n\beta_0 + \beta_1 \sum X_i; \Rightarrow \beta_0 = \bar{Y} - \beta_1 \bar{X}.$

Set the partial derivative of the loglikelihood with respect to $\beta_1$ equal to zero:

$0 = \sum X_i(Y_i - \beta_0 - \beta_1 X_i)/\sigma^2. \Rightarrow \sum X_i Y_i = \beta_0 \sum X_i + \beta_1 \sum X_i^2. \Rightarrow \sum X_i Y_i = (\bar{Y} - \beta_1 \bar{X}) \sum X_i + \beta_1 \sum X_i^2.$

$\Rightarrow \hat{\beta}_1 = \{\sum X_i Y_i - \bar{Y} \sum X_i\} / \{\sum X_i^2 - \bar{X} \sum X_i\} = (255 - (10)(24)) / (174 - (6)(24)) = 15/30 = 0.5.$

$\Rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 10 - (0.5)(6) = 7.$

Comment: Matches the linear regression model with an intercept.
For example, in deviations form:

$\bar{X} = 24/4 = 6. \quad x = X - \bar{X} = -4, -1, 2, 3. \quad \bar{Y} = 40/4 = 10. \quad y = Y - \bar{Y} = 0, -4, 1, 3.$

$\hat{\beta} = \sum x_i y_i / \sum x_i^2 = 15/30 = 0.5. \quad \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} = 10 - (0.5)(6) = 7.$
3.100. Set the partial derivative of the loglikelihood with respect to \( \sigma \) equal to zero:

\[
0 = \sum (Y_i - \beta_0 - \beta_1 X_i)^2 / \sigma^3 - n / \sigma.
\]

\[\Rightarrow \sigma^2 = \sum (Y_i - \beta_0 - \beta_1 X_i)^2 / n = \sum (Y_i - 7 - (0.5)X_i)^2 / 4 = (10 - 7 - (0.5)(2))^2 + (6 - 7 - (0.5)(5))^2 + (11 - 7 - (0.5)(8))^2 + (13 - 7 - (0.5)(9))^2 = 18.5 / 4 = 4.625.
\]

\[\Rightarrow \hat{\sigma} = \sqrt{4.625} = 2.15.
\]

3.101. & 3.102. Let \( x = X - \bar{X} = -4, -1, 2, 3 \), and \( y = Y - \bar{Y} = 0, -4, 1, 3 \).

Then,

\[
\sum X_i Y_i - \bar{Y} \sum X_i = \sum X_i Y_i - \sum Y_i \sum X_i / n = \sum Y_i (X_i - \bar{X}) = \sum Y_i x_i.
\]

Also,

\[
\sum X_i^2 - \bar{X} \sum X_i = \sum X_i (X_i - \bar{X}) = \sum X_i^2 + \bar{X} \sum x_i = \sum x_i^2 + \bar{X} (0)
\]

\[= \sum x_i^2.
\]

\[\hat{\beta}_1 = \{\sum x_i y_i - \bar{Y} \sum x_i\} / \{\sum x_i^2 - \bar{X} \sum x_i\} = \sum Y_i x_i / \sum x_i^2.
\]

\[\text{Var}[\hat{\beta}_1] = \text{Var}[\sum Y_i x_i / \sum x_i^2] = \sum \text{Var}[Y_i x_i] / (\sum x_i^2)^2 = \sum x_i^2 \text{Var}[Y_i] / (\sum x_i^2)^2 = \sigma^2 / \sum x_i^2
\]

\[= 4.625 / 30 = 0.1542. \text{ StdDev}[\hat{\beta}_1] = \sqrt{0.1542} = 0.393.
\]

\[\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = (Y_1 + Y_2 + Y_3 + Y_4) / 4 - (\sum Y_i x_i / \sum x_i^2)(6) = (Y_1 + Y_2 + Y_3 + Y_4) / 4 - (-4Y_1 - Y_2 + 2Y_3 + 3Y_4)(6/30) = 1.05Y_1 + 0.45Y_2 - 0.15Y_3 - 0.35Y_4.
\]

Recalling that the \( Y_i \) are independent and each have variance \( \sigma^2 \):

\[\text{Var}[\hat{\beta}_0] = \sigma^2 (1.05^2 + 0.45^2 + 0.15^2 + 0.35^2) = 1.45\sigma^2 = (1.45)(4.625) = 6.706.
\]

\[\text{StdDev}[\hat{\beta}_0] = \sqrt{6.706} = 2.59.
\]

Comment: Beyond what you should be asked on your exam.

One can show in general that \( \text{Var}[\hat{\beta}] = \sigma^2 \sum x_i^2 \) and \( \text{Var}[\hat{\alpha}] = \sigma^2 \sum x_i^2 / (N \sum x_i^2) \).

While the maximum likelihood results are similar, they do not match linear regression:

\[
\hat{Y} = \hat{\alpha} + \hat{\beta} X = 8, 9.5, 11, 11.5. \quad \hat{\epsilon} = Y - \hat{Y} = 2, -3.5, 0, 1.5. \quad \text{ESS} = \sum \hat{\epsilon}_i^2 = 18.5.
\]

\[s^2 = \text{ESS} / (N - 2) = 18.5 / (4 - 2) = 9.25.
\]

\[\text{Var}[\hat{\beta}] = s^2 / \sum x_i^2 = 9.25 / 30 = 0.3083. \quad s_\beta = \sqrt{0.3083} = 0.555.
\]

\[\text{Var}[\hat{\alpha}] = s^2 \sum x_i^2 / (N \sum x_i^2) = (9.25)(174) / ((4)(30)) = 13.41. \quad s_\alpha = \sqrt{13.41} = 3.66.
\]
3.103. Since they have all been fit to the same data and have the same number of parameters, the model with the smallest Deviance is best. 
Comment: If we were to apply either AIC or BIC, in this case the ranks of the models would be the same as that of their deviances.

3.104. When the variance is greater than the mean, we can use an overdispersed Poisson with $\phi > 1$. 
$\text{Var}[Y_i] = \phi \text{E}[Y_i]$. For $\phi > 1$, variance is greater than the mean. While this does not correspond to the likelihood of any exponential family, otherwise the GLM mathematics works.
Using an overdispersed Poisson (ODP), we get the same estimated betas as for the usual Poisson regression. However, the standard errors of all of the estimated parameters are multiplied by $\sqrt{\phi}$.
Comment: When the variance is greater than the mean, one could use a Negative Binomial Distribution, which has a variance greater than its mean.
Often the results of using an overdispersed Poisson and a Negative Binomial will be similar.

3.105. While a 5% probability value may seem small, it allows for a 1-in-20 chance of a variable being accepted as significant when it is not. Since in a typical insurance modeling project we are testing many variables, this threshold may be too high to protect against the possibility of spurious effects making it into the model.
For example, if we are testing the potential usefulness of 40 possible predictor variables, then if we use a p-value of 5%, even if none of the variables actually predict the outcome, on average two of these 40 variables will be selected as significant.
Comment: See page 9 of “Generalized Linear Models for Insurance Rating”.
“Spurious correlations exist when the historical correlation between two variables is random or coincidental. In these cases, one variable cannot reliably be used to inform a projection of the other variable going forward. For example, over the past year the number of California Department of Insurance rate regulation actuaries has increased, as has California average rainfall. Unfortunately, however, we cannot expect to influence future California rainfall by hiring additional actuaries.”

3.106. The partial residual plot seems linear; thus, no action is indicated.

3.107. A potential problem may exist where two or more predictors in a model may be strongly predictive of a third, a situation known as multicollinearity. Instability problems may result, since the information contained in the third variable is also present in the model in the form of the combination of the other two variables. However, the variable may not be highly correlated with either of the other two predictors individually, and so this effect will not show up in a correlation matrix, making it more difficult to detect.
3.108. Territories are not a good fit for the GLM framework. One should include the territory relativities produced by the separate model as an offset in the GLM used to determine classification relativities. Similarly, one should include classification relativities produced by the GLM as an offset in the separate model used to determine territory relativities.

3.109. A hold-out sample is data that was not used in the development of the model so that it could be used to test the effectiveness of the model. (This could either be a random sample of the original data, or an additional year of data.) One compares the expected outcome of the model with results on the hold-out sample. The extent to which the model results track closely to results on the hold-out sample for a large part of the portfolio is an indication of how well the model validates.

3.110. You can use age groups, but probably want to group fewer ages together for the younger ages. (Unfortunately, the volume of data is smaller for the very youngest ages, so there is a trade-off between homogeneity and credibility.) For ages above about 25, the affect of gender is relatively small and similar by age. In contrast, for younger ages the affect of gender is large and differs by age. Thus a simple multiplicative model with a single relativity for male compared to female will not work. One would need to have a gender relativity that varied by age. (This may be possible to accomplish this by having an interaction term in the GLM.)

3.111. (a) For $z_1 = 1$ and $z_2 = 30$, renewal probability is:

$$\frac{\exp[0.6 + (0.05)(1) + (0.02)(30)]}{1 + \exp[0.6 + (0.05)(1) + (0.02)(30)]} = 0.7773.$$ 

For $z_1 = 10$ and $z_2 = 30$, renewal probability is:

$$\frac{\exp[0.6 + (0.05)(10) + (0.02)(30)]}{1 + \exp[0.6 + (0.05)(10) + (0.02)(30)]} = 0.8455.$$ 

$0.7773 / 0.8455 = 0.919.$

(b) For $z_1 = 1$ and $z_2 = 50$, renewal probability is:

$$\frac{\exp[0.6 + (0.05)(1) + (0.02)(50)]}{1 + \exp[0.6 + (0.05)(1) + (0.02)(50)]} = 0.8389.$$ 

For $z_1 = 10$ and $z_2 = 50$, renewal probability is:

$$\frac{\exp[0.6 + (0.05)(10) + (0.02)(50)]}{1 + \exp[0.6 + (0.05)(10) + (0.02)(50)]} = 0.8909.$$ 

$0.8389 / 0.8909 = 0.942.$

Comment: Not intended as a realistic model of policy renewal. In general for a particular GLM, the relativities for one predictor variable can depend on the level(s) of the other predictor variable(s).

This model was based on the logit link function. If instead the log link function had been used, the model would have been multiplicative, and the indicated multiplicative relativities would not have depended on the other predictor variable. If instead the identity link function had been used, the model would have been additive, and the indicated additive relativities would not have depended on the other predictor variable.
3.112. The deviance residuals seem to decrease on average with $X_3$. The lack of independence of the deviance residuals and $X_3$ is not good. One should investigate refining the model.

3.113. Let $X_0$ correspond to the constant term. Let $X_1$ be 1 if there is a child. Let $X_2$ be the years of education.

\[
X = \begin{pmatrix}
1 & 0 & 12 \\
1 & 0 & 14 \\
1 & 0 & 15 \\
1 & 0 & 16 \\
1 & 0 & 17 \\
1 & 1 & 10 \\
1 & 1 & 11 \\
1 & 1 & 13 \\
1 & 1 & 15 \\
1 & 1 & 16 \\
\end{pmatrix}
\]

\[
Y = \begin{pmatrix}
1 \\
0 \\
1 \\
0 \\
1 \\
0 \\
0 \\
0 \\
1 \\
1 \\
\end{pmatrix}
\]

For a Bernoulli (yes/no) with parameter $p$, $f(y) = p^y(1-p)^{1-y}$.\[
\ln f(y) = y \ln p + (1-y) \ln(1-p) = y \ln(p/(1-p)) + \ln(1-p) = \\
y(\beta_0 + \beta_1 X_1 + \beta_2 X_2) - \ln[1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)].
\]

\[
\text{loglikelihood} = \sum y_i (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) - \sum \ln[1 + \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})].
\]

Setting the partial derivatives of the loglikelihood with respect to $\beta_0$, $\beta_1$, and $\beta_2$ equal to zero:

\[
0 = \sum y_i - \sum \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}]/(1 + \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}]).
\]

\[
0 = \sum y_i X_{1i} - \sum X_{1i} \exp[\beta_0 + \beta_1 X_{1i}]/(1 + \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}]).
\]

\[
0 = \sum y_i X_{2i} - \sum X_{2i} \exp[\beta_0 + \beta_1 X_{1i}]/(1 + \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}]).
\]

\[
\sum y_i = 1 + 0 + 1 + 0 + 1 + 0 + 0 + 1 + 0 + 1 = 5.
\]

\[
\sum y_i X_{1i} = 2.
\]

\[
\sum y_i X_{2i} = 12 + 15 + 17 + 13 + 16 = 73.
\]
The first equation becomes:
\[
5 = \frac{1}{1 + \exp[-\beta_0 - 12\beta_2]} + \frac{1}{1 + \exp[-\beta_0 - 14\beta_2]} + \frac{1}{1 + \exp[-\beta_0 - 15\beta_2]}
+ \frac{1}{1 + \exp[-\beta_0 - 16\beta_2]} + \frac{1}{1 + \exp[-\beta_0 - 17\beta_2]} + \frac{1}{1 + \exp[-\beta_0 - \beta_1 - 10\beta_2]}
+ \frac{1}{1 + \exp[-\beta_0 - \beta_1 - 11\beta_2]} + \frac{1}{1 + \exp[-\beta_0 - \beta_1 - 13\beta_2]} + \frac{1}{1 + \exp[-\beta_0 - \beta_1 - 15\beta_2]}
+ \frac{1}{1 + \exp[-\beta_0 - \beta_1 - 16\beta_2]}. 
\]

The second equation becomes:
\[
2 = \frac{1}{1 + \exp[-\beta_0 - \beta_1 - 10\beta_2]} + \frac{1}{1 + \exp[-\beta_0 - \beta_1 - 11\beta_2]} + \frac{1}{1 + \exp[-\beta_0 - \beta_1 - 13\beta_2]}
+ \frac{1}{1 + \exp[-\beta_0 - \beta_1 - 16\beta_2]}. 
\]

The third equation becomes:
\[
73 = \frac{12}{1 + \exp[-\beta_0 - 12\beta_2]} + \frac{14}{1 + \exp[-\beta_0 - 14\beta_2]} + \frac{15}{1 + \exp[-\beta_0 - 15\beta_2]}
+ \frac{16}{1 + \exp[-\beta_0 - 16\beta_2]} + \frac{17}{1 + \exp[-\beta_0 - 17\beta_2]} + \frac{10}{1 + \exp[-\beta_0 - \beta_1 - 10\beta_2]}
+ \frac{11}{1 + \exp[-\beta_0 - \beta_1 - 11\beta_2]} + \frac{13}{1 + \exp[-\beta_0 - \beta_1 - 13\beta_2]}
+ \frac{15}{1 + \exp[-\beta_0 - \beta_1 - 15\beta_2]} + \frac{16}{1 + \exp[-\beta_0 - \beta_1 - 16\beta_2]}. 
\]

Comment: In a practical application, one would have at least several hundred data points.

Using a computer, the fitted parameters are:
\[
\beta_0 = -3.65238, \beta_1 = -0.373673, \beta_2 = 0.275467.
\]

The fitted probabilities of workplace participation are:
0.4142, 0.5509, 0.6177, 0.6803, 0.7370, 0.2190, 0.2697, 0.3906, 0.5265, 0.5942.

For example, with a child and 10 years of education, the estimated probability of participating in the workforce is:
\[
\frac{\exp[-3.65238 - (1)(0.373673) + (10)(0.275467)]}{1 + \exp[-3.65238 - (1)(0.373673) + (10)(0.275467)]} = \frac{\exp[-1.271383]}{1 + \exp[-1.271383]} = 21.90\%.
\]

3.114. 1. The variables to be considered.
2. The distributional form of the errors.
3. The link function.
4. Whether he is modeling frequency, severity, or pure premium.
5. Whether he will be modeling all of the perils together or he will be modeling one of the major perils separately.

Comment: There are probably other good answers.
3.115. 1. Predictive accuracy: for the right panel graph, the plotted loss costs correspond more closely between the two lines than for the left panel graph, indicating that the proposed model seems to predict actual loss costs better than the current rating plan does.
2. Monotonicity: the current plan has a reversal in the 6th decile, whereas the model has no significant reversals.
3. Vertical distance between the first and last quantiles: the spread of actual loss costs for the current plan is about 0.6 to 1.2, which is not very much. The spread of the proposed model is larger.
Thus, by all three metrics, the new plan outperforms the current one.
Comment: Graphs taken from “Introduction to Predictive Modeling Using GLMs A Practitioner’s Viewpoint,” a presentation by Dan Tevet and Anand Khare.

Modeling fraud on claims.
Comment: Many other possible answers.

3.117. \[ \exp[8.8 + (-0.03)(30) - 0.15] = 2322. \]

3.118. \[ \text{mean} = \exp[8.8 + (-0.03)(40)] = 1998. \text{ Variance} = \phi \text{mean}^2 = (0.3)(1998^2) = 1,197,601\]

3.119. Approximately 95% of the time the actual relativity should be within the bands two standard errors on either side of the parameter estimate.
In the first graph the bands are relatively narrow. Also the relativities display an increase with vehicle symbol, which makes sense. Vehicle symbol appears to be a significant factor for the first model.
In the second graph, the bands are wide. Also the relativities display no consistent pattern with vehicle symbol. Vehicle symbol does not appear to be a significant factor for the second model. There are no parameter estimates more than two standard errors from zero.
In other words, the results are consistent with a multiplicative relativity of one for all symbols.
Comment: The graphs are taken from “A Practitioner's Guide to Generalized Linear Models,” by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi. We note that in the first graph the one-way (univariate analysis) comes up with different relativities than the GLM, presumably because vehicle symbol is correlated with other significant predictor variables in the GLM. The bottom righthand of the original of the first graph shows a p-value of 0%, indicating that vehicle symbol is significant. The original of the second graph shows a p-value of 52.5% indicating that vehicle symbol is not significant in this second model.

3.120. One way to combine separate models by peril in order to get a model for all perils:
1. Use the separate models by peril to generate predictions of expected loss due to each peril for some set of exposure data.
2. Add the peril predictions together to form a combined loss cost for each record.
3. Run a model on that data, using the combined loss cost calculated in Step 2 as the target, and the union of all the individual model predictors as the predictor variables.
3.121. (a) \( \mu = \exp[\alpha + \beta x] = \exp[\alpha_i] \exp[\beta x] \).
This is a multiplicative model, with relativities for gender and relativities for age.
The age relativities are the same for males and females.
If \( \beta < 0 \), then the relative frequencies decline exponentially with age.
(b) \( \mu = \exp[\alpha_i + \beta_i x] = \exp[\alpha_i] \exp[\beta_i x] \).
Similar to the previous model, except now the age relativities differ by gender.
For example, the relativity for age 20 relative to age 30 is:
\( \exp[20\beta_i] / \exp[30\beta_i] = \exp[-10\beta_i] \), which differs by gender.
(If \( \beta_1 = \beta_2 \), then this reduces to the previous model.)

**Comment:** Even for \( \beta_i < 0 \), this is not a realistic model of expected claim frequencies by driver age. Instead one would group the ages into for example, 17-20, 21-24, etc., and treat the age groups as categorical variables.

3.122. **D.** Histogram D most closely matches the Normal Distribution.

3.123. The results of a GLM depend on the choice of link functions. So perhaps the two models have different link functions. The results of a GLM depend on the choice of predictor variables. So perhaps the two models have different sets of predictor variables other than driver age. The results of a GLM depend on the choice of the assumed distributional form of the errors. So perhaps the two models have different distributional forms of their errors.

**Comment:** Commonly the actuary analyzes the relativities for driver age assuming all of the other predictor variables in the GLM are at the base level. If one varies the levels of the other predictor variables in the GLM, then relativities between driver ages will also usually vary.

3.124. Plot \((\Phi^{-1}[i/37]), x_{(i)}\).
\(Q_{9/37} = Q_{0.243} = -0.696\), since \(\Phi[-0.696] = 0.243\).
Thus the plotted point is: \((-0.696, 0.004)\).

3.125. A useful statistic for detecting multicollinearity is the variance inflation factor (VIF). The VIF for any predictor is a measure of how much the squared standard error for the predictor is increased due to the presence of collinearity with other predictors. It is determined by running a linear model for each of the predictors using all the other predictors as inputs, and measuring the predictive power of those models.

A common statistical rule of thumb is that a VIF greater than 10 is considered high. However, where large VIFs are indicated, it is important to look deeper into the collinearity structure in order to make an informed decision about how best to handle it in the model.
3.126. The new categorical variable has five categories, so adds 4 degrees of freedom. Test statistic is: 
\[ F = \frac{D_S - D_B}{\text{(number of added parameters)}} \phi^S \] 
\[ \phi^S = \frac{2196.1 - 2179.3}{4} / 2.09 = 2.010. \]

The number of degrees of freedom in the numerator is 4.
The number of degrees of freedom in the denominator is: number of observations minus the number of parameters in the smaller model.
We compare the test statistic to an F-distribution.
The null hypothesis is to use the simpler model.
We reject the null hypothesis if the test statistic is big.

3.127. Cross Validation is another technique for data splitting.
Split the data into for example 10 groups. Each group is called a fold. For each fold:
• Train the model using the other folds.
• Test the model using the given fold.
Several models can be compared by running the procedure for each of them on the same set of folds and comparing their relative performances for each fold.
However, cross validation is often of limited usefulness for most insurance modeling applications. Using cross validation in place of a holdout set is only appropriate where a purely automated variable selection process is used. The actuary usually applies a great deal of care and judgment in selecting the variables to be included in the model. If using cross validation, this actuarial judgement should be applied separately to each of the data sets created by leaving out one fold. This is not really practical.
For most actuarial modeling, the use of a holdout set is preferred to the use of cross validation.
Comment: See Section 4.3.4 of Goldburd, Khare, and Tevet.
Purely automated variable selection processes should be used with appropriate caution.

3.128. A common statistical rule of thumb is that a VIF greater than 10 is considered high. Thus, there is probably multicollinearity related to Weight; two or more predictors in the model are probably strongly predictive of Weight. This may cause instability problems with the model. This situation should be investigated further.
It may help to either remove Weight from the model or to preprocess the data using dimensionality reduction techniques such as principal components analysis.
Comment: The VIF of 6.33 for Body Surface Area may also warrant some investigation.
3.129. In the first graph for liability losses, the number of children seems to have a significant impact on frequency. The 95% confidence intervals do not include a log of the multiplier of 0; in other words the multiplier is significantly different from one. Also while one child increases the frequency compared to none, two children also increase the frequency compared to one. It seems as if the number of children in the household is a useful variable for modeling liability frequency for Homeowners.

In the second graph for wind losses, the number of children seems to have a insignificant impact on frequency. The 95% confidence intervals do include a log of the multiplier of 0; in other words the multiplier is not significantly different from one. Also while one child increases the frequency compared to none, two children decreases the frequency compared to one. The number of children in the household is not a useful variable for modeling wind frequency for Homeowners.

Comment: There is no logical relationship between the number of children and wind losses. A child (or any relative) who lives in the house is covered for any liability claim he or she causes. Also having children in the house may lead to more neighborhood children coming on your property with the potential for liability claims if they are injured on your property. Thus there is some logical relationship between the number of children in the household and the frequency of liability claims for Homeowners.

Presumably, the liability relativity for three children would be higher than for two children. (Three children was not shown in the graph in order to keep things simple.) One would want to apply statistical tests to see if the number of children in the household is a useful variable for modeling liability frequency. Also one would want to check the consistency over time of the indicated relativities.

3.130. With a Normal error function and an identity link function, this is the same as a multiple regression. The squared error is:

\[
800 (\beta_1 + \beta_2 + \beta_3 - 700,000/800)^2 + 600 (\beta_2 + \beta_3 - 400,000/600)^2 \\
+ 700 (\beta_1 + \beta_3 - 500,000/700)^2 + 500 (\beta_3 - 300,000/500)^2.
\]

We are given that \(\beta_3 = 570.356\), thus the squared error is:

\[
800 (\beta_1 + \beta_2 - 304.644)^2 + 600 (\beta_2 - 96.311)^2 + 700 (\beta_1 - 143.930)^2 + 500 (-29.644)^2.
\]

Setting the partial derivative with respect to \(\beta_1\) equal to zero:

\[
0 = 1600 (\beta_1 + \beta_2 - 304.644) + 1400 (\beta_1 - 143.930).
\]

\(\Rightarrow\) \(3000 \beta_1 + 1600 \beta_2 = 688,932\).

Setting the partial derivative with respect to \(\beta_2\) equal to zero:

\[
0 = 1600 (\beta_1 + \beta_2 - 304.644) + 1200 (\beta_2 - 96.311).
\]

\(\Rightarrow\) \(1600 \beta_1 + 2800 \beta_2 = 603,004\).

\(\Rightarrow\) \(\beta_2 = (603,004 - 1600 \beta_1) / 2800\).

Plugging back into the first equation: \(3000 \beta_1 + 1600 (603,004 - 1600 \beta_1) / 2800 = 688,932\).

\(\Rightarrow\) \(\beta_1 = \frac{(2800)(688,932) - (1600)(603,004)}{(3000)(2800) - (1600)(1600)} = 964,203,200 / 5,840,00 = 165.103\).

\(\Rightarrow\) \(\beta_2 = 121.014\).
3.131. BIC = (-2) (maximum loglikelihood) + (number of parameters)ln[60].
For example, BIC = (-2)(-220.18) + 2 ln[60] = 448.55.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>Loglikelihood</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>-220.18</td>
<td>448.55</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>-217.40</td>
<td>447.08</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>-214.92</td>
<td><strong>446.22</strong></td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>-213.25</td>
<td>446.97</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>-211.03</td>
<td>454.81</td>
</tr>
</tbody>
</table>

Since BIC is smallest for model C, model C is preferred.

3.132. One should also perform a statistical test to compare a model with year to a simpler model without year.
Before excluding year as a variable, it would be better to first try a model where you group the years into fewer categories, for example: 2010-2011, 2012, 2013-2014. (We may not have enough data from each year in order to be statistically confident of separate coefficients by year.)
Then if after fitting the new model the revised coefficients for years are still not significant, one could exclude year from the model.
Comment: The actuary would want to determine whether the pattern between years of the fitted coefficients makes any sense to him given his knowledge of the situation being modeled. Statistical tests are important, but just one tool. Actuarial judgement is also important.
3.133. The third plotted point is: (0.0353, 0.0079).
The last plotted point is: (0.9997, 0.9884).

Comment: In my graph, I have had the computer join the plotted points. Information was taken from the 1998 Massachusetts Wage Distribution Table. The Gini index is twice the area between the Lorenz Curve and the Line of Equality.
Simple quantile plots are created via the following steps:
1. Sort the dataset based on the Model A predicted loss cost (from smallest to largest).
2. Bucket the data into quantiles, such that each quantile has the same volume of exposures. Common choices are quintiles (5 buckets), deciles (10 buckets), or vigintiles (20 buckets).
3. Within each bucket, calculate the average predicted pure premium (predicted loss per unit of exposure) based on the Model A predicted loss cost, and calculate the average actual pure premium.
4. Plot, for each quantile, the actual pure premium and the pure premium predicted by Model A.
5. Repeat steps 1 through 4 using the Model B predicted loss costs.
There are now two quantile plots: one for Model A and one for Model B.
6. Compare the two quantile plots to determine which model provides better lift.
In order to determine the “winning” model, consider the following 3 criteria:
1. **Predictive accuracy.** How well each model is able to predict the actual pure premium in each quantile.
2. **Monotonicity.** By definition, the predicted pure premium will monotonically increase as the quantile increases, but the actual pure premium should also increase (though small reversals are okay).
3. **Vertical distance between the first and last quantiles.** The first quantile contains the risks that the model believes will have the best experience, and the last quantile contains the risks that the model believes will have the worst experience. A large difference (also called “lift”) between the actual pure premium in the quantiles with the smallest and largest predicted loss costs indicates that the model is able to maximally distinguish the best and worst risks.

Comment: See Section 7.2.1 of Goldburd, Khare, and Tevet.

(a) For example, using a discrimination threshold of 25%, one would be predicting fraud for any claim for which the GLM says the probability of fraud is greater than 25%. Alternately, choose a specific probability level, called the discrimination threshold, above which we will investigate the claim for fraud and below which we will not. This determination may be thought of as the model’s “prediction” in a binary (i.e., fraud/no fraud) sense.
(b) Using a lower threshold would detect more of the fraudulent claims, at the cost of also having to investigate more claims which turned out not to be fraudulent. Using a higher threshold would detect fewer of the fraudulent claims, but we would have to investigate fewer claims which turned out not to be fraudulent.
Alternately, there is trade-off: a lower threshold results in a higher sensitivity (true positive rate), while a higher threshold results in a higher specificity (and thus a lower false positive rate).

Comment: Similar to 8, 11/17, Q.6d.

See pages 75 to 77 of Generalized Linear Models for Insurance Rating.
“The selection of a discrimination threshold involves a trade-off: a lower threshold will result in more true positives and fewer false negatives than a higher threshold, but at the cost of more false positives and fewer true negatives.”
The mean modeled claim counts are:

<table>
<thead>
<tr>
<th></th>
<th>Terr. A</th>
<th>Terr. B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>$24,000 \exp[\beta_0]$</td>
<td>$15,000 \exp[\beta_0 + \beta_2]$</td>
</tr>
<tr>
<td>Female</td>
<td>$20,000 \exp[\beta_0 + \beta_1]$</td>
<td>$13,000 \exp[\beta_0 + \beta_1 + \beta_2]$</td>
</tr>
</tbody>
</table>

The likelihood function of a Poisson is: $\ln f(y_i; \mu_i) = \sum (\mu_i - y_i \ln[\mu_i] - \ln[y_i!])$

The loglikelihood ignoring terms that do not depend on the betas is:

$$-24,000 \exp[\beta_0] + 1200 (\beta_0) - 20,000 \exp[\beta_0 + \beta_1] + 800 (\beta_0 + \beta_1)$$  
$$- 15,000 \exp[\beta_0 + \beta_2] + 1100 (\beta_0 + \beta_2) - 13,000 \exp[\beta_0 + \beta_1 + \beta_2] + 900 (\beta_0 + \beta_1 + \beta_2).$$

Setting the partial derivative of the loglikelihood with respect to $\beta_1$ equal to zero:

$$- 20,000 \exp[\beta_0 + \beta_1] + 800 - 13,000 \exp[\beta_0 + \beta_1 + \beta_2] + 900 = 0.$$  
Given $\beta_0 = -3.0300$:  
$$1700 = 966.3 \exp[\beta_1] + 628.1 \exp[\beta_1] \exp[\beta_2].$$

Setting the partial derivative of the loglikelihood with respect to $\beta_2$ equal to zero:

$$- 15,000 \exp[\beta_0 + \beta_2] + 1100 - 13,000 \exp[\beta_0 + \beta_1 + \beta_2] + 900 = 0.$$  
$$\Rightarrow 2000 = 724.7 \exp[\beta_2] + 628.1 \exp[\beta_1] \exp[\beta_2].$$

Subtracting two equations: $300 = 724.7 \exp[\beta_2] - 966.3 \exp[\beta_1].$

$$\Rightarrow \exp[\beta_2] = 0.4140 + 1.3334 \exp[\beta_1].$$

$$\Rightarrow 1700 = 966.3 \exp[\beta_1] + 628.1 \exp[\beta_1] (0.4140 + 1.3334 \exp[\beta_1]).$$

Let $x = \exp[\beta_1]$.  
$$\Rightarrow 1700 = 966.3 x + 628.1 x (0.4140 + 1.3334 x).$$

$$\Rightarrow 837.5 x^2 + 1226.3 x - 1700 = 0.$$  
$$\Rightarrow x = \frac{-1226.3 \pm \sqrt{1226.3^2 - (4)(837.5)(-1700)}}{2(837.5)} = 0.8697,$$ taking the positive root.

$$\Rightarrow \beta_1 = \ln(0.8697) = -0.1396.$$  
$$\Rightarrow \exp[\beta_2] = 0.4140 + 1.3334 \exp[\beta_1] = 0.4140 + (1.3334)(0.8697) = 1.5737.$$  
$$\Rightarrow \beta_2 = \ln(1.5737) = 0.4534.$$  

Expected frequency of a female risk in Territory B is:  
$$\exp[\beta_0 + \beta_1 + \beta_2] = \exp[-3.0300 - 0.1396 + 0.4534] = 6.61\%.$$  

Comment: Similar to 8, 11/15, Q.3.

Using a computer, without being given $\beta_0$, the maximum likelihood fit is:  
$$\hat{\beta}_0 = -3.02999, \hat{\beta}_1 = -0.139599, \hat{\beta}_2 = 0.453335.$$  

The mean modeled frequencies are:

<table>
<thead>
<tr>
<th></th>
<th>Territory A</th>
<th>Territory B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>$\exp[-3.02999] = 4.83%$</td>
<td>$\exp[-3.02999 + 0.453335] = 7.60%$</td>
</tr>
<tr>
<td>Female</td>
<td>$\exp[-3.02999 - 0.139599] = 4.20%$</td>
<td>$\exp[-3.02999 - 0.139599 + 0.453335] = 6.61%$</td>
</tr>
</tbody>
</table>
3.137. When the effect of one predictor depends on the level of another predictor, and vice-versa, such an relationship is called an interaction. An example of an interaction term: $X_1X_2$.

In this example, $g(\mu) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + ...$

The effect of $X_1$ depends on the level of $X_2$ and vice-versa.

Comment: See Section 5.6 of Goldburd, Khare, and Tevet.

The actuary can use the GLM significance statistics in order to determine whether the inclusion of an interaction significantly improves the model.
3.138. (a) The percent of losses for A are 50%. So the Lorenz Curve has the point (50, 50).

The Lorenz curve is equal to the line of equality, and thus the area between them is zero. The Gini Index is twice that, or zero.

(b) The percent of losses for A are 0%. So the Lorenz Curve has the point (50, 0).

The region between the Lorenz curve and the line of equality is a triangle of base 50% and height 100%, and thus area: \((1/2)(50\%)(100\%) = 0.25\). The Gini Index is twice that, or 50%.

Comment: We looked at the two extreme cases, which will not occur in practice.

Here is a graph of the Gini Index versus the percent of total actual losses in Class A:
3.139. The curve corresponding to the text labeled VA has more area under it, so it is better than the test labeled NE.

3.140. The fitted parameter(s) are the same, while the standard errors are multiplied by $\sqrt{3.071}$.

The standard error of $\hat{\beta}_1$ is: $0.1978 \sqrt{3.071} = 0.3466$.

95% confidence interval for $\beta_1$: $5.624 \pm (1.96)(0.3466) = 5.624 \pm 0.679$.

Comment: One could instead use: $5.624 \pm (2)(0.3466) = 5.624 \pm 0.693$.

3.141. A simple quintile plot is a simple quantile plot with 5 buckets.
- Sort the dataset based on the model predicted pure premium from smallest to largest.
- Group the data into 5 buckets with equal volume.
- Within each group, calculate the average predicted pure premium based on the model, and the average actual pure premium.
- Plot for each group, the actual pure premium and the predicted pure premium.

Since we are not given the overall average pure premium, I will plot the pure premiums relative to average.

The saturated model has as many predictors as data points. Thus for the saturated model, the predictions exactly match the observations for each record.

The simple quintile plot:
The null model, has no predictors, only an intercept. Thus for the null model the prediction is the same for every record: the grand mean. Since every risk has the same prediction, one would assign them to buckets at random. Thus all of the actuals by quintile should be close to the grand mean, with small differences due to the randomness of assignments. The simple quintile plot:
“A model that could be used in practice”, would have the actuals increase monotonically, have good but not perfect predictive accuracy, and a reasonably large vertical distance between the actuals in the first and last quintiles. A simple quintile plot:

Comment: Similar to 8, 11/07, Q. 5.
There are many possible examples of the last plot.
Since the records are ordered by predicted values, the records in each bucket change for each graph. Thus, actuals are not the same between the graphs.
Quintile plots are sorted by predicted values from smallest to largest value. Thus the predicted values must be monotonically increasing (or in the case of the null model equal). Actuals need not be monotonically increasing, although that is desirable.
In every graph, the average of the actuals should be the grand mean.
In the final plot, the average of the predicteds should be close to if not equal to the grand mean; the GLM may have a small bias.
In the final plot, the predicted and actuals for the final quintile should each be less than in the saturated model. In the final plot, the predicted and actuals for the final quintile should each be more than in the null model.

3.142. I prefer the Gamma model, since the standardized deviance residuals are much closer to being Normally Distributed.
3.143. Since the proposed model is not able to segment the data into lower and higher loss ratio buckets, the proposed model is not significantly outperforming the current rating plan. 
Comment: See Section 7.2.3 of Goldburd, Khare, and Tevet.

3.144. AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
For example, AIC = (-2)(-9844.16) + (5)(2) = 19,698.32.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>Loglikelihood</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>-9844.16</td>
<td>19,698.32</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>-9822.48</td>
<td>19,664.96</td>
</tr>
<tr>
<td>C</td>
<td>15</td>
<td>-9815.70</td>
<td>19,661.40</td>
</tr>
</tbody>
</table>

Since AIC is smallest for model C, model C is preferred.

3.145. BIC = (-2) (maximum loglikelihood) + (number of parameters) ln[number of data points].
For example, BIC = (-2)(-9844.16) + 5 ln[5000] = 19730.91.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>Loglikelihood</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>-9844.16</td>
<td>19,730.91</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>-9822.48</td>
<td>19,730.13</td>
</tr>
<tr>
<td>C</td>
<td>15</td>
<td>-9815.70</td>
<td>19,759.16</td>
</tr>
</tbody>
</table>

Since BIC is smallest for model B, model B is preferred.
Comment: Similar to 8, 11/16, Q.7.
See Section 6.2.2 in Goldburd, Khare, and Tevet.
Most actuarial GLMs are fit to many more than 5000 data points.
“As most insurance models are fit on very large datasets, the penalty for additional parameters imposed by BIC tends to be much larger than the penalty for additional parameters imposed by AIC. In practical terms, the authors have found that AIC tends to produce more reasonable results. Relying too heavily on BIC may result in the exclusion of predictive variables from your model.”

3.146. The first model does a better job of fitting the data and is thus preferred.
3.147. Sort the risks from best to worst based on the model predicted pure premium.

<table>
<thead>
<tr>
<th>Risk</th>
<th>Model P.P.</th>
<th>Exposures</th>
<th>Cumulative Exposures</th>
<th>% of Exposures</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1000</td>
<td>7</td>
<td>7</td>
<td>7%</td>
</tr>
<tr>
<td>8</td>
<td>2000</td>
<td>24</td>
<td>31</td>
<td>31%</td>
</tr>
<tr>
<td>5</td>
<td>3000</td>
<td>12</td>
<td>43</td>
<td>43%</td>
</tr>
<tr>
<td>3</td>
<td>4000</td>
<td>8</td>
<td>51</td>
<td>51%</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>11</td>
<td>62</td>
<td>62%</td>
</tr>
<tr>
<td>6</td>
<td>6000</td>
<td>16</td>
<td>78</td>
<td>78%</td>
</tr>
<tr>
<td>1</td>
<td>7000</td>
<td>3</td>
<td>81</td>
<td>81%</td>
</tr>
<tr>
<td>7</td>
<td>8000</td>
<td>19</td>
<td>100</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk</th>
<th>Exposures</th>
<th>Actual P.P.</th>
<th>Actual Losses</th>
<th>Cumulative Losses</th>
<th>% of Losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>7</td>
<td>4000</td>
<td>28,000</td>
<td>28,000</td>
<td>5.6%</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>4000</td>
<td>96,000</td>
<td>124,000</td>
<td>24.8%</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>1000</td>
<td>12,000</td>
<td>136,000</td>
<td>27.2%</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>2000</td>
<td>16,000</td>
<td>152,000</td>
<td>30.4%</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>8000</td>
<td>88,000</td>
<td>240,000</td>
<td>48.0%</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>8000</td>
<td>128,000</td>
<td>368,000</td>
<td>73.6%</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>6000</td>
<td>18,000</td>
<td>386,000</td>
<td>77.2%</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>6000</td>
<td>114,000</td>
<td>500,000</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

On the x-axis, plot the cumulative percentage of exposures.  
On the y-axis, plot the cumulative percentage of actual losses.  
The plotted points are: (0, 0), (7%, 5.6%), (31%, 24.8%), ... , (81%, 77.2%), (100%, 100%).
Comment: Similar to 8, 11/16, 5a.
The Gini index is twice the area between the Lorenz Curve and the line of equality. The higher the Gini Index, the better the rating plan is at identifying risk differences.

3.148.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle Price</td>
<td>3</td>
</tr>
<tr>
<td>Vehicle Age</td>
<td>$8 - 1 = 7$</td>
</tr>
<tr>
<td>Driver age</td>
<td>$2 - 1 = 1$</td>
</tr>
<tr>
<td>Number of drivers</td>
<td>$3 - 1 = 2$</td>
</tr>
<tr>
<td>Gender</td>
<td>$2 - 1 = 1$</td>
</tr>
<tr>
<td>Interaction Gender &amp; Driver Age</td>
<td>1</td>
</tr>
</tbody>
</table>

Number of parameter is: $3 + 7 + 1 + 2 + 1 + 1 = 15$.
Comment: Similar to CAS S, 11/15, Q.35.

A model with only Vehicle Price would involve: $\beta_0 + \beta_1 (vp) + \beta_2 (vp)^2$.
The interaction of gender and driver age only uses one parameter since each of gender and driver age only use one parameter.
3.149. A double lift chart compares the current rating plan to a proposed model. Sort data by ratio of model prediction to current premium. Subdivide sorted data into quantiles with equal exposure. For each quantile calculate average actual loss cost, average model predicted loss cost and the average loss cost underlying the current manual premium. Index the quantile averages to the overall averages. Plot the results.

Comment: The “winning” model is the one that more closely matches the actual pure premiums.

3.150. The difference in degrees of freedom is: 18,175 - 18,169 = 6; we add 6 parameters.

Test statistic is: 
\[ F = \frac{D_S - D_B}{(\text{number of added parameters})} = \frac{8,905.6226 - 8,901.4414}{(6) (0.4523)} = 1.541. \]

The number of degrees of freedom in the numerator is 6.
The number of degrees of freedom in the denominator is: number of degrees of freedom for the simpler model = 18,175.
We compare the test statistic to the appropriate F-distribution.
The null hypothesis is to use the simpler model.
We reject the null hypothesis if the test statistic is sufficiently big.

Comment: Using a computer, the p-value is 16.0%. Thus at for example a 10% significance level, we do not reject the null hypothesis to use the simpler model.

3.151. Mean = \( \exp[5.07 + 0.48 - 0.36] = 179.5 \).
Variance = mean^2 / \alpha = 179.5^2 / 2.2 = 14,646.

Comment: Similar to CAS S, 5/16, Q.32.

3.152. The Gini index can be used to measure the lift of an insurance rating plan by quantifying its ability to segment the population into the best and worst risks.
The larger the Gini index, the better job the rating plan does of segmenting.
Thus the rating plan used in Model 1 has more lift than the rating plan used in Model 2.

3.153. One works with loss ratios with respect to the premiums for the current plan. To create a loss ratio chart:
1. Sort the dataset based on the model prediction.
2. Group the data into quantiles with equal volumes of exposures.
3. Within each group, calculate the actual loss ratio.

Comment: If the proposed model is able to segment the data into lower and higher loss ratio buckets, then the proposed model is better than the current model.
3.154. Driver age and number of years claims-free are positively correlated. Older drivers are likely to be claims-free for more years than younger drivers. Thus in order to avoid double counting effects, the GLM lessens the effect of each variable somewhat compared to a model that just used one of the two variables.

**Comment:** A graph of number of years claims-free versus driver age:

![Exposure Distribution (Age X NCD)](image)


If two variables are very highly correlated, which is not the case here, then the GLM will have trouble converging and the parameter estimates may be unreliable.

3.155. If the current rating plan were perfect, then all risks should have the same loss ratio. The fact that the proposed model is able to segment the data into lower and higher loss ratio buckets is a strong indicator that it is outperforming the current rating plan.

**Comment:** Graph taken from “Goodness of Fit vs. Goodness of Lift,” by Glenn Meyers and David Cummings, August 2009 Actuarial Review.

If one insurer were to use the current rating plan, while another insurer were to use the proposed rating plan, the second insurer should be able to attract better risks from the first insurer. The first insurer who continued to use the current plan would be subject to adverse selection.
3.156. The offset is: \( \ln(0.8) = -0.223 \).
The linear component is: \( 6 + (0.1)(13) + (-0.2)(3) = 6.7 \).
Modeled pure premium = \( \exp[6.7 - 0.223] = 650 \).
Alternately, the modeled pure premium is: \( (0.8) \exp[6 + (0.1)(13) + (-0.2)(3)] = 650 \).

3.157. Laurel is proposing the saturated model; it is overfit.
Hardy is proposing the null model; it is underfit.
A model with a number of parameters between the two would make sense.

3.158. The average relativity is:
\[
(1.50)(9\%) + (1.35)(4\%) + (1.18)(5\%) + (1)(82\%) = 1.0680.
\]
Thus compared to average, 3 or more years claims-free has a relativity of:
\[
1.00/1.0680 = 0.9363.
\]
Thus three years of data has a credibility of: \( 1 - 0.9363 = 6.4\% \).
Comment: Bailey and Simon look at the credibility of cars rather than drivers.
Also, Bailey and Simon were dealing with a very simple classification system.
The credibility depends on how refined the class system is; the more refined the classification
system, the less credibility is given to individual experience.
In order to estimate a credibility for one year of data, one would have to group drivers into those
that had at least one year claims free.

3.159. These two variables are likely significantly positively correlated.
(Those policies with three or more operators, usually list parents and one or more children as
drivers. The listed child will be a teenager or young adult living at home. Very few teenagers
have their own car on their own separate policy.)
Including two such variables in a model, can produce anomalous results.
In any case, due to the interaction of these two variables, it will be difficult to interpret the
relativities for the different levels of each variable.

3.160. Both AIC and BIC are penalized measures of fit; in each case a penalty is added to twice
the negative loglikelihood. In the case of AIC the penalty is twice the number of parameters in
the model, while in the case of BIC the penalty is \( p \log(n) \), where \( p \) is the number of parameters,
and \( n \) is the number of data points that the model is fit on
In both cases, a smaller statistic is better.
“In practical terms, the authors have found that AIC tends to produce more reasonable results.
Relying too heavily on BIC may result in the exclusion of predictive variables from your model.”
In other words, they believe that the use of BIC will tend to produce models that are underfit.

3.161. The “winning” model is the one that more closely matches the actual pure premiums.
The proposed model does a much better job than the current rating plan; thus the proposed
model is preferred.
3.162. 1. Model A does a better job of matching the actual than does Model B. Thus based on the criterion of predictive accuracy I prefer Model A. Both models satisfy the criterion of monotonicity: the actuals increase with quintile. Model A has a larger vertical distance between the actuals for the first and last quintiles than does Model B. Thus based on this criterion I prefer Model A. Thus overall I prefer Model A to Model B.

Comment: In order to determine the “winning” model, consider the following 3 criteria:
1. Predictive accuracy. How well each model is able to predict the actual pure premium in each quantile.
2. Monotonicity. By definition, the predicted pure premium will monotonically increase as the quantile increases, but the actual pure premium should also increase (though small reversals are okay).
3. Vertical distance between the first and last quantiles. The first quantile contains the risks that the model believes will have the best experience, and the last quantile contains the risks that the model believes will have the worst experience. A large difference (also called “lift”) between the actual pure premium in the quantiles with the smallest and largest predicted loss costs indicates that the model is able to maximally distinguish the best and worst risks.

3.163. I prefer the Inverse Gaussian model, since the standardized deviance residuals are much closer to being Normally Distributed.

3.164. \( X\beta = 1 + (8)(0.31) = 3.48. \)
The odds are: \( e^{3.48} = 32.5. \)

Alternately, for the logistic model: \( \hat{\pi} = e^{3.48} / (1 + e^{3.48}) = 0.9701. \)
The odds are: \( \hat{\pi} / (1 - \hat{\pi}) = 0.9701 / (1 - 0.9701) = 32.4. \)

Comment: Similar to MAS-1, 5/18, Q.27.
We have estimated that the probability of renewal is 32.5 times the probability of a nonrenewal.
### 3.165.

<table>
<thead>
<tr>
<th>Claim #</th>
<th>Fraud</th>
<th>Predict.</th>
<th>30% Threshold</th>
<th>60% Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N</td>
<td>Y</td>
<td>False Pos.</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>N</td>
<td>True Neg.</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>N</td>
<td>True Neg.</td>
<td>N</td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td>Y</td>
<td>True Pos.</td>
<td>Y</td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>N</td>
<td>True Neg.</td>
<td>N</td>
</tr>
<tr>
<td>7</td>
<td>Y</td>
<td>N</td>
<td>False Neg.</td>
<td>N</td>
</tr>
<tr>
<td>8</td>
<td>N</td>
<td>Y</td>
<td>False Pos.</td>
<td>Y</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>Y</td>
<td>True Pos.</td>
<td>Y</td>
</tr>
<tr>
<td>10</td>
<td>Y</td>
<td>Y</td>
<td>True Pos.</td>
<td>Y</td>
</tr>
<tr>
<td>12</td>
<td>Y</td>
<td>Y</td>
<td>True Pos.</td>
<td>N</td>
</tr>
<tr>
<td>14</td>
<td>N</td>
<td>Y</td>
<td>False Pos.</td>
<td>Y</td>
</tr>
<tr>
<td>15</td>
<td>N</td>
<td>Y</td>
<td>False Pos.</td>
<td>N</td>
</tr>
</tbody>
</table>

(a)

#### 30% Threshold

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Fraud</th>
<th>No Fraud</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud</td>
<td>true pos.: 4</td>
<td>false neg.: 1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>No Fraud</td>
<td>false pos.: 7</td>
<td>true neg.: 3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>4</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

#### 60% Threshold

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Fraud</th>
<th>No Fraud</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud</td>
<td>true pos.: 3</td>
<td>false neg.: 2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>No Fraud</td>
<td>false pos.: 2</td>
<td>true neg.: 8</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>
(b) Sensitivity = \( \frac{\text{True Positives}}{\text{Total Number of Events}} = \frac{\text{Correct Predictions of Fraud}}{\text{Total Number of Fraudulent Claims}} \).

Specificity = \( \frac{\text{True Negatives}}{\text{Total Number of Non-Events}} = \frac{\text{Correct Predictions of No Fraud}}{\text{Total Number of Nonfraudulent Claims}} \).

30% threshold: sensitivity = 4/5, and specificity = 3/10. Graph (1 - 3/10, 4/5).
60% threshold: sensitivity = 3/5, and specificity = 8/10 = 4/5. Graph (1 - 4/5, 3/5).

The ROC Curve, plus the 45-degree comparison line:

Comment: Similar to 8, 11/07, Q. 6a&b.
3.166. Let $t_i$ be the territory relativity for a given insured. Then the offset is $\ln(t_i)$.

As per Section 2.6 of “Generalized Linear Models for Insurance Rating”:

$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \text{offset}$.

Olaf would have a set of predictors of class relativity, and the model would be:

$\ln[\text{class relativity}] = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \ln(t_i)$.

class relativity $= \exp[\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \ln(t_i)] = \exp[\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}] t_i$.

The GLM would be fit as usual, where each $t_i$ is a known constant.

Comment: \[\frac{\text{class relativity}}{\text{known territory relativity}} = \exp[\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}]\].

Territories are not a good fit for the GLM framework.

See Section 8.2 of “Generalized Linear Models for Insurance Rating”.

One should include the territory relativities produced by the separate model as an offset in the GLM used to determine classification relativities. Similarly, one should include classification relativities produced by the GLM as an offset in the separate model used to determine territory relativities.

3.167. A model with approximately 6 degrees of freedom has the right balance, since it has the smallest test MSE.

A model with approximately 2 degrees of freedom is too simple, it has a larger test MSE.

A model with approximately 20 degrees of freedom is too complex, it has a larger test MSE; while the training MSE is smaller that is due to this model being overfit.

Comment: Similar to 8, 11/17, Q 4b.

See Figure 7 in Generalized Linear Models for Insurance Rating.

We are interested in how the GLM will perform at predicting the response variable on some future set of data rather than on the set of past data with which we are currently working.

Our goal in modeling is to find the right balance where we pick up as much of the signal as possible with minimal noise, represented in this case by model with about 6 degrees of freedom.
3.168. (a) i. One can bin driver age into groups.
   ii. Adding polynomial terms, such as $X^2$.
   iii. One can use hinge functions. A hinge function is of the form: $(X - c)_+ = \max(0, X - c)$. This will result in a piecewise linear function, with a change in slope at each breakpoint $c_i$.

(b) i. With binning: Continuity is not guaranteed.
   Variation within intervals is ignored.
   There may not be enough data in each bin to be credible.
   There could be non-intuitive results, such as reversals.
   ii. It is hard to interpret models that include powers.
   “One potential downside to using polynomials is the loss of interpretability. From the coefficients alone it is often very difficult to discern the shape of the curve; to understand the model’s indicated relationship of the predictor to the target variable it may be necessary to graph the polynomial function.”
   iii. Using hinge functions:
   The breakpoints must be selected by the user.
   Comment: Similar to 8, 11/18, Q.5.

In all cases, more parameters are added to the model; the principal of parsimony states that we prefer a simpler model with fewer parameters, all else being equal.

3.169. \[ \frac{\exp[2.182 + 1.137 - 0.422]}{1 + \exp[2.182 + 1.137 - 0.422]} = 94.77\% . \]

Comment: Similar to CAS S, 5/16, Q.33.

3.170. (a) Adding driver age using 5 bins, adds 4 parameters.
   i. Deviance = $2 \{ \log\text{likelihood for the saturated model} - \log\text{likelihood for the fitted model} \}$. $D_A = (2) \{-100 - (-130)\} = 60$. $D_B = (2) \{-100 - (-123)\} = 46$.
   $F = \frac{(D_A - D_B)}{\text{(number of added parameters)}} = \frac{(60 - 46)}{4} / 0.61 = 5.738$.

We compare to the given critical value of 5.707. Since $5.738 > 5.707$, model B is significantly better than model A. Driver age should be included in the rating plan.

ii. $AIC_A = (-2)(-130) + (2)(5) = 270$.
    $AIC_B = (-2)(-123) + (2)(5+4) = 264$.

Since $AIC_B < AIC_A$, model B is better than model A. Driver age should be included in the rating plan.

iii. $BIC_A = (-2)(-130) + (5) \ln(50) = 279.56$.
    $BIC_B = (-2)(-123) + (5+4) \ln(50) = 281.21$.

$BIC_A < BIC_B$, and model A is better than model B. Driver age should not be included in the rating plan.

Comment: Similar to 8, 11/18, Q. 6a.

The degrees of freedom for Model A = number of observations minus number of fitted parameters for model A = 50 - 5 = 45.

The F-statistic has degrees of freedom 45 and 2. The given critical value is the 10% critical value for 45 and 2 degrees of freedom.
3.171. \( X\beta = 1 + (2)(0.50) + (-5)(0.29) = 0.55 \).

The Gamma has the inverse as its canonical link function.

\( 1/ (X\beta) = 1/0.55 = 1.818 \).

The Poisson has the log as its canonical link function.

\( \exp[X\beta] = e^{0.55} = 1.733 \).

The Normal has the identity as its canonical link function.

\( X\beta = 0.55 \).

\( 0.55 < 1.733 < 1.818 \). Thus the correct ordering is: \( \text{III} < \text{II} < \text{I} \).

Comment: Similar to MAS-1, 5/18, Q.25.

The Binomial has the logit as its canonical link function.

3.172. (a) As per Section 2.6 of “Generalized Linear Models for Insurance Rating”:

\[ g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \text{offset}. \]

Thus the offset for Policy 1 is: \( \ln\left( \frac{0.023}{1 - 0.023} \right) = -3.749 \).

The offset for Policy 2 is: \( \ln\left( \frac{0.112}{1 - 0.112} \right) = -2.070 \).

The offset for Policy 3 is: \( \ln\left( \frac{0.045}{1 - 0.045} \right) = -3.055 \).

(b) In each case we add the offset to the linear component from the insurance score.

For Policy 1: \( 1.581 + (-0.032)(34) - 3.749 = -3.256 \).

Probability of a claim is: \( \frac{\exp[-3.256]}{1 + \exp[-3.256]} = 3.71\% \).

For Policy 2: \( 1.581 + (-0.032)(66) - 2.070 = -2.601 \).

Probability of a claim is: \( \frac{\exp[-2.601]}{1 + \exp[-2.601]} = 6.91\% \).

For Policy 3: \( 1.581 + (-0.032)(88) - 3.055 = -4.290 \).

Probability of a claim is: \( \frac{\exp[-4.290]}{1 + \exp[-4.290]} = 1.35\% \).

Comment: Similar to 8, 11/18, Q. 7a and 7b.

3.173. One-way or univariate analysis does not accurately take into account the effect of other rating variables. It does not consider exposure correlations with other rating variables.
3.174. a. Linear Model:
- Random Component: Each component of Y is independent and normally distributed. Their means may differ, but they have common variance.
- Systematic Component: The covariates are combined to produce the linear predictor $\eta = X\beta$.
- Link Function: The relationship between the random component and the systematic component is specified with the identity link function: $E(Y) = \mu = \eta$.
  (if $g$ is the identity link function, $g^{-1}(\eta) = \eta$.)

Generalized Linear Model:
- Random Component: Each component of Y is independent and a member of an exponential family. (While the Normal is one possibility, there are others.)
- Systematic Component: The covariates are combined to produce the linear predictor $\eta = X\beta$.
- Link Function: The relationship between the random component and the systematic component is specified with the link function, which is differentiable and monotonic such that: $E(Y) = \mu = g^{-1}(\eta)$.
  (While the identity link function is one possibility, there are others.)

b) 1) The assumption of normality with common variance is often not true.
2) Sometimes the response variable may be restricted to be positive, but normality with the identity link function violates this.

3.175. i. Classical Linear Model: Response variable is normally distributed.
Generalized Linear Model: Response variable is from the exponential family.
ii) Classical Linear Model: The variance is constant but the mean is allowed to vary.
Generalized Linear Model: The variance is a function of the mean (exponential family).
3.176. a. Intrinsic aliasing is a linear dependency between covariates due to the definition. For example, if we have only black, red and blue cars, the red cars can be determined from \([\text{total cars}] - \text{black} - \text{blue}\). As another example, age of vehicle would alias with model year, since if you know one you can determine the other.
Extrinsic aliasing is a linear dependency between covariates that arises due to the nature of the data rather than inherent properties of the covariates themselves. For example, if in the data all cars with unknown color also have an unknown number of doors, and vice-versa.

b. We have that \([\text{all cars}] - \text{large cars} - \text{medium cars} = \text{small cars}\), so we can say that
\[
X_{\text{small}} = 1 - X_{\text{large}} - X_{\text{medium}}.
\]
If we do not have a base level, then we could have two size variables such as Large and Medium, plus all four territories.
We have that \([\text{all cars}] - \text{North} - \text{South} - \text{West} = \text{East}\), so we can say that
\[
X_{\text{East}} = 1 - X_{\text{North}} - X_{\text{South}} - X_{\text{West}}.
\]
If we do not have a base level, then we could have three territory variables such as North, South, and West, plus all three sizes.
Alternately, we can eliminate \(\beta_{\text{small}}\) and \(\beta_{\text{East}}\) from the model and include an intercept term; Small / East would be the base level. Intercept plus 2 size and 3 territory variables.

Comment: The current syllabus reading does not distinguish between intrinsic and extrinsic aliasing.

In part (b) we should end up with 6 variables in total.
If we have an intercept term, we would have in addition three territory levels and two size levels.
Aliasing occurs when there is a linear dependency among the observed covariates. Equivalently, aliasing can be defined as a linear dependency among the columns of the design matrix \(X\).
Near aliasing is a common problem and occurs when two or more factors contain levels that are almost, but not quite, perfectly correlated. This same problem comes up when performing multiple linear regressions.
3.177. There are many possible ways to set this up. Taking North and Medium as the base levels as instructed. Let \( X_1 \) correspond to the intercept term. It is one for all cells. Let \( X_2 \) correspond to South. \( X_2 = 1 \) if South and 0 otherwise. Let \( X_3 \) correspond to East. \( X_3 = 1 \) if East and 0 otherwise. Let \( X_4 \) correspond to West. \( X_4 = 1 \) if West and 0 otherwise. Let \( X_5 \) correspond to Small. \( X_5 = 1 \) if Small and 0 otherwise. Let \( X_6 \) correspond to Large. \( X_6 = 1 \) if Large and 0 otherwise. Then the design matrix, \( X \), and response vector \( Y \) are:

\[
X = \begin{pmatrix}
1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 \\
\end{pmatrix}
\]

\[
Y = \begin{pmatrix}
100 \\
150 \\
250 \\
80 \\
110 \\
290 \\
90 \\
170 \\
200 \\
180 \\
260 \\
540 \\
\end{pmatrix}
\]

The vector of parameters is:

\[
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5 \\
\beta_6 \\
\end{pmatrix}
\]

Comment: While one could fit a Poisson to pure premiums, and treat the result as a discrete approximation, it is more common to fit a Tweedie Distribution. Using a computer the fitted parameters are:

\[
\beta_1 = 4.95978, \beta_2 = -0.040822, \beta_3 = -0.0833816, \beta_4 = 0.672944, \beta_5 = -0.427444, \beta_6 = 0.617924.
\]

We have a multiplicative model with relativities:

South: \( \exp[-0.040822] = 0.960 \), East: \( \exp[-0.0833816] = 0.920 \), West: \( \exp[0.672944] = 1.960 \), Small: \( \exp[-0.427444] = 0.652 \), Large: \( \exp[0.617924] = 1.855 \).

For example, the fitted value for South and Small is:

\[
\exp[\beta_1 + \beta_2 + \beta_3] = \exp[4.95978 - 0.040822 - 0.427444] = 89.26.
\]
The fitted values are:

<table>
<thead>
<tr>
<th>Territory</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>92.98</td>
<td>142.56</td>
<td>264.46</td>
<td>500</td>
</tr>
<tr>
<td>South</td>
<td>89.26</td>
<td>136.86</td>
<td>253.88</td>
<td>480</td>
</tr>
<tr>
<td>East</td>
<td>85.54</td>
<td>131.16</td>
<td>243.31</td>
<td>460.01</td>
</tr>
<tr>
<td>West</td>
<td>182.23</td>
<td>279.42</td>
<td>518.35</td>
<td>980</td>
</tr>
<tr>
<td>Total</td>
<td>450.01</td>
<td>690</td>
<td>1280</td>
<td>2420.01</td>
</tr>
</tbody>
</table>

Subject to rounding, the totals for the fitted match those for the data.

In general, the estimates will be in balance as they were here, when one uses the canonical link function; the canonical link function for the Poisson is the log link function.

3.178. a) \( Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e \).

\[ Y_1 = 400 = \beta_1 + 0 + \beta_3 + e_1. \]

\[ Y_2 = 250 = \beta_1 + 0 + 0 + e_2. \]

\[ Y_3 = 200 = 0 + \beta_2 + \beta_3 + e_3. \]

\[ Y_4 = 100 = 0 + \beta_2 + 0 + e_4. \]

Sum of Squared Errors = \( e_1^2 + e_2^2 + e_3^2 + e_4^2 \)

= \((400 - \beta_1 - \beta_3)^2 + (250 - \beta_1)^2 + (200 - \beta_2 - \beta_3)^2 + (100 - \beta_2)^2\).

Set equal to zero the partial derivatives with respect to betas:

\[ 2(400 - \beta_1 - \beta_3)(-1) + 2(250 - \beta_1)(-1) = 0 \Rightarrow 2\beta_1 + \beta_3 = 650. \]

\[ 2(200 - \beta_2 - \beta_3)(-1) + 2(100 - \beta_2)(-1) = 0 \Rightarrow 2\beta_2 + \beta_3 = 300. \]

\[ 2(400 - \beta_1 - \beta_3)(-1) + 2(200 - \beta_2 - \beta_3)(-1) = 0 \Rightarrow \beta_1 + \beta_2 + 2\beta_3 = 600. \]

Solve for the betas.

b) i. constant variance. However, the variance is often a function of the mean.

ii. The components of the response variable are normally distributed.

For example, the response variable may be restricted to non-negative values, violating normality.

iii. Additivity of effects. Many factors in reality have multiplicative effects.

Comment: GLMs relax all of the three assumptions in part (b).

In the additive model in the question, we are taking Rural as the base; we have three categorical variables that each can take on the values zero or one, although when \( X_1 = 1 \) we must have \( X_2 = 0 \) and vice-versa.

The solution to the three equations is: \( \beta_1 = 525/2, \beta_2 = 175/2, \) and \( \beta_3 = 125. \)

The resulting estimates are:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>525/2 + 125 = 387.5</td>
<td>525/2 = 262.5</td>
</tr>
<tr>
<td>Female</td>
<td>175/2 + 125 = 212.5</td>
<td>175/2 = 87.5</td>
</tr>
</tbody>
</table>

The corresponding minimum sum of squared errors is:

\((400 - 387.5)^2 + (250 - 262.5)^2 + (200 - 212.5)^2 + (100 - 87.5)^2 = 625.\)
3.179. a. As per the exam question, take Male (X₁), Female (X₂), Urban (X₃).

Then the design matrix, X, and response vector Y are:

\[
X = \begin{pmatrix}
1 & 0 & 1 \\
1 & 0 & 0 \\
0 & 1 & 1 \\
0 & 1 & 0 \\
\end{pmatrix} \quad \text{Male Urban} \\
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
\end{pmatrix} \quad \text{Male Rural} \\
\begin{pmatrix}
0 & 1 & 1 \\
0 & 1 & 0 \\
1 & 0 & 1 \\
\end{pmatrix} \quad \text{Female Urban} \\
\begin{pmatrix}
0 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 0 \\
\end{pmatrix} \quad \text{Female Rural}
\]

\[
Y = \begin{pmatrix}
400 \\
250 \\
200 \\
100 \\
\end{pmatrix}
\]

The vector of parameters is:

\[
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\end{pmatrix}
\]

b. For the Gamma Distribution, \( f(y) = \theta^{-\alpha} y^{\alpha-1} e^{-y/\theta} / \Gamma(\alpha) \).

\[
\ln f(y) = (\alpha-1)\ln(y) - y/\theta - \alpha \ln(\theta) - \ln[\Gamma(\alpha)] = (\alpha-1)\ln(y) - y/(\mu/\alpha) - \alpha \ln(\mu/\alpha) - \ln[\Gamma(\alpha)]
\]

\[
= (\alpha-1)\ln(y) - \alpha y/\mu - \alpha \ln(\mu) + \alpha \ln(\alpha) - \ln[\Gamma(\alpha)].
\]

With the identity link function: \( \mu = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \).

Thus the loglikelihood is:

\[
(\alpha-1)\ln(400) - \alpha 400/(\beta_1 + \beta_3) - \alpha \ln(\beta_1 + \beta_3) + \alpha \ln(\alpha) - \ln[\Gamma(\alpha)] + \\
(\alpha-1)\ln(250) - \alpha 250/(\beta_1) - \alpha \ln(\beta_1) + \alpha \ln(\alpha) - \ln[\Gamma(\alpha)] + \\
(\alpha-1)\ln(200) - \alpha 200/(\beta_2 + \beta_3) - \alpha \ln(\beta_2 + \beta_3) + \alpha \ln(\alpha) - \ln[\Gamma(\alpha)] + \\
(\alpha-1)\ln(100) - \alpha 100/(\beta_2) - \alpha \ln(\beta_2) + \alpha \ln(\alpha) - \ln[\Gamma(\alpha)].
\]

Setting the partial derivative with respect to \( \alpha_1 \) equal to zero:

\[
0 = \alpha 400/(\beta_1 + \beta_3)^2 - \alpha/(\beta_1 + \beta_3) + \alpha 250/(\beta_1)^2 - \alpha/(\beta_1). \Rightarrow \\
400/(\beta_1 + \beta_3)^2 + 250/\beta_1^2 = 1/(\beta_1 + \beta_3) + 1/\beta_1.
\]

Setting the partial derivative with respect to \( \beta_2 \) equal to zero:

\[
0 = \alpha 200/(\beta_2 + \beta_3)^2 - \alpha/(\beta_2 + \beta_3) + \alpha 100/(\beta_2)^2 - \alpha/(\beta_2). \Rightarrow \\
200/(\beta_2 + \beta_3)^2 + 100/\beta_2^2 = 1/(\beta_2 + \beta_3) + 1/\beta_2.
\]

Setting the partial derivative with respect to \( \beta_3 \) equal to zero:

\[
0 = \alpha 400/(\beta_1 + \beta_3)^2 - \alpha/(\beta_1 + \beta_3) + \alpha 200/(\beta_2 + \beta_3)^2 - \alpha/(\beta_2 + \beta_3). \Rightarrow \\
400/(\beta_1 + \beta_3)^2 + 200/(\beta_2 + \beta_3)^2 = 1/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3).
\]
c. For the Gamma Distribution, \( f(y) = \theta^{-\alpha} y^{\alpha-1} e^{-y/\theta} / \Gamma(\alpha) \).

\[
\ln f(y) = (\alpha-1)\ln(y) - \frac{y}{\theta} - \alpha \ln(\theta) - \ln[\Gamma(\alpha)] = (\alpha-1)\ln(y) - \frac{y/\mu}{\alpha} - \alpha \ln(\mu/\alpha) - \ln[\Gamma(\alpha)]
\]

With the inverse link function: \( 1/\mu = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \).

Thus the loglikelihood is:

\[
(\alpha-1)\ln(400) - \alpha \ln(\beta_1 + \beta_3) + \alpha \ln(\beta_1 + \beta_3) + \alpha \ln(\alpha) - \ln[\Gamma(\alpha)] + \alpha \ln(\beta_1) + \alpha \ln(\alpha) - \ln[\Gamma(\alpha)] + \alpha \ln(\beta_2) + \alpha \ln(\alpha) - \ln[\Gamma(\alpha)] + (\alpha-1)\ln(100) - \alpha \ln(\beta_2) + \alpha \ln(\alpha) - \ln[\Gamma(\alpha)].
\]

Setting the partial derivative with respect to \( \beta_1 \) equal to zero:

\[
0 = -\alpha 400 + \alpha/(\beta_1 + \beta_3) - \alpha 250 + \alpha/(\beta_1). \Rightarrow 650 = 1/(\beta_1 + \beta_3) + 1/\beta_1.
\]

Setting the partial derivative with respect to \( \beta_2 \) equal to zero:

\[
0 = -\alpha 200 + \alpha/(\beta_2 + \beta_3) - \alpha 100 + \alpha/(\beta_2). \Rightarrow 300 = 1/(\beta_2 + \beta_3) + 1/\beta_2.
\]

Setting the partial derivative with respect to \( \beta_3 \) equal to zero:

\[
0 = -\alpha 400 + \alpha/(\beta_1 + \beta_3) - \alpha 200 - \alpha/(\beta_2 + \beta_3). \Rightarrow 600 = 1/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3).
\]
(d) \( f(x) = \sqrt{\frac{\theta}{2\pi}} \exp\left[ -\frac{\left( x - \frac{1}{\mu} \right)^2}{2x} \right] x^{1.5} \).

Ignoring terms that do not involve \( \mu \), \( \ln f(x) = -\frac{\theta \left( \frac{x}{\mu} - 1 \right)^2}{2x} = -\frac{\theta}{2x} \left( \frac{x^2}{\mu^2} - 2 \frac{x}{\mu} + 1 \right) \)

\[ = -\frac{\theta x}{2\mu^2} + \frac{\theta}{\mu} - \frac{\theta}{2x}. \]

Using the squared reciprocal link function: \( 1/\mu^2 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \).

Thus ignoring terms that do not include \( \mu \), the loglikelihood is:

\[ -\frac{\theta}{2} \{ 400(\beta_1 + \beta_3) + 250(\beta_1) + 200(\beta_2 + \beta_3) + 100(\beta_2) \} + \theta \left( \sqrt{\beta_1 + \beta_3} + \sqrt{\beta_1} + \sqrt{\beta_2 + \beta_3} + \sqrt{\beta_2} \right). \]

Setting the partial derivative with respect to \( \beta_1 \) equal to zero:

\[ 0 = -\frac{\theta}{2} \{ 400 + 250 \} + \frac{\theta}{2} \{ 1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_1} \}. \Rightarrow 650 = 1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_1}. \]

Setting the partial derivative with respect to \( \beta_2 \) equal to zero:

\[ 0 = -\frac{\theta}{2} \{ 200 + 100 \} + \frac{\theta}{2} \{ 1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2} \}. \Rightarrow 300 = 1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2}. \]

Setting the partial derivative with respect to \( \beta_3 \) equal to zero:

\[ 0 = -\frac{\theta}{2} \{ 400 + 200 \} + \frac{\theta}{2} \{ 1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_2 + \beta_3} \}. \Rightarrow 600 = 1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_2 + \beta_3}. \]

Comment: Using a computer, the fitted parameters in part b are:

\( \beta_1 = 263.236, \beta_2 = 98.160, \beta_3 = 110.129. \)

For example, the fitted value for Female and Urban is: 98.160 + 110.129 = 208.29.

The fitted values in part b are:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>373.36</td>
<td>263.24</td>
</tr>
<tr>
<td>Female</td>
<td>208.29</td>
<td>98.16</td>
</tr>
</tbody>
</table>

Using a computer, the fitted parameters in part c are:

\( \beta_1 = 0.00447623, \beta_2 = 0.00789904, \beta_3 = -0.0021321. \)

For example, the fitted value for Female and Urban is: \( 1/ (0.00789904 - 0.0021321) = 173.40. \)
The fitted values in part c are:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Urban</th>
<th>Rural</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>426.60</td>
<td>223.40</td>
<td>650</td>
</tr>
<tr>
<td>Female</td>
<td>173.40</td>
<td>126.60</td>
<td>300</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>350</td>
<td>950</td>
</tr>
</tbody>
</table>

The totals for the fitted match those for the data.

In general, the estimates will be in balance as they were here, when one uses the canonical link function; the canonical link function for the Gamma is the reciprocal link function.


Note that when the weights differ by cell, this balance involves weighted averages.

Using a computer, the fitted parameters in part d are:

\[ \beta_1 = 0.0000218789, \beta_2 = 0.000053899, \beta_3 = -0.0000166235. \]

For example, the fitted value for Female and Urban is:

\[ 1 / \sqrt{0.000053899 - 0.0000166235} = 163.79. \]

The fitted values in part d are:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Urban</th>
<th>Rural</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>436.21</td>
<td>213.79</td>
<td>650</td>
</tr>
<tr>
<td>Female</td>
<td>163.79</td>
<td>136.21</td>
<td>300</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>350</td>
<td>950</td>
</tr>
</tbody>
</table>

Since the canonical link function for the Inverse Gaussian is the squared reciprocal link function, again the totals for the fitted match those for the data.
3.180.  a. Let $\beta_1$ represent territory A.  Let $\beta_2$ represent territory B.  
Let $\beta_3$ represent private passenger.  Let $\beta_4$ represent light trucks.
(These are not the only choices.  We have chosen medium trucks as the base level.)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Row 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Row 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first row of (1, 0, 1, 0) corresponds to the cell for Territory A ($\beta_1$) and private passenger ($\beta_3$).
The second row of (1, 0, 0, 1) corresponds to the cell for Territory A ($\beta_1$) and light truck ($\beta_4$).
The third row of (1, 0, 0, 0) corresponds to the cell for Territory A ($\beta_1$) and medium truck.
(There are other ways to arrange the design matrix.)
The corresponding vector of betas (parameters) is:

$\beta_1 + \beta_3, \beta_1 + \beta_4, \beta_1, \beta_2 + \beta_3, \beta_2 + \beta_4, \beta_2$.

b. For a poisson error structure variance is a function of the expected value, while under the gamma error structure the variance of an observation is a function of its mean squared.
c. Determine the form of the density for the chosen error structure (distribution of errors.)
Using this density and the chosen link function take a product of the chances of the observations; this is the likelihood as a function of the parameters.
Maximize the log of the likelihood function by setting the partial derivatives with respect to each of the parameters equal to zero.
Solve the resulting system of equations for the fitted parameters.
Compute the predicted values.

Comment: In part (a) one could have instead for example taken:
Let $\beta_1$ be an intercept. Let $\beta_2$ represent territory A.
Let $\beta_3$ represent light trucks. Let $\beta_4$ represent medium trucks.

In that case, we have taken Territory B / Heavy Trucks as the base level.
Instead, other combinations of territory and truck weight could have been chosen as the base level.
If we use an intercept term, then we can have only one coefficient for territory, and two coefficients for vehicle type. Including the intercept, we still have a total of four coefficients in our model.
In more complicated situations one would not be able to solve the equations for the parameters in closed form. Fortunately, there are commercial packages of computer software specifically designed to solve and analyze GLMs.
3.181. One-way analysis doesn’t consider:
1. Correlations between rating variables, in other words correlations of exposures by cell.
   For example, young people drive older cars more often. Worse loss ratios for older cars can be
   partially driven by the larger proportion of youthful drivers.
   For example, age may be correlated with territory if a greater proportion of senior citizens live in
   certain parts of a state. The relative loss ratios for such territories will be better due to the higher
   proportion of drivers who are senior citizens.
2. Interdependencies among rating variables.
   For example, the rate differentials between male and female drivers vary by age.
   For example, young drivers who have expensive cars may be poor risks, but old drivers who
   have expensive cars may be good risks.
3.182. a) In the absence of any other information, I would choose a Poisson error function which is commonly used for frequency. The frequencies look like they might follow a multiplicative model; the ratios of the columns look kind of similar and the ratios of the rows look kind of similar.  
(In contrast, the differences in the columns look kind of different and the differences in the rows look kind of different. Thus I would not choose an additive model and the identity link function.) Therefore, I will use a log link function corresponding to a multiplicative model.

\[ g(x) = \ln(x) \quad g^{-1}(x) = e^x. \]

One needs to pick a base level. It is likely that Yes/Yes has the most exposures, so I will pick that as the base level.

The vector of model parameters:
Let \( \beta_0 \) be the intercept term, which is a parameter which applies to all observations.
Let \( \beta_1 \) correspond to no for homeowners.
Let \( \beta_2 \) correspond to no for auto policy.

Then the response vector would be:
\[
\begin{pmatrix}
3 \\
5 \\
8 \\
12
\end{pmatrix}
\]

The design matrix would be:
\[
\begin{pmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1
\end{pmatrix}
\]
\[
X_1 \left\{ \begin{array}{l}
\text{No H.O. = 1} \\
\text{Otherwise = 0}
\end{array} \right.
\quad X_2 \left\{ \begin{array}{l}
\text{No Auto = 1} \\
\text{Otherwise = 0}
\end{array} \right.
\]

Alternately, the vector of model parameters: Let \( \beta_1 \) correspond to yes for auto.
Let \( \beta_2 \) correspond to no for auto. Let \( \beta_3 \) correspond to yes for homeowners.

Then the response vector would be:
\[
\begin{pmatrix}
3 \\
5 \\
8 \\
12
\end{pmatrix}
\]
The design matrix would be:

\[
\begin{pmatrix}
1 & 0 & 1 \\
1 & 0 & 0 \\
0 & 1 & 1 \\
0 & 1 & 0 \\
\end{pmatrix}
\]

\[X_1 \leftrightarrow \begin{cases} 
\text{Yes Auto} = 1 \\
\text{Otherwise} = 0
\end{cases} \quad X_2 \leftrightarrow \begin{cases} 
\text{No Auto} = 1 \\
\text{Otherwise} = 0
\end{cases} \quad X_3 \leftrightarrow \begin{cases} 
\text{Yes H.O.} = 1 \\
\text{Otherwise} = 0
\end{cases} \]

b) Missing data can lead to aliasing. For the missing data, if in most cases both whether it had an auto and homeowner policy is missing, then there is the potential problem of aliasing or near aliasing. No data auto and no data homeowners would be either perfectly or highly correlated. With aliasing the model parameters make no sense. Near aliasing creates problems with convergence of the model.

A solution would be to exclude these missing data records from modeling. Another solution is to eliminate the unknown level from one of the factors so there are no linear dependencies. In other words, one further covariate needs to be removed; this could either be the “unknown” auto covariate or the “unknown” homeowners covariate.

Comment: One can make other choices in part (a) for the vector of parameters and get full credit provided the design matrix is consistent.
3.183. a) Increase of 7.0% \(\Leftrightarrow\) -0.4172. 2 phone calls \(\Leftrightarrow\) -0.4239.

\[ -0.4172 - 0.4239 + 1.793 = 0.9519. \]

Using the logistic model, probability of renewal is: \(\exp(0.9519) / \{1 + \exp(0.9519)\} = 72.15\%.\)

b) I would not use this strategy:

1. There is no reasonable connection between insurance losses and the number of times an insured calls an insurer. Charging those who call the insurer more would not be acceptable to the public.

Insurance regulators are very unlikely to allow the use of this rating variable.

While causality is not required, this is an extreme case of the opposite of causality.

2. The proposed variable is easily manipulated by the insured.

3. The proposed variable lacks constancy; the number of phone calls from an insured is likely to change from year to year.

4. We do not know why the rate of renewal decreases with number of phone calls made by the insured to the insurer. Could this be because when they call, insureds get impolite or incompetent service? In that case, a better strategy would be to improve the insurer’s service so that they do not lose so many customers.

5. The number of phone calls is likely related to other variables which are more directly related to renewal probability, such as moving or age. The actuary should go back and try to find variables that are the underlying reasons for the model results.

6. If many of these calls are from insureds who are making claims, then perhaps the lower renewal is due to poor claims service. It would be a better strategy to improve claims service.

7. By raising the rate of insureds who made phone calls, you are making their future renewal rate even lower. The insurer is likely to lose a lot of insureds if it followed this strategy.

Alternately, I would use this strategy:

1. When pricing based on the lifetime of a policy, it makes sense to take into account the expected renewal rate. Those with a lower expected renewal rate, such as those who make several phone calls to the insurer, should be charged more, all else being equal.

2. Those who call the insurer more often are likely to be reporting a claim. Those with claims in the past, have a higher expected future claim frequency. So it makes sense to charge those with more calls more, since their future average claim frequency is higher than average.

3. Those who call more often in the past are more likely to call more often in the future, resulting in higher expense for the insurer.

4. Those who call the insurer more often are more likely to make a small claim when they suffer a small loss, and thus have higher expected future loses.

Comment: I found it much, much easier in this case, to argue against using the strategy.

(My reasons in favor other than the first, have nothing to do with the given model.)

On your exam, pick whichever side of the argument allows you to quickly come up with two good reasons.

Without diagnostics there is no way to check the statistical significance of the modeled result. Some of the extra phone calls may be from insureds who got big increases and are calling to complain or to see if this insurer will match a quote from another insurer. Thus the two variables in the model may be correlated.
3.184. (a) The first column refers to $\beta_1$ whether or not we have a male, the second column refers to $\beta_2$ whether or not we are in Territory A, the third column refers to $\beta_3$ the intercept, and thus is all ones.

$$X = \begin{pmatrix}
1 & 1 & 1 \\
0 & 1 & 1 \\
1 & 0 & 1 \\
0 & 0 & 1 \\
\end{pmatrix}$$

$\quad$ or $X = \begin{pmatrix}
1 & 1 & 1 \\
1 & 0 & 1 \\
0 & 1 & 1 \\
0 & 0 & 1 \\
\end{pmatrix}$

(b) $Y = \begin{pmatrix}
700/1400 \\
400/1000 \\
600/1000 \\
420/1200 \\
\end{pmatrix} = \begin{pmatrix}
0.50 \\
0.40 \\
0.60 \\
0.35 \\
\end{pmatrix}$

$\quad$ or $Y = \begin{pmatrix}
0.50 \\
0.60 \\
0.40 \\
0.35 \\
\end{pmatrix}$

(c) With a Normal error function and an identity link function, this is the same as a multiple regression.
Assuming $\beta_3 = 0.35$, then the squared error is:

$$1400 (\beta_1 + \beta_2 + 0.35 - 0.5)^2 + 1000 (\beta_2 + 0.35 - 0.4)^2 + 1000 (\beta_1 + 0.35 - 0.6)^2 + 1200 (0.35 - 0.35)^2 =$$

$$1400 (\beta_1 + \beta_2 - 0.15)^2 + 1000 (\beta_2 - 0.05)^2 + 1000 (\beta_1 - 0.25)^2.$$ Setting the partial derivative with respect to $\beta_1$ equal to zero:

$$0 = 2800(\beta_1 + \beta_2 - 0.15) + 2000(\beta_1 - 0.25). \implies 4800 \beta_1 + 2800 \beta_2 = 920.$$ Setting the partial derivative with respect to $\beta_2$ equal to zero:

$$0 = 2800(\beta_1 + \beta_2 - 0.15) + 2000(\beta_2 - 0.05). \implies 2800 \beta_1 + 4800 \beta_2 = 520.$$ \Rightarrow \beta_2 = (520 - 2800 \beta_1) / 4800. \implies \beta_2 = -0.0052.

Plugging back into the first equation: $4800 \beta_1 + 2800 (520 - 2800 \beta_1) / 4800 = 920.$ \implies \beta_1 = 0.1947. \implies \beta_2 = -0.0052.
Alternately, without taking into account exposures by cell, the squared error is:

\[(\beta_1 + \beta_2 + 0.35 - 0.5)^2 + (\beta_2 + 0.35 - 0.4)^2 + (\beta_1 + 0.35 - 0.6)^2 + (0.35 - 0.35)^2 =
(\beta_1 + \beta_2 - 0.15)^2 + (\beta_2 - 0.05)^2 + (\beta_1 - 0.25)^2.\]

Setting the partial derivative with respect to \(\beta_1\) equal to zero:

\[0 = 2(\beta_1 + \beta_2 - 0.15) + 2(\beta_1 - 0.25). \Rightarrow 4 \beta_1 + 2 \beta_2 = 0.8.\]

Setting the partial derivative with respect to \(\beta_2\) equal to zero:

\[0 = 2(\beta_1 + \beta_2 - 0.15) + 2(\beta_2 - 0.05). \Rightarrow 2 \beta_1 + 4 \beta_2 = 0.4.\]

\[\Rightarrow \beta_2 = (0.4 - 2 \beta_1) / 4 = 0.1 - 0.5 \beta_1.\]

Plugging back into the first equation: \(4 \beta_1 + 2 (0.1 - 0.5 \beta_1) = 0.8.\)

\[\Rightarrow \beta_1 = 0.2. \Rightarrow \beta_2 = 0.\]

Alternately, in either case one can fit via maximum likelihood and get the same result as by minimizing the squared errors.

For the Normal Distribution, \(f(x) = \frac{\exp[- (x-\mu)^2 / 2\sigma^2]}{\sigma \sqrt{2\pi}}.\)

\[\ln [f(x)] = - \frac{(x-\mu)^2}{2\sigma^2} - \ln[\sigma] - \ln[2\pi]/2.\]

Without taking info account exposures by cell, the loglikelihood is:

\[- (\beta_1 + \beta_2 + 0.35 - 0.5) / (2\sigma^2) - (\beta_1 + 0.35 - 0.4) / (2\sigma^2) - (\beta_2 + 0.35 - 0.6) / (2\sigma^2)
- (0.35 - 0.35) / (2\sigma^2) - 4 / \sigma.\]

Setting the partial derivative of the loglikelihood with respect to \(\beta_1\) equal to zero:

\[0 = -(\beta_1 + \beta_2 - 0.15) / \sigma^2 - (\beta_1 - 0.25) / \sigma^2. \Rightarrow 2 \beta_1 + \beta_2 = 0.4.\]

Setting the partial derivative with respect to \(\beta_2\) equal to zero:

\[0 = -(\beta_1 + \beta_2 - 0.15) / \sigma^2 - (\beta_2 - 0.05) / \sigma^2. \Rightarrow \beta_1 + 2 \beta_2 = 0.2.\]

Solving two equations in two unknowns: \(\beta_1 = 0.2, \text{ and } \beta_2 = 0.\)

Setting the partial derivative with respect to \(\sigma\) equal to zero:

\[0 = (\beta_1 + \beta_2 - 0.15)^2 / \sigma^3 - (\beta_1 - 0.05)^2 / \sigma^3 - (\beta_2 - 0.25)^2 / \sigma^3 - 4 / \sigma.\]

\[\Rightarrow \sigma^2 = \{(0.2 + 0 - 0.15)^2 + (0.2 + 0 - 0.05)^2 + (0.2 + 0 - 0.25)^2\} / 4 = 0.006875.\]
(d) 1. The Normal Distribution allows negative values, while the Poisson Distribution does not. Since claim frequencies are never negative, the Poisson error structure is preferable here.

2. The Normal error structure assumes that the process variances of the frequency in the four cells are equal. In contrast, a Poisson error structure assumes that the process variances of the frequency in the four cells are equal to their means. I would expect that those cells with higher expected frequencies would have higher process variances than those with lower expected frequencies, and thus would prefer the Poisson error structure to the Normal.

3. The log link function would assume multiplicative relativities, while the identity link function assumes additive relativities. If the relationship is approximately multiplicative, then the log link function would do a better job than the identity link function.

Comment: In parts (a) and (b), the order in which one lists the rows is arbitrary; it would be a good idea to label what you did.

In part (c), if one includes the exposures in the sum of squared errors, that is equivalent to using exposures as the prior weights in the GLM, or using exposures in an offset term. Including the exposures is equivalent to doing a weighted multiple regression.

The observed frequencies are:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Territory A</th>
<th>Territory B</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.50</td>
<td>0.60</td>
<td>0.10</td>
</tr>
<tr>
<td>Female</td>
<td>0.40</td>
<td>0.35</td>
<td>-0.05</td>
</tr>
<tr>
<td>Difference</td>
<td>-0.10</td>
<td>-0.25</td>
<td></td>
</tr>
</tbody>
</table>

The differences between territories are not similar for the two genders. The differences between genders are not similar for the two territories.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Territory A</th>
<th>Territory B</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.50</td>
<td>0.60</td>
<td>1.2</td>
</tr>
<tr>
<td>Female</td>
<td>0.40</td>
<td>0.35</td>
<td>0.875</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.80</td>
<td>0.583</td>
<td></td>
</tr>
</tbody>
</table>

The ratios between territories are not similar for the two genders. The ratios between genders are not similar for the two territories. Thus perhaps neither an additive nor a multiplicative relationship is appropriate.
3.185. (a) $\phi$ is the scale or dispersion parameter, which scales the variance. $\omega_i$ is a (prior) weight, representing the amount of data we have for observation $i$; the variance is inversely proportional to the volume of data.

(b) i. Gamma Distribution is most commonly used to model the error structure for severity; it works well in many situations based on diagnostics. The Gamma is continuous with support from zero to infinity. The gamma distribution also has an intuitively attractive property for modeling claim amounts since it is invariant to measures of currency. In other words measuring severities in dollars and measuring severities in cents will yield the same results using a gamma multiplicative GLM. (This is not true of some other distributions such as Poisson, but would be for the Inverse Gaussian.)

For the Gamma: $V(\mu_i) = \mu_i^2$.  

ii. For policy renewal a Bernoulli or Binomial is used, since policy renewal is a yes/no process. For the Bernoulli: $V(\mu_i) = \mu_i (1 - \mu_i)$.  

For the Binomial representing $\mu$ trials ($\mu$ policies): $V(\mu_i) = \mu_i (1 - \mu_i) / m$.  

(c) 1. For severity, $\omega_i$ would be the number of claims, the measure of how much data we have.  

2. For policy renewal, if using the Bernoulli, $\omega_i$ would be the number of policies. If using the Binomial, $\omega_i = 1$.

3.186. Smaller Bayesian Information Criterion is better.  

BIC = -2 (maximum loglikelihood) + $p \ln(n)$, where $n = 1000$ is the sample size and $p$ is the number of parameters. Since the deviance = (2) (saturated max. loglikelihood - maximum likelihood for model), we can compare between the models: Deviance + $p \ln(n) = \text{Deviance} + p \ln(1000)$. (The maximum likelihood for the saturated model is the same in each case.)

<table>
<thead>
<tr>
<th>Model #</th>
<th>$p$</th>
<th>Deviance</th>
<th>Deviance + $p \ln(1000)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1085.0</td>
<td>1098.82</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1084.8</td>
<td>1105.52</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1083.0</td>
<td>1103.72</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1081.9</td>
<td>1109.53</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1081.6</td>
<td>1116.14</td>
</tr>
</tbody>
</table>

The smallest Deviance + $p \ln(n)$ is for Model 1.
3.187. Smaller Akaike Information Criterion is better.
AIC = -2 (maximum loglikelihood) + (number of parameters)(2).
Since the deviance = (2) (saturated max. loglikelihood - maximum likelihood for model),
we can compare between the models:
Deviance + p 2 = Deviance + (number of parameters)(2).
(The maximum likelihood for the saturated model is the same in each case.)

<table>
<thead>
<tr>
<th>Model #</th>
<th>p</th>
<th>Deviance</th>
<th>Deviance + (number of parameters)(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1085.0</td>
<td>1089.0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1084.8</td>
<td>1090.8</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1083.0</td>
<td>1089.0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1081.9</td>
<td>1089.9</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1081.6</td>
<td>1091.6</td>
</tr>
</tbody>
</table>

The smallest AIC is a tie between Model 1 and Model 3.

3.188. Estimated mean severity for a rural male is: \( \exp[2.32 - 0.64 + 0.76] = 11.473 \).
For the Gamma Distribution, \( \text{Var}[Y] = \phi \mu^2 = (2)(11.473^2) = 263.3 \).

3.189. \( \exp[\beta x] = \exp[-1.485 + 0 - 1.175 - 0.101] = e^{-2.761} = 0.06323 \).
For the logit link function: \( \mu = \frac{e^{\beta x}}{e^{\beta x} + 1} = 0.06323 / (0.06323 + 1) = 5.95\% \).

3.190. \( \mu = \exp[-2.633 + 0.132 + 0] = 0.07957 \).

3.191.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle Price</td>
<td>4</td>
</tr>
<tr>
<td>Driver age</td>
<td>2 - 1 = 1</td>
</tr>
<tr>
<td>Number of drivers</td>
<td>4 - 1 = 3</td>
</tr>
<tr>
<td>Gender</td>
<td>2 - 1 = 1</td>
</tr>
<tr>
<td>Interaction Gender &amp; Driver Age</td>
<td>1</td>
</tr>
</tbody>
</table>

Maximum number of parameter is: 4 + 1 + 3 + 1 + 1 = 10.
Comment: A model with only Vehicle Price would involve: \( \beta_0 + \beta_1 (\text{vp}) + \beta_2 (\text{vp})^2 + \beta_3 (\text{vp})^3 \).
The interaction of gender and driver age only uses one parameter since each of gender and
driver age only use one parameter.

3.192. Smaller AIC is better, so we prefer Model 1.
\( \exp[\beta x] = \exp[-3.264 + (12)(0.212) + 0.727] = e^{0.007} = 1.007 \).
For the logit link function: \( \mu = \frac{e^{\beta x}}{e^{\beta x} + 1} = 1.007 / (1.007 + 1) = 50.2\% \).
3.193. Let $x$ be the number of additional parameters for the new model.
Let $\ell_1$ be the loglikelihood for the original model, and $\ell_2$ be the loglikelihood for the model including the new variable.

Deviance = (2) (saturated max. loglikelihood - maximum likelihood for model).
Thus the change in model deviance is: $-2(\ell_2 - \ell_1) = -53$.

AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
Thus the change in AIC is: $-2(\ell_2 - \ell_1) + 2x = -53 + 2x = -47$. $\Rightarrow x = 3$.

BIC = (-2) (maximum loglikelihood) + (number of parameters) $\ln$(number of data points).
Thus the change in BIC is: $-2(\ell_2 - \ell_1) + x \ln(n) = -53 + 3 \ln(n) = -32$. $\Rightarrow n = e^7 = 1097$.

3.194. We have to assume equal exposures in each of the four cells.
The mean modeled frequencies are:

<table>
<thead>
<tr>
<th>State</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\exp[\beta_1 + \beta_3]$</td>
<td>$\exp[\beta_2 + \beta_3]$</td>
</tr>
<tr>
<td>B</td>
<td>$\exp[\beta_1]$</td>
<td>$\exp[\beta_2]$</td>
</tr>
</tbody>
</table>

The loglikelihood ignoring terms that do not depend on the betas is:

$-\exp[\beta_1 + \beta_3] + 0.0920 (\beta_1 + \beta_3) - \exp[\beta_2 + \beta_3] + 0.1500 (\beta_2 + \beta_3) - \exp[\beta_1] + 0.0267 \beta_1 - \exp[\beta_2] + 0.0500 \beta_2$.

Setting the partial derivative of the loglikelihood with respect to $\beta_1$ equal to zero:

$-\exp[\beta_1 + \beta_3] + 0.0920 - \exp[\beta_1] + 0.0267 = 0$.

Given $\beta_3 = 1.149$: $-\exp[\beta_1] e^{1.149} + 0.0920 - \exp[\beta_1] + 0.0267 = 0$.

$\Rightarrow \exp[\beta_1] = (0.0920 + 0.0267) / (1 + e^{1.149}) = 0.02857$.

$\Rightarrow \exp[\beta_1 + \beta_3] = 0.02857 e^{1.149} = 0.0901 = \text{expected frequency of a male risk in State A}$.

Comment: Similar to 8, 11/13, Q.2c.

What the exam questions calls “the likelihood function” is the loglikelihood function.

$\hat{\beta}_1 = \ln(0.02857) = -3.555$.

Setting the partial derivative of the loglikelihood with respect to $\beta_2$ equal to zero:

$-\exp[\beta_2 + \beta_3] + 0.1500 - \exp[\beta_2] + 0.0500 = 0$.

Given $\beta_3 = 1.149$: $-\exp[\beta_2] e^{1.149} + 0.1500 - \exp[\beta_2] + 0.0500 = 0$.

$\Rightarrow \exp[\beta_2] = (0.1500 + 0.0500) / (1 + e^{1.149}) = 0.04813$. $\Rightarrow \hat{\beta}_2 = -3.034$.

Using a computer, without being given $\beta_3$, the maximum likelihood fit is:

$\hat{\beta}_1 = -3.5555$, $\hat{\beta}_2 = -3.0338$, and $\hat{\beta}_3 = 1.1490$.

The mean modeled frequencies are:

<table>
<thead>
<tr>
<th>State</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\exp[-3.5555 + 1.1490] = 9.01%$</td>
<td>$\exp[-3.5555] = 2.86%$</td>
</tr>
<tr>
<td>B</td>
<td>$\exp[-3.0338 + 1.1490] = 15.19%$</td>
<td>$\exp[-3.0338] = 4.81%$</td>
</tr>
</tbody>
</table>
3.195. In order to solve for the unknown intercept, we use the given probability of accident for a driver in age group 2, from area C and with vehicle body type Other.

\[
0.22 = \exp[x + 0.064 - 0.371] / \{\exp[x + 0.064 - 0.371] + 1\}.
\]

\[
\Rightarrow \exp[x + 0.064 - 0.371] = 0.22 / (1 - 0.22) = 0.28205.
\]

\[
\Rightarrow x + 0.064 - 0.371 = \ln(0.28205) = -1.2657. \Rightarrow x = -0.9587.
\]

Thus for a driver in age group 3, from area C and with vehicle body type Sedan, the odds (ratio) is:

\[
\pi / (1 - \pi) = \exp[-0.9857 + 0 + 0 + 0] = 0.3834.
\]

Comment: The probability of having an accident for a driver in age group 3, from area C and with vehicle body type Sedan is: \(0.3834 / (1 + 0.3834) = 0.277\).

Note that \(0.277 / (1 - 0.277) = 0.383\).

3.196. \[
\frac{\exp[-2.358 + 0.905]}{1 + \exp[-2.358 + 0.905]} = 0.190.
\]

3.197. \[
\]

3.198. For an observation from Zone 4, with Vehicle Class Sedan and Driver Age Middle age, the mean is: \(\exp[2.1] = 8.166\).

For the Gamma Distribution the variance is: \(\phi \mu^2 = (1) (8.166^2) = 66.7\).

3.199. \[
\frac{\exp[1.530 + 0.735 - 0.031]}{1 + \exp[1.530 + 0.735 - 0.031]} = 90.33\%.
\]

3.200. Since Model 2 has one fewer parameter than model 3, model 2 has 9 degrees of freedom.

\[
\begin{align*}
\text{AIC} &= (-2) (\text{maximum loglikelihood}) + (\text{number of parameters})(2), \\
\text{BIC} &= (-2) (\text{maximum loglikelihood}) + (\text{number of parameters}) \ln(\text{number of data points}).
\end{align*}
\]

Therefore from Model 1: \(95,473.61182 = (-2)(-47,704) + (5) \ln(n). \Rightarrow n = 500,000\).

For Model 2, \(\text{AIC} = -47,495 + (9)(2)\).

For Model 2, \(\text{BIC} = -47,495 + (9) \ln(500,000)\).

The absolute difference between the AIC and the BIC for Model 2 is: \(1 (9) \ln(500,000) - 18 I = 100.1\).

3.201. Graph one shows an increasing variance with fitted value. Homoscedasticity would be constant variance, so statement 1 is false; statement 2 is true.

The residuals in Graph 2 are not symmetric around zero; there are more extreme positive values than there are extreme negative values. This indicates that the residuals are not normally distributed.

Statement 3 is true.

Comment: In Graph 2 it is not clear the meaning of the horizontal lines. A Normal Q-Q Plot would have been much more useful than Graph 2.
3.202. The model with the smallest AIC is usually the best model in model selection process, all other things being equal. Statement A is not true.
The model with the smallest BIC is usually the best model in model selection process, all other things being equal. Statement B is not true.
The model with the smallest deviance is usually the best model in model selection process, all other things being equal. Statement C is not true.
AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
BIC = (-2) (maximum loglikelihood)) + (number of parameters) ln(number of data points).
The penalty for AIC is (number of parameters)(2).
The penalty for BIC is (number of parameters) ln(number of data points).
So the penalties are equal for: 2 = ln(number data points). ⇒ number of data points = e^2 = 7.4.
Thus, other things equal, when number of observations ≥ 8, BIC penalizes more for the number of parameters used in the model than AIC. Thus statement E is true.
Comment: Since statements D and E are opposites, it is likely that one of them is true.

3.203. Change in AIC is: (2) (number of parameters added).
Change in BIC is: ln(1500) (number of parameters added).
We want: ln(1500) (number of parameters added) > (2) (number of parameters added) + 25.
⇒ Number of parameters added > 4.7. ⇒ Number of added parameters is at least 5.
⇒ Minimum possible number of levels in the new categorical variable is: 5 + 1 = 6.

3.204. \[100,000 \exp[-15 - 1.2 + (0.15)(25) + (0.004)(25^2) + (0.012)(25)]\]
\[= 100,000 \, e^{-9.65} = 6.44 \text{ deaths.}\]

3.205. i. Where the variable in question relates to a policy option selected by the insured, having its factor reflect anything other than the excess losses due to higher limit is not a good idea. One can get counterintuitive results such as charging less for more coverage. Even if the indicated result is not counterintuitive, to the extent that the factor differs from the pure effect on loss potential, it will affect the way insureds choose coverage options in the future. Thus, the selection dynamic will change and the past results would not be expected to replicate for new policies. For this reason it is recommended that factors for coverage options such as increased limit factors be estimated outside the GLM, using traditional actuarial techniques. (The resulting factors should then be included in the GLM as an offset.)
ii. I assume what is intended is that the number of coverage changes during the current policy period will be used to help rate the policy during its next policy period. (We are not given any information on whether the number of coverage changes in a policy period is related to the insurance costs the following period compared to otherwise similar insureds.) The number of changes during a given policy period is not a good classification variable. It is something that is likely to be zero for many policy periods, and vary somewhat randomly over time. If those with more coverage changes are changed more it is unlikely to be acceptable to insurance regulators and the public. If those with more coverage changes are changed more, then it will give insureds less incentive to make necessary coverage changes during a policy period; some of these coverage changes would have resulted in additional premiums for the insurer.

Alternately, the information will not be available for new business since we are building a GLM for the prospective period.

Alternately, the number of coverage changes is likely to change from what it is in the current policy period and thereafter year by year.

iii. Territories are not a good fit for the GLM framework. You may have thousands of zipcodes to consider and aggregating them to a manageable level will cause you to lose a great deal of important signal. If one does not aggregate the large number of zipcodes, then there are too many parameters which can lead to overfitting.

Using a spatial smoothing technique would be a more appropriate technique; one would then include the value determined for ZIP code as an offset term in the GLM.

(b) 1. One can get counterintuitive results such as charging more for less coverage.
2. Even if the indicated result is not counterintuitive, to the extent that the factor differs from the pure effect on loss potential, it will affect the way insureds choose coverage options in the future. Thus, the selection dynamic will change and the past results would not be expected to replicate for new policies.
3. Deductibles should lower frequency (small losses below deductible not reported) but usually increase severity (since claims that do get reported are higher average cost). This violates the assumption for the Tweedie Distribution, that a lower pure premium is due to both a lower frequency and a lower severity.

(c) One can calculate deductible relativities from loss elimination ratios.

\[
\text{Deductible Relativity} = \frac{(1 - \text{LER for chosen deductible})}{(1 - \text{LER for Base Deductible})}.
\]

Loss elimination ratios can be estimated from size of loss data.

\[
\text{Loss elimination ratio} = \frac{\text{Limited Expected Value at Deductible Amount}}{\text{Mean}}.
\]

In the GLM, one would then include an offset of \(\ln[\text{deductible relativity}]\).

Comment: While the average size of non-zero payment, equal to the mean residual life, usually increases as the size of deductible increases, this is not always the case.

Deductible factors may produce higher relativities at higher deductibles due to factors other than pure losses elimination:

1. Insureds at high loss potential and high premiums may be more likely to elect high deductibles in order to reduce their premium.
2. Underwriters may force high deductibles on riskier insureds.
3.206.  a. Sort the risks from best to worst based on the model predicted loss.

<table>
<thead>
<tr>
<th>Risk</th>
<th>Model Predicted Loss</th>
<th>Actual Loss</th>
<th>Cumulative Losses</th>
<th>Cumulative % of Losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>200</td>
<td>400</td>
<td>400</td>
<td>8.0%</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>220</td>
<td>620</td>
<td>12.4%</td>
</tr>
<tr>
<td>4</td>
<td>800</td>
<td>850</td>
<td>1,470</td>
<td>29.4%</td>
</tr>
<tr>
<td>3</td>
<td>1,500</td>
<td>1,480</td>
<td>2,950</td>
<td>59.0%</td>
</tr>
<tr>
<td>1</td>
<td>2,000</td>
<td>2,050</td>
<td>5,000</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>5000</td>
<td></td>
</tr>
</tbody>
</table>

On the x-axis, plot the cumulative percentage of exposures.
On the y-axis, plot the cumulative percentage of actual losses.

b. The Gini index is twice the area between the Lorenz Curve and the line of equality. The higher the Gini Index, the better the rating plan is at identifying risk differences, in other words the rating plan has more lift.

“The Gini index can also be used to measure the lift of an insurance rating plan by quantifying its ability to segment the population into the best and worst risks.”

Comment: See Section 7.2.4 including Figure 21 in Goldburd, Khare, and Tevet. Usually, one would be working with thousands of risks.
3.207. (a) \(\exp[0.910 + (3)(0.013) + \ln[25,000] (-0.187) + (8)(0.062)] = e^{-0.4357} = 64.7\%\).

(b) One could take the coefficients of the new business model as a given, other than \(b_0\), which will be re-estimated.

Let the prior year claim count be \(x\) for renewal business.
Then the renewal business model is:
\[\mu = \exp[\beta_0 + 0.013 \text{ age} + (-0.187) \ln \text{prem} + 0.62 \ln \text{locont} + \beta_4 x].\]

We would fit the model via maximum likelihood to the data for renewal business, taking into account the form of density for the Tweedie Distribution.

Alternately, one can fit a single model to the data for new and renewal business.

Let the prior year claim count be \(x\) for renewal business.
Let \(D = 0\) if new business and \(1\) if renewal business.
Then the combined model is:
\[\mu = \exp[\beta_0 + \beta_1 \text{ age} + \beta_2 \ln \text{prem} + \beta_3 \ln \text{locont} + D \beta_4 + D \beta_5 x].\]
We would fit the model via maximum likelihood to the combined data, taking into account the form of density for the Tweedie Distribution.

(c) 1. Time-consistency. One can fit the model to the data for separate years and compare the coefficients. If the fitted coefficients are similar, that indicates stability over time.

Alternately, one could introduce dummy variables into the model for the various years of data.
For example, if we have data from 2012, 2013 and 2104,
then we could take 2012 = base year, \(x_5 = 1\) if 2013, \(x_6 =1\) if 2014.

Then test whether the coefficients of these variables are significantly different from zero. If one or more of the fitted coefficients are significantly different than zero, that indicates instability over time.

2. Bootstrapping. Create multiple datasets from the initial dataset by sampling with replacement. Run the model on each sampled set. Assess stability of estimates of coefficients by comparing the results from each run.

3. Cross-Validation. Split the data into \(k\) parts and run the model on the \((k-1)\) parts, then validate the result on the remaining part. Compare how similar the estimates are from the \(k\) iterations to assess variable stability.

4. Validation on Holdout Dataset. Split the data into two subsets, training and holdout. Determine the best model on the training set. Ideally, this model should fit well the holdout data.

5. Cook’s Distance. Sort the observations based on their Cook’s Distance value (higher distance = more influence on the model.) Remove one or more of the most influential observations and rerun the model on this new set of data to see the effect on estimated parameters.
3.208. (a) For the base model:
AIC = (-2)(-750) + (2)(10) = 1520.
BIC = (-2)(-750) + 10 \ln[1,000,000] = 1638.2.
For the new model:
AIC = (-2)(-737.5) + (2)(15) = 1505.
BIC = (-2)(-737.5) + 15 \ln[1,000,000] = 1682.2.
(b) AIC is preferable. As here, most actuarial models involve a lot of data points. Therefore, the
penalty for more parameters is very large for the BIC. Using BIC will tend to result in too simple
models. In contrast, AIC does not depend on the number of data points.
(c) Based on part (b), I will rely on AIC.
Smaller AIC is better, so I will recommend the new model.
Comment: See Section 6.2.2 in Goldburd, Khare, and Tevet.
If one instead relied on BIC, the base model would be preferred.
Deviance = 2 (loglikelihood of saturated model - loglikelihood of model).
Thus equivalently to using AIC, one could compare models using: Deviance + 2p.
For the base model, Deviance + 2p = 500 + (2)(10) = 520.
For the new model, Deviance + 2p = 475 + (2)(15) = 505.
Since 505 < 530, we prefer the new model based on this criterion.
Equivalently to using BIC, one could compare models using: Deviance + p \ln[N].
For the base model, Deviance + 2p = 500 + 10 \ln[1 million] = 638.16.
For the new model, Deviance + 2p = 475 + 15 \ln[1 million] = 682.23.
Since 638.16 < 682.23, we prefer the base model based on this criterion.
3.209. (a) 1. Attempting to test the performance of any model on the same set of data on which the model was built will produce overoptimistic results. Using the training data to compare this model to any model built on different data would give our model an unfair advantage.  
2. As we increase the complexity of the model, the fit to the training data will always get better. In contrast, for data the model fitting process has not seen, additional complexity may not improve the performance of a model; as the model gets more complex its performance on the holdout data (test data) will eventually get worse, as shown in the figure in this question.  
(b) Model 2 has the right balance, since it has the smallest test MSE.  
Model 1 is too simple (fewer degrees of freedom than Model 2), while model 3 is too complex (more degrees of freedom than Model 2).  
(c) “Out-of-time validation is especially important when modeling perils driven by common events that affect multiple policyholders at once. An example of this is the wind peril, for which a single storm will cause many incurred losses in the same area. If random sampling is used for the split, losses related to the same event will be present in both sets of data, and so the test set will not be true unseen data, since the model has already seen those events in the training set. This will result in overoptimistic validation results. Choosing a test set that covers different time periods than the training set will minimize such overlap and allow for better measures of how the model will perform on the completely unknown future.”  
Alternately, as in Couret and Venter, one may select either the even or odd years of data as the training set and the other as the holdout set, in order to be neutral with respect to trend and maturity.  
Comment: See Section 4.3 of Generalized Linear Models for Insurance Rating.  
The figure shown is very similar to Figure 7 in Generalized Linear Models for Insurance Rating. We are interested in how the GLM will perform at predicting the response variable on some future set of data rather than on the set of past data with which we are currently working. Our goal in modeling is to find the right balance where we pick up as much of the signal as possible with minimal noise, represented in this case by Model 2.
3.210. A simple quintile plot is a simple quantile plot with 5 buckets.
- Sort the dataset based on the model predicted fraud rate from smallest to largest.
- Group the data into 5 buckets with equal volume. (In this case 2000 claims in each.)
- Within each group, calculate the average predicted fraud rate based on the model, and the average actual fraud rate.
- Plot for each group, the actual fraud rate and the predicted fraud rate.

The saturated model has as many predictors as data points. Thus for the saturated model, the predictions exactly match the observations for each claim. In this case, 1000 of the claims involve fraud, and would all be placed in the last quintile. Thus the last quintile would consist of 1000 claims with fraud and 1000 claims without fraud.

The simple quintile plot:
The null model, has no predictors, only an intercept. Thus for the null model the prediction is the same for every record: the grand mean. In this case, the overall probability of fraud is: $\frac{1,000}{10,000} = 10\%$. Since every risk has the same prediction, one would assign them to buckets at random. Thus all of the actuals by quintile should be close to the grand mean, with small differences due to the randomness of assignments. The simple quintile plot:
“A model that could be used in practice”, would have the actuals increase monotonically, have
good but not perfect predictive accuracy, and a reasonably large vertical distance between the
actuals in the first and last quintiles. A simple quintile plot:

Comment: See Section 7.2.1 and page 59 of “GLMs for Insurance Rating.”
Combines separate ideas in the syllabus reading.
There are many possible examples of the last plot.
Since the records are ordered by predicted values, the records in each bucket change for each
graph. Thus, actuals are not the same for each graph.
Quintile plots are sorted by predicted values from smallest to largest value. Thus the predicted
values must be monotonically increasing (or in the case of the null model equal). Actuals need
not be monotonically increasing, although that is desirable.
In every graph, the average of the actuals should be the grand mean of 10%.
In the final plot, the average of the predicteds should be close to if not equal to 10%; the GLM
may have a small bias.
In the final plot, the predicted and actuals for the final quintile should each be less than the 50%
in the saturated model. In the final plot, the predicted and actuals for the final quintile should
each be more than the 10% in the null model.
### 3.2.11.

<table>
<thead>
<tr>
<th>Claim #</th>
<th>Fraud</th>
<th>Predict.</th>
<th>25% Threshold</th>
<th>Predict.</th>
<th>50% Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y</td>
<td>N</td>
<td>False Neg.</td>
<td>N</td>
<td>False Neg.</td>
</tr>
<tr>
<td>2</td>
<td>N</td>
<td>N</td>
<td>True Neg.</td>
<td>N</td>
<td>True Neg</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>N</td>
<td>True Neg.</td>
<td>N</td>
<td>True Neg</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>Y</td>
<td>False Pos.</td>
<td>Y</td>
<td>False Pos.</td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td>Y</td>
<td>True Pos.</td>
<td>Y</td>
<td>True Pos</td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td>Y</td>
<td>True Pos.</td>
<td>N</td>
<td>False Neg.</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>N</td>
<td>True Neg.</td>
<td>N</td>
<td>True Neg</td>
</tr>
<tr>
<td>8</td>
<td>Y</td>
<td>Y</td>
<td>True Pos.</td>
<td>Y</td>
<td>True Pos</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>Y</td>
<td>False Pos.</td>
<td>Y</td>
<td>False Pos.</td>
</tr>
<tr>
<td>10</td>
<td>N</td>
<td>Y</td>
<td>False Pos.</td>
<td>N</td>
<td>True Neg</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th></th>
<th>25% Threshold</th>
<th></th>
<th>50% Threshold</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted</td>
<td>Actual</td>
<td></td>
<td>Predicted</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fraud</td>
<td>No Fraud</td>
<td>Total</td>
</tr>
<tr>
<td>Fraud</td>
<td>true pos.: 3</td>
<td>false neg.: 1</td>
<td>4</td>
<td>true pos.: 2</td>
</tr>
<tr>
<td>No Fraud</td>
<td>false pos.: 3</td>
<td>true neg.: 3</td>
<td>6</td>
<td>false pos.: 2</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>4</td>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

### Notes:
- **Fraud** indicates whether the claim is fraudulent.
- **Predict.** indicates the prediction made by the model.
- **25% Threshold** and **50% Threshold** refer to the model's decision criteria.
(b) Sensitivity = \frac{\text{True Positives}}{\text{Total Number of Events}} = \frac{\text{Correct Predictions of Fraud}}{\text{Total Number of Fraudulent Claims}}.

Specificity = \frac{\text{True Negatives}}{\text{Total Number of Non-Events}} = \frac{\text{Correct Predictions of No Fraud}}{\text{Total Number of Nonfraudulent Claims}}.

25\% \: \text{threshold: sensitivity} = \frac{3}{4}, \: \text{and specificity} = \frac{3}{6} = \frac{1}{2}. \quad \text{Graph} \: (1 - \frac{1}{2}, \frac{3}{4}).

50\% \: \text{threshold: sensitivity} = \frac{2}{4} = \frac{1}{2}, \: \text{and specificity} = \frac{4}{6} = \frac{2}{3}. \quad \text{Graph} \: (1 - \frac{2}{3}, \frac{1}{2}).

The ROC Curve, plus the 45-degree comparison line:

(c) Using a 25\% threshold results in more predictions of fraud than using a 50\% threshold. Therefore, the 25\% threshold has greater sensitivity, more true positives, which is good; however, this is at the cost of lower specificity, more false positives, which is bad. Alternately, Advantage: You will catch more actual fraud claims because you will have a higher true positive rate. Disadvantage: You will have a higher false positive rate as well, which means you will waste resources to review claims that are not fraudulent.
There are few claims, but they are large. Thus we are very willing to spend money investigating claims for possible fraud; we do not want to miss any true positives and are willing to live with false positives. Therefore, we would prefer the lower threshold of 25%, which has greater sensitivity.

Alternately, a threshold of 0.25 is more appropriate. The high severity makes the cost of not investigating a fraudulent claim very high. The low frequency means that the number of additional claims that will need to be investigated is not very large. The cost of investigating these few additional claims is far less than the cost of potentially missing a few fraudulent claims at a higher discrimination threshold.

Comment: See Table 13 and Figure 22 in “GLMs for Insurance Rating.”

According to the CAS Examiner’s Report, in part (a) one was required to show a table similar to the one I have, showing the origin of the true positives, false positives, true negatives, and false negatives.

3.212. \( X\beta = 2 + (2)(1) + (1)(-1) = 3. \)
The Gamma has the inverse as its canonical link function.
\( 1/(X\beta) = 1/3. \)
The Poisson has the log as its canonical link function.
\( \exp[X\beta] = e^3 = 20.09. \)
The Binomial has the logit as its canonical link function.
\( \exp[X\beta] / \{1 + \exp[X\beta]\} = e^3 / (1 + e^3) = 0.9526. \)
\( 1/3 < 0.9525 < 20.09. \) Thus the correct ordering is: I < III < II.
Comment: The Normal has the identity as its canonical link function.

3.213. The base rate is charged to Male and Territory Q.
Using the log link function, \( \beta_{\text{Terr}} = \ln[545/148] = 1.3036. \)
\( \beta_{\text{Gend}} = \ln[446/148] = 1.11031. \)
\( \ln[4024/148] = \beta_{\text{Terr}} + \beta_{\text{Gend}} + \beta_{\text{Inter}}. \)
\( \Rightarrow 3.3028 = 1.3036 + 1.11031 + \beta_{\text{Inter}} \Rightarrow \beta_{\text{Inter}} = 0.8961. \)
Comment: 148 \( \exp[1.3036 + 1.11031 + 0.8961] = 4024. \)
\( \ln(\mu) = \beta + \beta_{\text{Terr}} X_R + \beta_{\text{Gend}} X_F + \beta_{\text{Inter}} X_R X_F, \)
where \( X_R = 1 \) if territory R and zero otherwise, and \( X_F = 1 \) if Female and zero otherwise.

3.214. \( X\beta = 5 + (-0.65)(5) = 1.75. \) The odds are: \( e^{1.75} = 5.75. \)
Alternately, for the logistic model: \( \hat{\pi} = e^{1.75} / (1 + e^{1.75}) = 0.852. \)
The odds are: \( \hat{\pi} / (1 - \hat{\pi}) = 0.852 / (1 - 0.852) = 5.75. \)
Comment: We have estimated that the probability of renewal is 5.75 times the probability of a nonrenewal.
3.215. (a) i. One can bin driver age into groups. For this example three bins may work well: 18 to 25, 26 to 80, above 80.
ii. One can use hinge functions. A hinge function is of the form: 
\[(X - c)_+ = \max(0, X - c)\].
This will result in a piecewise linear function, with a change in slope at each breakpoint \(c_i\).
For this example, I would choose breakpoints at 25 and 80.
(b) i. With binning: Continuity is not guaranteed. Variation within intervals is ignored.
There may not be enough data in each bin to be credible.
There could be non-intuitive results, such as reversals.
ii. Using hinge functions:
The breakpoints must be selected by the user.
Comment: In both cases, more parameters are added to the model; the principal of parsimony states that we prefer a simpler model with fewer parameters, all else being equal.
3.216. (a) Assuming model A has an intercept, adding driver age using a second order polynomial adds two parameters.

i. Deviance = 2 \{\text{loglikelihood for the saturated model} - \text{loglikelihood for the fitted model}\}.

\[ D_A = (2) \{-1000 - (-1500)\} = 1000. \quad D_B = (2) \{-1000 - (-1465)\} = 930. \]

\[ F = \frac{(D_A - D_B) / (\text{number of added parameters})}{\hat{\phi}_A} = \frac{(1000 - 930) / 2}{1.75} = 20. \]

We compare to the given critical value of 19.5. Since 20 > 19.5, model B is significantly better than model A. Driver age should be included in the rating plan.

ii. Let \( p \) be the number of fitted parameters for Model A.

\[ \text{AIC}_A = (-2)(-1500) + 2p = 3000 + 2p. \]

\[ \text{AIC}_B = (-2)(-1465) + 2(p+2) = 2934 + 2p. \]

Since \( \text{AIC}_B < \text{AIC}_A \), model B is better than model A. Driver age should be included in the rating plan.

iii. Let \( n \) be the number of data points (for each of the models).

\[ \text{BIC}_A = (-2)(-1500) + p \ln(n) = 3000 + p \ln(n). \]

\[ \text{BIC}_B = (-2)(-1465) + (p+2) \ln(n) = 2930 + (p+2) \ln(n). \]

\[ \text{BIC}_A - \text{BIC}_B = 70 - 2 \ln(n). \]

This difference is positive for \( n < e^{35} = 1.586 \times 10^{15} \).

Thus \( \text{BIC}_B < \text{BIC}_A \), and model B is better than model A. Driver age should be included in the rating plan.

(b) When parameters are added to a model, the deviance improves (gets smaller). Thus using deviance alone would lead to overfitting. The issue is whether the deviance gets significantly better. Using AIC or BIC is more appropriate, as they penalize for adding new parameters.

Comment: The degrees of freedom for Model A = number of observations minus number of fitted parameters for model A. The F-statistic has degrees of freedom equal that of Model A and 2. The given critical value is the 5% critical value for 100 and 2 degrees of freedom.

According to the CAS Examiner’s report, for part b common mistakes included: “Giving some of the limitations of deviance such as needing to have the same underlying dataset with the same distribution. This limitation is not restricted to deviance alone.” Besides this limitation on using the deviance, Section 6.1.3 of “Generalized Linear Models for Insurance Rating” also says: “For any comparisons of models that use deviance, in addition to the caveat above, it is also necessary that the assumed distribution and the dispersion parameter (basically, everything other than the coefficients) must be identical as well.” However, the question asked why the deviance statistic alone should not be used to assess model fit. As seen in the solution to part (a), deviance together with the difference in number of parameters and the estimated dispersion parameter can be used to assess model fit.
3.217. (a) As per Section 2.6 of “Generalized Linear Models for Insurance Rating”:
\[ g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \text{offset}. \]

Thus the offset for Policy 1 is: \( \ln\left( \frac{0.013}{1 - 0.013} \right) = -4.330 \).

The offset for Policy 2 is: \( \ln\left( \frac{0.203}{1 - 0.203} \right) = -1.368 \).

The offset for Policy 3 is: \( \ln\left( \frac{0.025}{1 - 0.025} \right) = -3.664 \).

(b) In each case we add the offset to the linear component from the insurance score.

For Policy 1: \( 1.250 + (-0.020)(78) - 4.330 = -4.640 \).

Probability of a claim is: \( \frac{\exp[-4.640]}{1 + \exp[-4.640]} = 0.96\% \).

For Policy 2: \( 1.250 + (-0.020)(92) - 1.368 = -1.958 \).

Probability of a claim is: \( \frac{\exp[-1.958]}{1 + \exp[-1.958]} = 12.37\% \).

For Policy 3: \( 1.250 + (-0.020)(35) - 3.664 = -3.114 \).

Probability of a claim is: \( \frac{\exp[-3.114]}{1 + \exp[-3.114]} = 4.25\% \).

(c) The logit function is: \( \ln\left( \frac{x}{1 - x} \right) \), for \( 0 < x < 1 \).

The logit function has range from \(-\infty\) to \(\infty\).

The logistic function is: \( \frac{e^x}{1 + e^x} \), for \( -\infty < x < \infty \).

The logistic function has range from 0 to 1.

(d) Since the range of the logistic function is 0 to 1, using its inverse the logit as a link function guarantees that the response is in the correct range for probabilities, zero to one.