Casualty Actuarial Society
E-Forum, Winter 2020

The CAS *E-Forum*, Winter 2020

The Winter 2020 edition of the CAS *E-Forum* is a cooperative effort between the CAS *E-Forum* Committee and various CAS committees, task forces, working parties and special interest sections. This *E-Forum* contains four submissions in response to a call for essays on the topic of communications to senior management issued by the Committee on Theory of Risk (COTOR). Also included are two independent research papers.

CAS *E-Forum*, Winter 2020

Table of Contents

For information on submitting a paper to the *E-Forum*, visit http://www.casact.org/pubs/forum/.

Communication of Technical Results to Senior Management: The Art of Storytelling

Jonathan Charak, FCAS, MAAA, CPL

Motivation. This paper was written in response to a 'Call for Papers' on Communication of Technical Results to Senior Management
Method. This essay relies on personal experience which has worked for me
Conclusions. Structured and brief communications are key to communicate with senior leadership as time may be limited

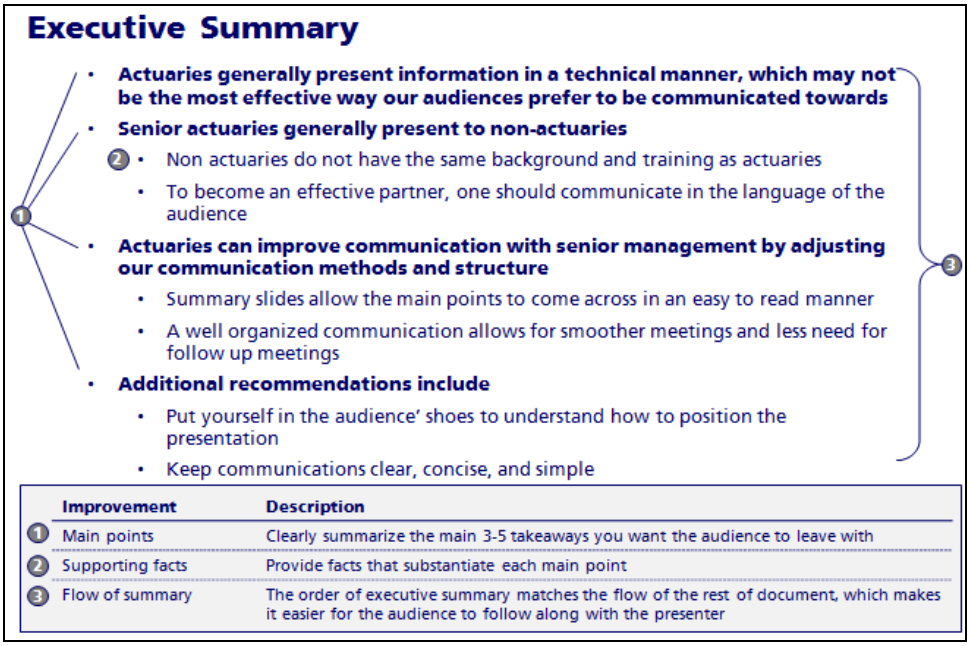
Keywords. Communication, Structured Thinking

As we progress in our careers as actuaries, our first challenges are exams. After this accomplishment, we may move into mentoring other actuaries, training them on actuarial principals, and managing a team of actuaries is a potential. Eventually, some actuaries will find themselves in front of the market-facing leaders and senior management of their company. Senior management may have a different background than actuaries and haven't spent years agonizing over ELFs/ILFs, tail factors, GLMs, and other 'technical' details that actuaries thrive in. Sharing actuarial insights is crucial for an insurance company's success. Effective communication should impart knowledge, nudge/influence decisions, and assist senior management come to the conclusion that betters the company. In my opinion, the ability to do this differentiates a good actuary from a great actuary.

Communication to senior management can come in many forms. While the details of how a slide deck, an email, or a document differs, they generally follow a similar format. There should be an **Executive Summary**, **Context/Background**, **Analysis of Finding**, **Recommendations**, and **Next Steps**. A slide deck provides the best format for communication with senior management. It allows one to use visuals and bullets to create an effective communication both in person and when senior management wants to view the information on their phone or tablet.

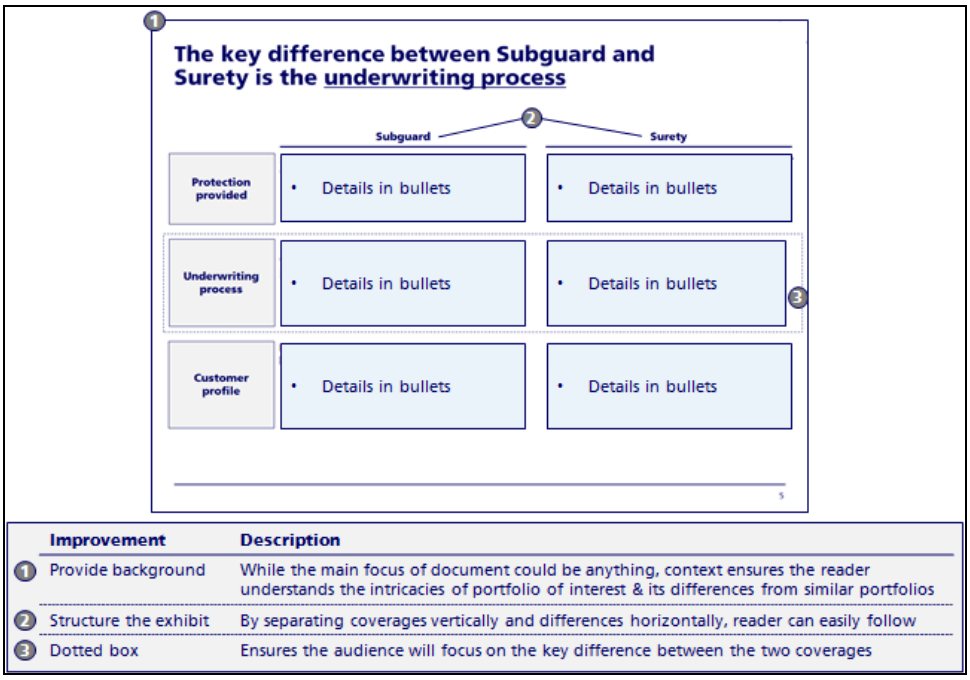
Structure	Description
A **Executive summary**	• First slide of a presentation • Use to clearly and concisely summarize the main takeaways for the audience • Order of main points and bullets should match the flow of the rest of the document
B **Context / background**	• Ensure the audience knows what you are talking about and why you are discussing it • Example items to include: The problem you are trying to solve, the facts that demonstrated there was an issue and the objectives of the analysis
C **Analysis and findings**	• Explain the key findings from the analysis • Support each finding with facts
D **Recommendations**	• Deliver your recommendation (and rationale)
E **Next steps**	• Clearly articulate next steps, including due dates, owners and dependencies

An executive summary should be able to summarize the entire communication into something senior management can read in a minute or two and understand. This could be a couple of paragraphs in a document, a three to five bullet email, or one slide in a slide deck (e.g. PowerPoint). The executive summary should cover three points: **Situation**, **Context**, and **Resolution**. The rationale to start with the situation is simple, senior management is busy. They may or may not recall the history, why this is pertinent for them (and the company), and why they should dedicate time from their busy calendar. Context provides an outline to why this is a problem and how the company is currently dealing with the said issue. The resolution is the action that the communication is putting forth. An email may only contain an executive summary with the details attached as a slide deck or a document. While the executive summary is the first part of a communication and guides the audience, you may choose to write this last.



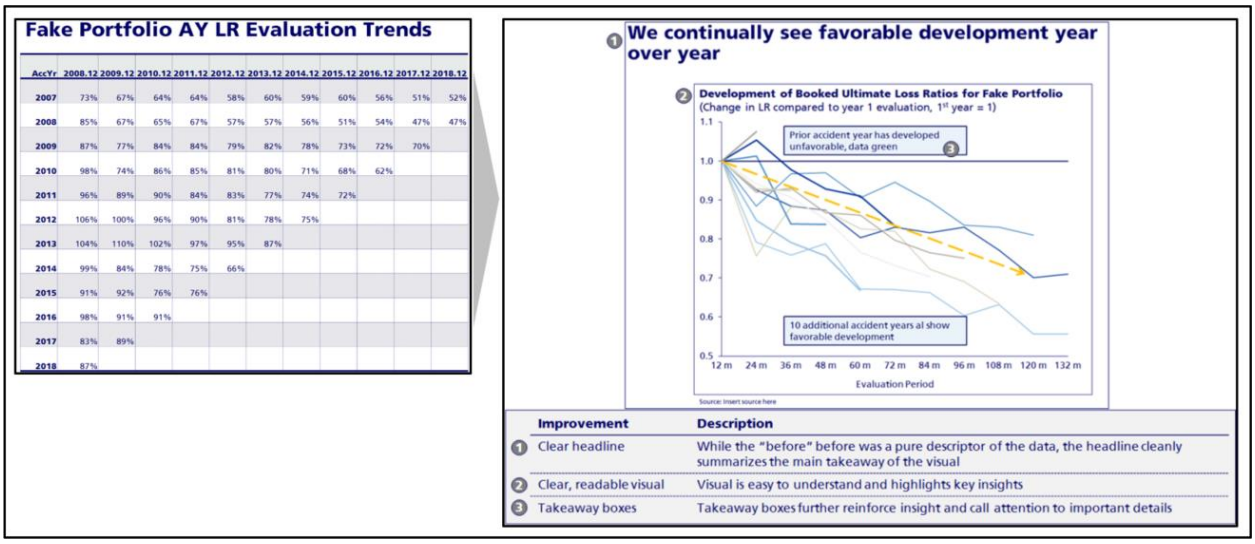
After the executive summary, the following sections include additional details. To ensure engagement of senior management, there should be a mix of visuals and well-organized bullets. Formatting content with call-out boxes, flow-charts, chevrons, and so forth will make a 'wall of text' much easier to comprehend. Further, when one creates visuals/graphs, be sure they are purpose-built and not merely screenshots of already existing visuals; making effective communication to senior management requires additional care. Keep principals of data visualization in mind, such as clean graphics are better than overly complex ones. Creating organized text and bespoke graphs will direct management's attention and allow you to drive the conversation. Finally, the lead on a slide should be an active lead. It assists in telling the story and guides the recipient of the communication.

The example below takes lots of information and structures it in a clear manner so senior management can easily gain context, even if they're not familiar to the details of the subject in question. The active lead provides clear details of the takeaway message.



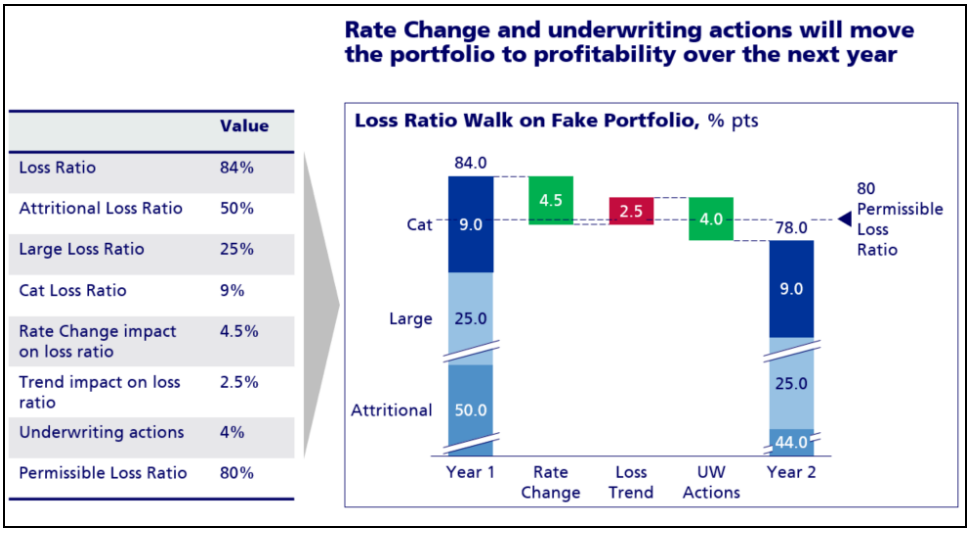
	Improvement	Description
1	Provide background	While the main focus of document could be anything, context ensures the reader understands the intricacies of portfolio of interest & its differences from similar portfolios
2	Structure the exhibit	By separating coverages vertically and differences horizontally, reader can easily follow
3	Dotted box	Ensures the audience will focus on the key difference between the two coverages

Senior leadership now has a base level of knowledge and is ready to proceed to the analysis and findings. Generally, senior management's interests are in the conclusions and the rationale behind the conclusion rather than all the details of the analysis. When working with senior management, an actuary may need to shift how they think about data and try to create visualizations to depict a finding. Instead of a chart, visualizations could easily be cleaner and more descriptive. Below are a couple of examples. The first one starts with a loss ratio triangle. A triangle is full of useful details, however more than senior management needs. Also, the lead does not provide any explanation. By creating a bespoke graphic, the conversation will naturally move toward the consistent favorable development in each accident year's loss ratios, which the lead corroborates.



	Improvement	Description
1	Clear headline	While the "before" before was a pure descriptor of the data, the headline cleanly summarizes the main takeaway of the visual
2	Clear, readable visual	Visual is easy to understand and highlights key insights
3	Takeaway boxes	Takeaway boxes further reinforce insight and call attention to important details

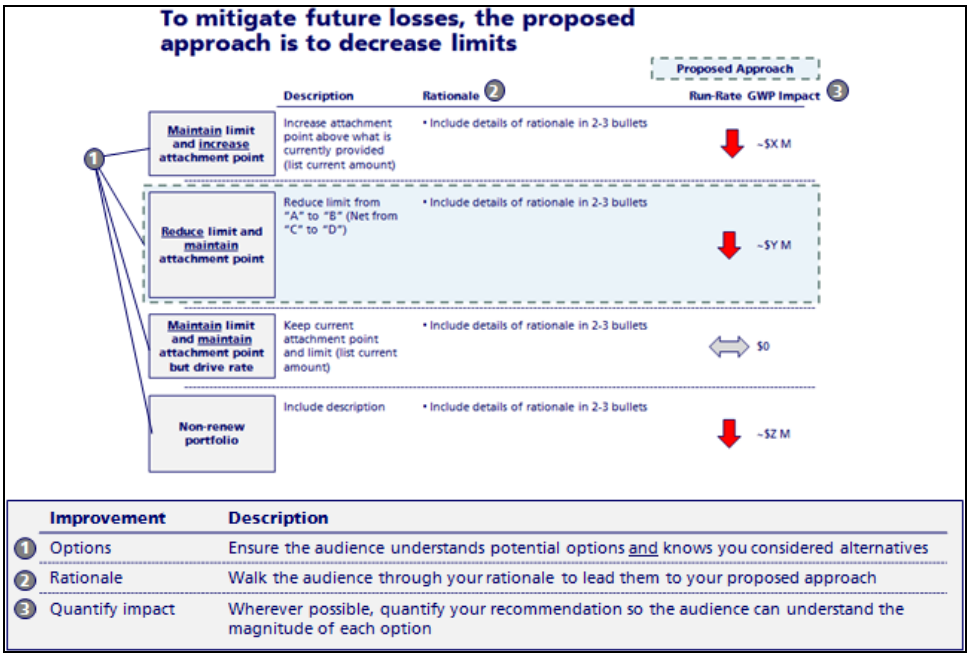
In another example, the presentation of key financial metrics is in a chart. By creating a clear visualization and a descriptive lead, one can direct senior management to see the changes proposed in the portfolio. Follow up slides, perhaps in the appendix, will explain rate achievement and other assumptions built into this projection.



Now that senior management understands what your analysis concludes the logical next question is: "What do we do next?" The next slide or two outlines a set of recommendations. Senior leadership wants to know that their actuaries reviewed all reasonable scenarios before they propose a recommendation. As such, communication to senior management should demonstrate that you (and your team) analyzed multiple options to make your proposal. All scenarios tested should include a description, have a rationale for inclusion to explore, and provide a projected impact on financial metrics.

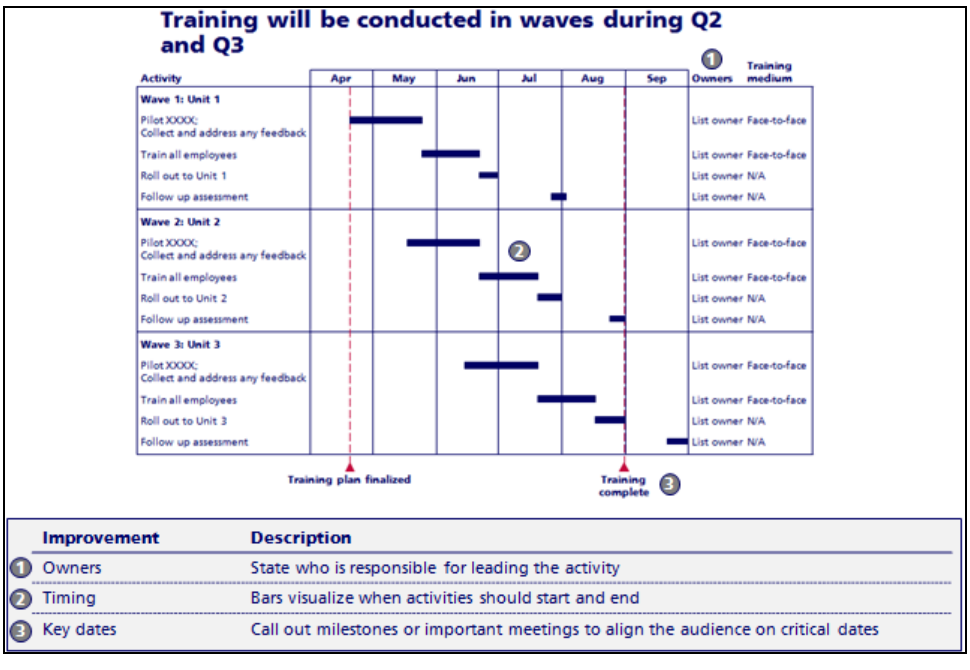
Using the mnemonic **SMART** when creating your recommendations will ensure proposed actions meet best practices. SMART was first coined by George Doran in a 1981 issue of *Management Review* and it remains a useful tool. Recommended actions must be **specific** in activities and who is participating, without any ambiguity. There must be a way to **measure** the expected impact (e.g. GWP, NWP, loss ratio, operating profit…). The action must be **achievable** within the resource and time constraints. Recommendations must be **relevant** to solve the problem. And the actions must be **time bound** with a timeline on completion of the objective.

The example below shows four recommendations on a portfolio. Different scenarios test potential actions on attachment points and limits; each scenario has a description, rationale, and projected impact. This communication shows all scenarios and identifies the proposed go-forward approach with a box to highlight. Details on the four scenarios are not in the main deck but in the appendix.



	Improvement	Description
1	Options	Ensure the audience understands potential options and knows you considered alternatives
2	Rationale	Walk the audience through your rationale to lead them to your proposed approach
3	Quantify impact	Wherever possible, quantify your recommendation so the audience can understand the magnitude of each option

Next, the communication discusses operationalizing the recommendation. The 'next steps' section will walk senior management through the actions needed to operationalize the proposed recommendation. Instead of a series of bullets, a Gantt chart, per the example below, can show the tactical steps needed to execute. Whatever method of communicating the next steps, one should always include the activities, owners, and timelines.



	Improvement	Description
1	Owners	State who is responsible for leading the activity
2	Timing	Bars visualize when activities should start and end
3	Key dates	Call out milestones or important meetings to align the audience on critical dates

Finally, one may need an appendix and other supporting files (such as a data visualization tool like PowerBI, Tableau, or R Shiny). The appendix houses all the additional details, including some of

the more technical actuarial work; generally, it includes work necessary to create the recommendations to senior management. Details included here most likely isn't what senior management will dig into, though you should prepare to talk about the details if need be. After all, this analytical work is why senior management hires actuaries. They value actuaries who can translate the analysis and distill easily digestible actionable intelligence.

The visual below includes a quick guide of best practices when structuring your communications to senior management. Following the structure described above and the best practices below will aid you in creating effective communication.

Structure	Best Practice
0 **Pre-Work**	• Identify the "key questions" you will need to answer • Create a shell on a piece of paper, focusing on the storyline before doing design work
A **Executive summary**	• This is the "embodiment" of your narrative arc • Try not to go any smaller than 16 point font (14 if absolutely necessary) • Use **bolding** to identify key points (your bolded sentences should read clearly)
B **Context / background**	• Try not to clump text together – use spacing and divider to break up the flow • Use callout boxes to hone in on key messages
C **Analysis and findings**	• Visuals should be clear, well labeled, and sourced • Limit the insights/takeaway messages per slide, too many makes it difficult to focus
D **Recommendations**	• Recommendations should always be accompanied by a **rationale** and an **impact within a certain timeframe** – in other words they should be SMART actions – Specific, Measurable, Achievable, Relevant, and Time bound
E **Next steps**	• A next steps page should always be included, even if it is only a bullet-point list

Executive communication syndicates information and helps senior management to make strategic decisions. Structuring a story as previously described, one shares the correct level of detail with a healthy mix of visuals. The advice above will refine your communication to leadership. Content is key and good structure alone won't make effective communication. Finally, if a communication is effective in one hundred words, don't say it in five hundred as brevity will allow you to retain the attention of senior management (or any business professional).

Biography of the Author

Jonathan is VP and Emerging Solutions Director at Zurich North America and Innovation lead for Technical Underwriting. His responsibilities include identifying emerging risks, evaluating potential solutions, and working with cross-functional teams to bring new products to market. He has held multiple roles of increasing responsibility across a variety of actuarial and non-actuarial functions across both the USA and Australia.

Jonathan volunteers with the CAS as the vice-chair of the Automated Vehicle Task Force, CAS Media spokesperson, Learning Enhancement Process Mentor, and additional engagements; from this he has presented at multiple industry events both on a national and international platform.

Jonathan holds a Bachelor of Science in Mathematics and a Bachelor of Science in Biology from Illinois Wesleyan University, is a Fellow of the Casualty Actuarial Society (FCAS), Member of the American Academy of Actuaries (MAAA), and a Certified Program Leader (CPL).

Communicating in Crisis Situations

Rick Gorvett, FCAS, CERA, MAAA, FRM, ARM, Ph.D.
Chris Morse, Ph.D.
Julie Volkman, Ph.D.

Abstract: Communicating technical information, especially in a crisis situation and particularly when the audience does not share the technical background, is a challenge that actuaries frequently face. This essay describes the dynamics and issues involved in crisis communications and provides some recommendations for actuaries confronting such a situation.

Keywords. Communication, crisis management, senior management

Actuaries, like practitioners in any profession that involves significant quantitative or technical expertise, have a reputation for sometimes being substandard communicators. To the extent this is true, it is probably less a matter of lacking basic communication skills, than it is the inherent difficulty in communicating technical material to audiences that generally do not share that background. Communicating in such an asymmetric environment presents a natural challenge. When, on top of this, an actuary is attempting to communicate bad news or a potential crisis situation, the task of communicating effectively is doubly difficult. We hope this essay will help actuaries to better understand the dynamics and issues involved in crisis communications.

Potentially, actuaries may confront at least two types of crises. To the extent that actuaries are executives and leaders in organizations, they may well have responsibilities in a high-profile crisis situation such as a cyberattack or an incident that somehow threatens the company's reputation. More often, though, actuaries need to operate and communicate in crises of a more subtle, actuarial nature. Many actuaries have had to deliver bad news or present and educate company executives regarding threatening situations. Just a few of many possible examples include:

- Results of an actuarial analysis indicate that the organization is insolvent, or that its financial condition is worse than had been anticipated.
- An emerging or ongoing natural catastrophe, unhedged financial risk, or other event is about to play havoc with the company's finances, operations, capital adequacy, liquidity, etc.

- A new type of risk has emerged, and the evolving litigatory environment surrounding that risk suggests that the organization will very soon experience significant losses that were previously unanticipated and were not contemplated in the ratemaking process.

Effectively communicating in a highly technical and quantitative environment, with an asymmetry between the communicating partners regarding an understanding of and familiarity with the analytics underlying the findings, is difficult enough. Where a particularly significant or crisis-level indication is concerned, all the difficulties involved in communicating in a crisis are also piled on. Indeed, post-mortem analysis of crisis situations often reveals that communications could have been handled better. While there could be several reasons for poor communications in crisis situations, we argue that a majority of miscommunication can be attributed to two main causes. First, audiences in a crisis behave differently than they do normally, so adjustments must be made [3]. Second, we as crisis communicators often overestimate our delivery ability, which can further cause issues. In this essay, we attempt to highlight some of the major factors within these two areas, as well as offer some advice for actuaries and other crisis communicators to overcome them.

The nature of a crisis impacts individuals' abilities to process information, requiring them to alter the ways that they cognitively operate in such a situation compared to their normal approach. In these cases, crisis communicators who do not alter their messages will often encounter problems, or at the very least fail to convey the importance of their information in a way that the audience understands. The result can be a failure to take the crisis seriously, a lack of motivation to act on the information, or an under-impression of the potential impact of the crisis on the company or organization.

In terms of audience behavior, crisis communicators must be aware of three key issues. First, in a crisis, individuals tend to find themselves in situations of high stress and are often being presented with large amounts of information in a short period of time. In cases such as this, research has suggested that individuals have trouble with message retention, oversimplify the message content often missing key pieces, and misinterpret goals articulated by the crisis communicator [4], [8]. Second, a crisis represents a situation in which uncertainty is created as an individual's understanding of the world is challenged or that person's ability to predict what is going to happen next is compromised. In cases such as this, individuals often find themselves clinging to "what they know is true." This means that people will often default to long-held beliefs about the world and how it works, or "tried and true" ways of handling things instead of alternative plans or

ideas [2]. Audience members will often reject "new" information in favor of what they have normally encountered. In cases of crisis, this would suggest that crisis communicators who present novel information or ideas, might be ignored by their audience in favor of "what has normally happened," or what has occurred in the past. Third, feelings of uncertainty will often result in negative emotional states such as anxiety, fear, and anger [1]. Emotional states such as these have been argued to create "action tendencies" or behavioral responses in individuals, that if left unaccounted for may present additional problems with a crisis communicators message. Fear, for example, has been linked with a tendency for "flight" responses while anger has often presented an "attack" response [7]. In the context of crisis communication, this could translate into a tendency for the audience to avoid a crisis message, either by ignoring it or discounting it, or they could challenge the message, questioning its validity. In either case, heightened emotional states can cause failure in the crisis communicator achieving her/his goal by having the audience be less receptive than anticipated.

While the impact that a crisis has on an audience is problematic, so too is the way in which crisis consultants convey the information. In many cases, people who are tasked with conveying information make assumptions about both their message as well as who they are talking to, which often causes confusion or reduced understanding. Unfortunately, in the case of a crisis, these assumptions can have severely negative impacts. One particularly problematic issue – particularly for actuaries – is a communicator's use of jargon. Oftentimes in work specializations, individuals develop and use terms that are not common vocabulary to those outside those specializations. Unfortunately, given the often-siloed nature of the workforce, and individuals being in constant contact with others who also speak with a similar vocabulary, people can often forget that these terms are not commonplace, or at the least make erroneous assumptions that "everyone else gets what I mean." In fact, the use of jargon impedes one's ability to effectively communicate with non-experts [5]. When conveying information to others, especially in high stress situations, individuals thus can overestimate the "simplicity" with which they are speaking. In cases where the audience is already experiencing the issues mentioned above, this can result in a speaker believing that a successful message was completed, while the audience member becomes lost or ignores what is being said.

There is an additional issue that should be of concern to crisis communicators. Literature involving primacy effects suggests that the first piece of information that people are presented with will be used to interpret and compare all future information [6]. Therefore, the first message that an

individual is presented with in a crisis tends to carry the most weight. This significantly increases the importance of presenting not only correct information to an audience but of making sure that it incorporates the issues stated above. If the message is designed without consideration of these issues, then not only can the decision making of individuals be compromised, but how people view the crisis will also be very hard to change from their initial erroneous impression.

Taken together, the above comments suggest that, when dealing with a crisis, the actuaries and other individuals doing the communicating cannot approach the task as simply "conveying information." By its very nature, a crisis impacts an audience, altering the way that they process and interpret information. Furthermore, some of the tendencies that speakers have, which might be normally overcome in everyday conversation, can have negative impacts when exhibited in a crisis conversation. It is important for individuals to remember that they must be simplistic and repetitive in the conveying of their information. They must be prepared to deal with audiences wanting to avoid what they are saying or challenging it. While the speaker may feel that they are speaking "plainly" they must examine their use of jargon and appreciate the experience level of those they are speaking to. Finally, while a speaker may believe that the solution being presented is logical and practical, he/she must understand that if the proposed solution deviates too much from the established norm, the audience may reject it as their uncertainty causes them to fall back on what has been done before – or, at the very least, what is comfortable and safe.

REFERENCES

[1] Afifi, W.A., and C.R. Morse, "Explaining the Role of Emotion in the Theory of Motivated Information Management," In T. D. Afifi and W. A. Afifi (Eds.), *Uncertainty, Information Management and Disclosure Decisions: Theories and Applications*, New York: Routledge, **2009**

[2] Andreason, A.R., *Marketing Social Change: Changing Behavior to Promote Health, Social Development, and the Environment*, San Francisco: Jossey-Bass Publishers, **1995**

[3] Centers for Disease Control and Prevention (CDC), *Crisis and Emergency Risk Communication (CERC) Manual*, **2012**

[4] Darke, S., "Effects of Anxiety on Inferential Reasoning Task Performance," *Journal of Personality and Social Psychology* **1988**, Vol. 55(3), 499-505

[5] de Bruin, W.B., and A. Bostrom, "Assessing What to Address in Science Communication," *Proceedings of the National Academy of Sciences* **2013**, Vol. 20 (Supp 3), 14062-8

[6] Haugtvedt, C.P., and D.T. Wegener, "Message Order Effects in Persuasion: An Attitude Strength Perspective," *Journal of Consumer Research* **1994**, Vol. 21(1), 205-218

[7] Lazarus, R.S., *Emotion and Adaptation*, New York: Oxford University Press, **1991**

[8] Sengupta, J., and G.V. Johar, "Contingent Effects of Anxiety on Message Elaboration and Persuasion," *Personality and Social Psychology Bulletin* **2001**, Vol. 27(2), 139-150

Biographies of the Authors

All of the authors are on the faculty of Bryant University in Smithfield, RI.

Rick Gorvett is Professor and Chair of the Mathematics Department. He has a PhD from the University of Illinois at Urbana Champaign and is a Fellow of the Casualty Actuarial Society.

Chris Morse is a professor in the Communication Department. He has a PhD from the Pennsylvania State University.

Julie Volkman is a professor in the Communication Department. She has a PhD from the Pennsylvania State University.

Setting the Scene for Communicating Technical Results to Senior Management

Christopher Smerald, FCAS, FIA, MAAA

Abstract: In this essay, we look at enablers to effective technical communication with senior management. Good planning and concise writing is essential, but in this essay, we argue that both analysist and recipient also need to work collaboratively towards ensuring the analysis is tuned to the recipient's needs. This is especially true, because actuarial method is often very different from management decision-making approaches. The actuary and, ideally, also management need to go the extra mile to ensure they understand the other's language and work context. To help with this, simple rules of thumb (heuristics) are suggested as part of a good communication process.

Keywords. communication, reports, culture, senior management, personal leadership

1. INTRODUCTION

Just imagine going to a play where the production spent most of its time writing the script and only spent a little time, at the end, thinking about how they might connect with their audience, making sure their set works, and preparing to speak their lines. It might just work with a simple play (or if the audience is another scriptwriter who can fill in the abstract gaps with their imagination), but to most it would seem incomplete or worse.

Transcribing the simple play for insurance, imagine two short actuarial studies required for "Andy", the CFO. One done by "Lucy" who worked alone to the last-minute preparing exhibits but did not plan what to say to Andy. "Ken" did the other. However, he had a cup of coffee with Andy to confirm what was needed before creating the exhibits, and he left himself time to prep for Andy. Who was invited back for an encore project?

Perhaps Andy also backs plays and he is funding a professional show. He knows public fashion demands lots of audience participation, thicker subplots, elaborate sets, and no mistakes. Things are much more complicated. Without a good process, much is at risk. Andy must understand how it all works and be more involved. His producer needs to understand Andy's and public needs better. Independent work with only a few short meetings is no longer adequate.

Now imagine a more challenging insurance situation where CFO Andy and management are under pressure for deeper / more agile business insights, improved risk management and governance, and "ownership" of the numbers. A nice focused table and simple clear words

may no longer be enough. The actuarial analysis may demand more thoroughness, transparency and efficiency using new methods with more data (or more pressure on old methods) and more controls, plus enhanced disclosures around selections, uncertainties and drivers.

For this more complicated actuarial work, a highly collaborative and participatory process is needed. We have broken the important aspects of this into four elements through the acronym CUPS: Culture, Understanding, Practice, Suggestions.

1. **Culture** relates to the principles and customs underlying the relationship. In this case a willingness to make things work and being collaborative in the relationship by listening well and allowing time for informal communication as well as formal.
2. **Understanding** is about knowing context and goals. This includes other participant's: language and values, working and thinking process, and priorities and pressures.
3. **Practical** relates to things which can be done to simplify communication by a rule of thumb toolkit (heuristics) once culture and understanding are established.
4. **Suggestions** are just that. The more complicated things are, the harder it is to manage or improve alone. Adjustments are made based on feedback that is specifically requested.

These ideas will be discussed separately in more depth below, followed by a few end comments.

Culture

According to The Barrett Values Centre, who help build values-driven organizations, "The culture of a group of people is a reflection of the values and beliefs … that are embedded in the structures, policies, systems, procedures and incentives of the group"[1]

The sort of culture we are seeking includes a strong personal leadership element and is founded on positive business and personal values. -Where each is committed to making the relationship work to the best advantage of all concerned. This includes willingness and skills

to work across boundaries, with curiosity and being open to challenge.

The needed culture and underlying values are likely already there and may need only a little reflection to be lived more authentically and effectively, so that the forms followed in engagement are aligned more closely with their function. The actual form will likely vary considerably among organizations, so this section focuses more on the values which underlay culture which can be universal. By thinking about values and how they are lived through culture, we can connect and communicate better with others.

Here are seven good values examples from 6Q Blogger Heryati R[2]

1. Stewardship,
6. Integrity,
10. Diversity (the source gives a fuller description),
16. Quality
20. Good Citizenship,
41. Leadership: The courage to shape a better future,
87. Togetherness and enthusiasm.

Culture development starts with thinking about actions which would support these values. For example, making time for informal conversations, which of course takes time, but may increase efficiency in the longer run. This is because informal conversations can carry wider bandwidth of meaning as opposed to emails, agenda packed meetings and video conferences, where messages are more compressed.

This personal connection aspect is also echoed in a list by Miranda Anderson[3] which suggests additional procedures:

- Create a shared ritual like a cup of coffee informally
- Agree to your commitments early and often and help facilitate commitments of other key stakeholders
- Be There When It's Hardest. Pick up the phone (or text, if necessary) the minute

there's a whiff of something awry, and then to do whatever it takes to make the situation right.

Understanding

Communication between actuaries and senior management is complex. Each focus on different aspects of the business, has different goals and past experiences, and may internally process things quite differently in language terms or units of thought. This may cause them to assign different meanings to the same underlying information[4]. So, understanding all this context, especially when the messages and uses are complex is especially important.

It helps to consider how actuaries solve business problems using actuarial method. This can be more of an iterative art than a science, especially if data is missing and simplifications or extrapolations are needed. Tools may include any of the following: logic, statistics, heuristics (rule of thumb methods and models), and professional judgement. The iterations and uncertainties can leave an actuary feeling they have not really completed the analysis. So, the actuary may be tempted to explain too much their steps and unresolved issues, and not why the selections make sense and what the key issues are.

In contrast to this deductive work, senior management might be reflecting more on similarities and differences in opinions from diverse experts while deciding on a course of action. The more objective the opinions and the more they use a common language, the easier it may be to decide. So, if the actuarial information is too abstract or tentative, they may not be able to synthesize it with more objectively framed opinions from sources like ERM, finance, investments, underwriting, etc. Thus, actuarial information is not always something which stands alone. It may be used as part of a larger process, so actuaries need to work to make it be more objective and comparable with other business information.

Working collaboratively and being able to see both sides of the of the situation is particularly relevant here. The actuary needs to discover the manager's objective and decision context. The manager needs to understand actuarial method and actual workings better, since not all of them can be translated efficiently into normal management language. This may take time before it becomes natural to both, but it is worth it, and it does take two.

Clear lines of responsibility and accountability are also important to the process of understanding. An actuary does not just produce "the answer" and a manager does not just make decisions. Each are responsible for their share of ensuring good risk management and

for contributing to governance and social protection[5]. These make communication more complicated. By recognizing these parallel and complementary roles and breaking up communication along these lines (of decisions, risk management and governance), messages can be simplified. This splitting out may also help find an optimal level of disclosures on uncertainties, controls and caveats, because they have been untangled from other goals.

Finally, listening is a strong part of this understanding. To be a good listener, you need to set aside your own reactions, ignore sparked tangential thoughts, and take good notes generally. In order to pick up nuances (especially where a conversation is on unfamiliar ground), it helps to research and plan what is likely to be said by you or others. This preparation sets expectations, so surprises are captured well. This is a verbal version of tracking actual vs. expected. Allow silences to happen. Silences → reflection → understanding. Reflective listening is also good. Say what you thought you heard or what you understand they want. This builds trust that you are listening and shakes out misunderstandings.

Practice

Practice is best approached in a principles-based way with ideas to try to fit the situation. The below framework is based on work of The Good Actuarial Report Working Party which the author has been leading.

The framework centers around truly understanding user needs and includes five parts[6]:

1. Prework. Communication may fail if user needs are not properly understood from the outset. This is partly covered in the preceding section, but planning time is needed:
    a. Really understanding managements goals and expected uses.
    b. Selecting / planning proportionality and priorities, and
    c. Planning the scope of the work to be performed.

The proportionality heuristic[7] is beyond the scope of this article, but just as actuaries have methods to simplify complexity in problem solving, analogous simplification can help with communication and work planning.

2. During Analysis. Complete the work focusing on what is important, having kept

notes on what was opaque and what was clear. A good practice is to rank your findings and interim assumptions by your level of belief. Was the fact pattern clear enough that you have a firm recommendation? Is it more of a best guess, or was it speculative where the model said "X", but you cannot validate it?

3. Communication planning: Keeping user needs in mind, plan what is most important to communicate in advance.
4. Writing/Communicating: Ensure it is relevant to user needs, highlighting what is important. Be Concise, less is more.
    a. Instead of: "I took this data, applied these methods, and got these results", you could try: "Your business needs fixing / is doing great, as these results show, and this is how you can see for yourself."
    b. Write for Flow, by writing with flow:
        i. The flow for the reader who discovers what is important through following clear logic.
        ii. Flow as a writing technique where you get a sprint of content down and before overwriting the first sprint, write the next part, then the next... Then, with first draft quickly finished, you can overwrite and refine, reorder, fill in gaps, reduce, etc. Don't start the iterations of improvement too early as you may burn too much time.
    c. Avoid Jargon – Use your authentic voice instead, avoid acronyms and technical terms
5. Feedback. See the "Suggestions" section below.

Suggestions (and Feedback)

Suggestions and feedback are important for complex situations, because without them, it is difficult to judge how to improve. The actuary needs to know what new thing worked, what did not, and based on management's experience what they might try next. Management needs to know if their actions are outside of the actuary's comfort zone, and what they might need to do to understand things better. It is easy to see how they are part of a good culture. Without listening and co-ownership of success it may not happen or be constructive.

Conclusions

The idea for this essay came from attending a workshop where non-executive directors and chief actuaries discussed successes and challenges in formal actuarial communication for UK actuarial function reports. I was struck by the lengths to which either the actuary or NED went to understand the other's language, and by the importance they placed on good lines of informal communication. -So that the actuary would not be socially constrained if issues were to arise later. These cultural aspects helped cement all the communication research I have been involved in. I encourage readers to look for their own examples of good practice and to conscientiously copy them, as I have done, wherever it makes sense.

Thank you for reading this essay and I look forward to your suggestions.

REFERENCES

[1] Barrett Values Centre, "What is Culture", https://www.valuescentre.com/mapping-values/culture
[2] 190 Brilliant Examples of Company Values, https://inside.6q.io/190-examples-of-company-values/, accessed on 2-Feb-19
[3] pulled from Miranda Anderson's, https://www.fastcompany.com/40548797/how-to-build-long-term-business-relationships-one-coffee-at-a-time
[4] This is a loose reference to Text and Conversation Theory https://en.wikipedia.org/wiki/Text_and_conversation_theory accessed 30-Dec-2018. and Taylor, J.R., Cooren, F., Giroux, N., & Robichaud, D. (1996). The communicational basis of organization: Between the conversation and the text. Communication Theory, 6, 1-39
[5] Christopher Smerald, The Good Actuarial Report Working Party Looks at Meeting Standards, Unpublished paper.
[6] Christopher Smerald, Matt Byrne and the Good Actuarial Report Working Party, "A Holistic Process for Producing Good Actuarial Reports", Institute and Faculty of Actuaries, GIRO presentation October 2018.
[7] Ibid note 5.

Biography of the Author

Chris Smerald is an actuary with over 34 years' experience in corporate environments. He has held several volunteer research positions with the CAS and the UK's Institute and Faculty of Actuaries (IFoA) in the areas of risk, reserving, communication and actuarial philosophy. He has experience of reserving, pricing, risk and analytics for most lines of business in the US and internationally. -Particularly within Excess and Surplus Lines. He has a degree in Mathematics from the University of Delaware, is a Fellow of the CAS and the IFoA, and is a Member of the American Academy of Actuaries.

How to Present Technical Results to Managers without Either Side Feeling Stupid

Jim Weiss, FCAS, MAAA, CSPA, CPCU

Abstract. The following essay is a response to the CAS Theory of Risk Committee call for essays on the topic of Communications to Senior Management. The essay argues some of the prevailing thinking regarding interactions between managers and technicians may reinforce counterproductive tendencies and that a more critical but rarely discussed challenge is both parties' fear of looking stupid. The essay offers practical suggestions to acknowledge and overcome this fear both short and long term.

Keywords. Communications; Management; Fear of Looking Stupid (FOLS).

1. INTRODUCTION

Many discussions between technicians and managers go less than ideally. Some of the structural elements contributing to this misery are self-evident or amply explored in literature. For example, management's congested schedules make it impractical to engage them in nuances required to understand pros and cons of different techniques and approaches; it is sometimes difficult to abstract how mathematical results translate into actions with real world impact; and each cohort possesses different skills, experiences, and peer groups and is not used to interacting with the other. All these factors are straightforward enough that if any represented the true problem, then the Casualty Actuarial Society Theory of Risk Committee would not sponsor an essay contest on communicating results to senior management – and I would not submit an entry arguing the real issue both managers and technicians must address is their mutual fear of looking stupid (FOLS).[1] Once each party understands and plans for its own and the other's FOLS, they can all begin to experience more fruitful, less stressful interactions.

2. FOLS … BY CHOOSING THE WRONG ANSWER

An inaccurate subtext to studies like the present one is that there exists some sort of fundamental difference between managers and technicians, when in fact technicians can and often do become highly effective leaders in their organizations. There is arguably much

[1] Possible origin of term FOLS is Torrence (2017).

more that (horizontally) differentiates the frame of reference of, say, a medical or legal professional from that of an insurance professional of any kind, than there is that (vertically) differentiates an actuary's or data scientist's perspective from that of a chief underwriting or chief financial officer at the same insurance company. The latter differences in outlook tend to relate more to individual motivations and incentives rather than knowledge or experiences.

Individuals do not (usually) consciously prioritize individual needs over those of their organizations, but biases come into play at a subconscious level. The Peter Principle argues that individuals receive promotions until their successes turn to failures.[2] Having a success story to one's name involves taking chances, because it is relatively rare to experience pure and unearned good fortune. However, once an organization rewards successful risk taking with a management opportunity, the individual's incentives change. Salary and accountability increase, and advancement opportunities become more elusive. Reputation sometimes becomes as powerful an asset as skill or ability. There is greater individual financial freedom to be patient for the perfect opportunity, and greater adverse consequences for unsuccessful risk taking. Meanwhile, those whose initial risk taking does not pay off have less to lose from further risk taking.

The circumstances in which technicians and managers typically find themselves interacting exacerbates this subconscious conflict of interest. Technicians' presence at the table suggests that problems at hand are insufficiently addressable or understandable by more qualitative, instinctive, or fundamental approaches, and that heavier artillery such as math is required. Managers may prefer lighter artillery. This is exactly where overplayed advice for technicians to "lighten up" their message misplaces focus. Digestion is prudent, but it does not change the essential nature of most technical recommendations – which is to exit the comfort zone. The best chance at breaking through to a manager on this front is by illustrating that risks of inaction exceed the risks of potential actions implied by the analysis.

To illustrate, consider an insurance company whose goal is to break even. Their actuary's analysis suggests expected expenses exceed expected revenues by 25% for the upcoming year. The chief underwriting officer receiving the analysis is likely less concerned with how efficaciously the actuary derived the 25% than with risking his or her own reputation among

[2] Wagner (2018) reviews recent academic research surrounding the reality of the Peter Principle.

policyholders, producers, and regulators with intervention.[3] Providing a defense of the analysis casts the conversation as a technical referendum rather than a comparison between one approach implied by the analysis and another of doing nothing.[4] The actuary can avoid this trap by volunteering probabilities of breaking even under either alternative – say, 60% with the recommendations and 20% otherwise. In this way, the actuary assumes the burden of defending not only the recommendations but also the CUO's default position.[5] This, in turn, aligns the actuary's narrative with the CUO's FOLS, by objectively presenting inertia as a very risky alternative.

3. FOLS … BY NOT UNDERSTANDING THE DETAILS

Aligning incentives is one way to protect managers and technicians from emotions deriving from FOLS. However, numerous inadvertent slights still permeate most interactions between managers and technicians, often because the former are terse and the latter are verbose. For example, some managers reportedly spend over 20 hours per week in meetings.[6] As a result, they may not have time to send detailed e-mails when they wish to obtain information from a technician, and may send a note that says, "We need to talk." The technician will likely then worry about what requires discussion and why the note could not specify what it is. He or she will begin to analyze how to respond to several of the endless possibilities, ultimately becoming exhausted and anxious by the time the manager becomes free. The manager will then feel overwhelmed by the technician's anxiety and preparation advantage when discussion commences, which puts he or she too on the defensive. A vicious cycle ensues.

The cycle is easily generalized. Per the previous section, little more separates how some managers and technicians obtain their stations than the chance results of prior risk taking. Yet both parties often identify with tropes that one "gets business" while the other "gets numbers." These tropes can be useful for identifying project roles, specifically who is handling various tasks such as final decision-making – but they also leave all parties feeling

[3] Warrell (2013) describes various fears triggered by the possibility of taking a risk.
[4] Of course, doing nothing is often the most reasonable strategy – see Taleb (2017).
[5] Balani (2018) points to resource limitations as one reason why doing nothing is often a default position.
[6] Perlow et al (2017) suggests managers' time spent in meetings has more than doubled since the 1960s.

underestimated. For this reason, communications strategies that pander to tropes reinforce negative emotion. For example, some dimly suggest that technical content must be simplistic and catchy to engage "non-technical" audiences such as managers. Yet a natural reaction to receiving information presented in this way may be, "s/he must think I'm stupid!" This then leads managers to ask questions that illustrate technicians are equally "stupid" when it comes to the business. Each focuses more on perceived capability than problems at hand.

In contrast, being yourself is easier than "selling" others, and all parties should focus on presenting the truest versions of their work rather than altered versions of themselves.[7] For technicians, sharing a report in advance of a face-to-face discussion shows confidence in a manager's ability to interpret it, and the latter probably will not have time to give it more than a skim anyway. Rather than investing in a second career in digital marketing, the technician should invest in simple format changes to ensure the skim properly orients the manager to discuss further, not unlike how they might make the same changes to spruce up the document for a technical peer. For example, a data scientist may accentuate calls to action in decisive red, while banking valuable positive emotional capital by highlighting areas of present strength in a more tranquil blue.[8] S/he can use white space as relaxing intermezzos between key points. None of this is hard or requires altering the substance of a report and maintains a technical vernacular to the report that in turn preserves the glory of identifying a business solution for the manager.

Returning to our earlier example of an insurer whose projected expenses exceed its revenues, it does not take an advanced mathematics degree to identify a basic inequality, nor does it require extensive business acumen to know how to plug a revenue shortfall. Some may argue that technical presentations to managers should cut to the chase and focus on what findings mean for the business. This depends in part on the personalities involved, but these behaviors mostly just reinforce the cycle. For example, the actuary may have used gradient boosting (or any other mysterious-sounding algorithm of the reader's choosing) to isolate the shortfall to a specific segment of the book, and surrogate models to identify variables that describe that segment. It may be as obvious to the actuary that a rate increase or non-renewal strategy is necessary for the targeted cohort, as it is to the manager that the

[7] Peñarredonda (2018) describes the importance of psychological safety in the workplace.
[8] Williams (2007) reviews examples of the moods created by different colors.

explanatory variables intuitively correlate with risk. The actuary can "lead a horse to water" with prompts and visuals but should resist the temptation to make him or her drink. By staying in their respective lanes, neither the manager nor the technician looks "stupid" by having a perceived novice explain how to do their jobs. The manager looks smarter by asking intelligent questions and extracting business insights from math, as does the technician by anticipating questions and having answers ready. The unfortunate tropes survive, but neither party overcorrects for them, minimizing their harm.

4. BEING SMART … BY "GETTING STUPID"

Though well meaning, conversations like the one we are having do more to harm dynamics between managers and technicians than they do to help. They create a mythos that the two parties are fundamentally different, and they create unreasonably high expectations for the interactions. Because they often focus on the technician's role, they absolve managers of responsibility to make such interactions positive. The absolution in turn disempowers managers, as if they are incapable of doing anything to make life easier. The conversation makes everyone fearful of looking stupid. To speak technically, it divides us, multiplies hard feelings, and subtracts from self-worth. This essay adds one more opinion to a pile of existing and conflicting literature referenced throughout the document.

So how do we solve the problem of the less than ideal interactions, aside from fewer essays? Above I have outlined some simple steps technicians may consider in the short term to better empathize with managers' FOLS (and recognize their own FOLS) – by assuming shared responsibility for managers' risk aversion and unselfishly ceding opportunities to draw logical conclusions.[9] In a literal sense, one is more work for the technician, the other less. Longer term, all parties may consider a colorful slang expression called "getting stupid," which is defined as wild, unscripted dancing -- in other words, pure joy. The best managers and technicians take incredible joy in using their strengths to solve problems and celebrating their impact together.[10] "Geeking out" over a killer technical analysis and/or business strategy may not be proper decorum, and some may call it a waste of time. This is our FOLS talking. The more technicians start "getting stupid," the more senior managers

[9] 72% of CEOs feel the state of empathy in their organizations needs to evolve (businessolver 2019).
[10] Morgan (2011) points out that a presenter's passion helps makes dry subject matter interesting.

will follow their example, and the sooner waves of joy will overcome barriers of fear in their businesses. We all will be smarter when that day arrives.

5. REFERENCES

[1] J. Balani, "The Key to Presenting to Senior Executives," *Forbes*, **April 27, 2017**.
[2] N. Morgan, "5 Tips to Present Boring Information So It Isn't Boring," *Forbes*, **June 8, 2011**.
[3] J. Penarredonda, "Yes, you should really 'be yourself' at work," *BBC*, **November 29, 2018.**
[4] L. Perlow, C.N. Hadley, and E. Eun, "Stop the Meeting Madness," *Harvard Business Review,* **July 2017.**
[5] "State of Workplace Empathy," *businessolver*, **2019**.
[6] N. Taleb, "On Interventionistas and Their Mental Defects," *Medium*, April 21, 2017.
[7] E. Torrence, "Pushing Past the Fear of Looking Stupid," *Thin Difference*, **March 28, 2017**.
[8] R. Wagner, "New Evidence The Peter Principle is Real – And What to Do About It," *Forbes*, **April 10, 2018.**
[9] M. Warrell, "Take a Risk: The Odds Are Better Than You Think," *Forbes*, **June 18, 2013**.
[10] J. Williams, "Your Brand's True Colors," *Entrepreneur*, **March 7, 2007**.

Abbreviations and notations
FOLS, fear of looking stupid

Biography of the Author

Jim Weiss is senior actuary at Crum & Forster, where he is responsible for commercial lines predictive modeling. Previously he held various pricing, modeling, and data management roles at Insurance Services Office. Jim is past president of the Casualty Actuaries of Greater New York and former chairperson of the CAS Working Party on Microinsurance. He is a Fellow of the CAS, a Member of the American Academy of Actuaries, a Certified Specialist in Predictive Analytics, and Chartered Property Casualty Underwriter.

A by Layer Approach Algorithm for Computing Increased Limits Factors -- with Adjustments for Varying Policy Limits and Other Common Concerns

Joseph A. Boor, FCAS, CERA, Ph.D.

Abstract: ***When computing increased limits for short or medium tail lines of business, it is common to begin with loss data from a combination of policies with different policy limits. This has ramifications not only for the computation of the increased limits factors based on the experience data, but also for computing the credibility at different limits. The paper shows how analyzing the costs of the various layers easily accommodates corrections for the varying policy limits underlying the data. It contains an approach that replaces classical credibility with best estimate credibility. It also shows how Miccolis test disparities in the credibility weighted data may be readily and objectively resolved by using interpolation along a Pareto-based increased limit curve. Of note, large portions of this paper simply relate current best practices, although the discussions using best estimate credibility and interpolation along the curve are arguably new.***

Keywords: ***increased limits factors, policy limits truncation, credibility, classical credibility, best estimate credibility, interpolation along a curve***

1. INTRODUCTION

Because large claims are fairly rare, when actuaries develop or revise increased limits factors (hereafter "ILF"s) they often find that they do not have enough internal data to reliably predict the upper layers. However, it is important to effectively use the data that is available. For short or medium tail lines of business, ILFs may be developed directly from the claim-by-claim data from accident years where all the claims are closed, rather than by curve fitting or other more complex processes. The long lag until claims are closed complicates the process in long-tail lines, though. The data at the upper limits is usually distorted[1] by the fact that the policies with lower limits do not cover claims at those upper limits. Hence the data in the upper limits is "censored". Also, external benchmark ILFs are often used to supplement the experience data of the company using a credibility process. As will be shown in the remainder of the paper, the best approach to deal with these issues involves looking at loadings for each "layer" between one policy limit and the next limit above it. In effect, one would look at the losses that exceed each lower limit but cap each loss in or above the next limit at the difference or "length" between the two limits.

[1] Issues that might be mentioned, that fall outside the immediate scope of this model, arise when large deductibles eliminate, say, the basic layer (suggesting that, for purposes of this analysis, they should be excluded from the analysis) or when upper layer costs are effectively specified by an specific excess of loss reinsurance contract.

Additional aspects of the algorithm are needed to resolve common issues. Hence, they are covered as well. For example, if credibility is used, it is important to reflect all claims that contribute to an ILF layer but only those claims that contribute to an ILF layer. Further, since the ILFs arising from a credibility process may be inconsistent, a process to correct the inconsistencies is sometimes needed. Similarly, an approach for best estimate credibility rather than classical credibility is needed to provide a full algorithm to deal with most increased limits ratemaking situations.

The organization of this paper is as follows: first, the key features of the algorithm will be listed, then the reasoning behind each feature will be presented, with a comprehensive example showing how each works in practice.

2. THE KEY FEATURES OF THE ALGORITHM

2.1 A LIST OF THE KEY FEATURES

There are six major aspects of the scheme that are discussed within the paper:

1) Rather than computing the ILFs directly from data capped at the various limits, compute the layers separately, and combine them to create the ILFs.
2) Offset policy limit truncation by applying adjustments for the ratio of the policies that cover each layer to all the policies sold;
3) Credibility weight the layer factors (layer relativities to the basic limit), not the ground-up ILFs;
4) When you count the claims for credibility, count all the claims that pass into or through each layer, not just those that have a final value in the layer;
5) Replace inconsistent values (that fail the "Miccolis test"[2]) in the higher layers with interpolated or extrapolated values computed by interpolation/extrapolation along a Pareto severity curve[3]; and
6) To improve the accuracy, consider using a best estimate credibility process.

When circumstances suggest it, some issues may be combined in some sections. The specific items covered in each section are included in the title.

Sample calculations accompany the exposition. The data in Table 1 will be used. It is itself an example of the type of raw data summary that should be reasonably possible to produce from an insurer's statistical system.

[2] Per Robert Miccolis' 1977 paper.
[3] Using the procedure in Boor 2014.

Note that the rationale or formula for each item is shown below each column number, although in this case everything is input data. Note also that items of a given type are grouped by vertical lines. Those conventions will continue throughout this article.

Table 1: Raw Data

	Experience Data			Other Data	
(1)	(2)	(3)	(4)	(5)	(6)
Data	Data	Data	Data	Data	Data
Limit	Number of Insureds (Exposures) (or Premium) at This Limit	Claims in Layer Ending at Limit	Aggregate Ground-Up Cost of Those Claims	Current ILF	Benchmark ILF
$250,000	100	200	$40,000,000	1.00	1.00
500,000	200	40	15,000,000	1.60	1.90
1,000,000	300	15	10,500,000	2.50	3.60
2,000,000	100	3	4,200,000	3.50	4.00
5,000,000	50	2	7,000,000	4.00	5.00
Total	750	260	$76,700,000		

2.2 COMPUTING THE LOSSES BY LAYER (Items 1 and 2 above)

As mentioned earlier, when the policies in a dataset have different limits of liability, the losses in the upper layers are not directly comparable to the losses in the lower layers. For example, if only half the policies have limits above $250,000, then the losses excess of $250,000 should be half [4] (or less for higher limits) what the cost would be if policy limits were not an issue. If only two thirds of those with policy limits of $250,000 or more have policy limits over $500,000, then of course the losses above $500,000 come from two thirds as many policies as those in the $250,000 to $500,000 layer. Taking it one step further, they would be two thirds of one half (one third---or less) what they would be if they were not "truncated" by lower policy limits of either $250,000 or $500,000 on many policies. Thus, there are strong reasons not to begin by capping the unadjusted losses.

The key is to look at the losses by layers, and correct for the different levels of policies (or other exposure units) by layer. So, one would have sets of losses that are adjusted so that they reflect the losses one would observe if the limit had the same number of policies, etc. as the lowest limit.

The first step is to compute the losses in each layer. For example, if a product offers liability limits with increasing values of $A, $B, $C, or $D, then

[4] As it turns out maybe less in dollars, considering the truncation in the succeeding layers.

- The first layer "$A" is covered by all the policies, so all the losses from all the policies that amount to $A or less are included. All the claims, whether their size is $A or above are included. However, the amount of each is of course capped at $A.
- The second layer from $A to $B, is an intermediate layer. Only claims from policies with limits of $B, $C, or $D potentially fall into this layer. The sizes of each of the claims must be at least $A, or they would not be in this layer. So, to isolate the portion that is in this layer, one would subtract $A from each claim size. Lastly, the excess amounts that result would be capped at $B-$A to reflect the "length" of the layer, and the results aggregated to produce the total losses in the band from $A to $B.
- Only claims from policies with limits of $C or $D fall into the next intermediate layer (from $C to $D). A similar process is employed here.
- $D is the topmost limit layer. For each claim over $C in size, one would subtract $C from the loss, to place it in the layer, and the results would be aggregated. Presumably[5] no loss is larger than the limit $D.

Once the aggregate loss dollars in each layer are computed, one must recognize that large claims (above the policy limit) on policies with lower policy limits are truncated below the limit---at their lower policy limit. Per the policy conditions, the losses in each layer must come from policies with limits equal to the top of the layer or higher. Any court judgments against the insured that are high enough to reach to the next higher layer are "censored" (essentially eliminated from the data) unless the policy has coverage for the higher limit.

That creates inconsistent exposures among the various limits. Logically, if half the policies (or premiums or some other exposure units) are written at a limit of $A, then the losses in the layer between $A and $B were generated by only half the exposures as those capped at $A. Therefore, to make the losses in the layer between $A and $B comparable to the base layer, one must multiply the computed losses in the layer by the ratio of the exposure passing in or through the base layer ($A) to the exposure passing in or through this second layer (limits of $B, $C, or $D). Of course, since all policies have a limit of at least $A, one should multiply the computed loss in the second layer by a "basic limits equivalence factor"[6] or "BLEF". That factor would be computed by dividing the total exposures in the data (limits of $A or higher) to the exposures with limits of $B or higher. Similarly, one would multiply the losses computed for the layer from $B to $C by the ratio of the total exposure units to those with limits of $C or $D. and for the layer between $C and $D, one would use a similar ratio with the number of exposure units with a limit of $D in the denominator. In this paper, policy counts are used for simplicity. However, something such as earned premiums at the present rate level might create more accurate results. Table 2A illustrates the computation of the BLEFs.

Before the BLEF can actually be applied, though, one must compute the losses by layer present in the unadjusted data. To facilitate the explanation of the process, it is helpful to define a couple of terms. First, one may call the groupings of ground-up losses (for example, all individual claims in

[5] In practice so called "extra-contractual obligations" may result in losses beyond policy limits, but those are beyond the scope of this discussion.

[6] The author recognizes that the basic limit is not always the lowest limit in the data. If it preferred, one may think of this as the "lowest limit equivalence factor". The case where the basic limit is not the lowest limit receives little focus in this paper because, although the same principles hold, it distracts from the key issues in the paper.

the data costing between $250,000 and $500,000) "size groups". Then the portions of individual claims, of whatever size group, that enter or exceed a given range are said to fall in that "layer". For example, on a $510,000 claim and $400,000 claim the portions in the $250,000 excess of $250,000 layer (the layer between $250,000 and $500,000) are $250,000 and $150,000, respectively. The total losses in the layer are the sum of all those claim-by-claim amounts in the layer. Because each size group includes "ground up" claims costs (including those in all the lower layers), the costs in each layer will not equal the costs in the corresponding size group. Generally, the costs will be shifted to the lower layers, as is proper.

Table 2A: Calculation of Basic Limits Adjustment Factors

(1)	(2)	(3)	(4)
Table 1 c.2	Table 1 c. 2	Sum (2) +[All (2) Following]	[Total (2)]/(3)
Limit	Number of Insureds (Exposures) at This Limit	Number of at This Limit at This or Higher Limit	Correction for Limits Truncation "BLEF"
$250,000	100	750	1.00
500,000	200	650	1.15
1,000,000	300	450	1.67
2,000,000	100	150	5.00
5,000,000	50	50	15.00
Total	750		

An example involving multiple layers starts with a total amount of $813,000. So, it is within the $500,000 to $1,000,000 size group. Then it generates the full $250,000 in the basic limit layer; a full $250,000 second layer ($250,000 excess $250,000, or from $250,000 to $500,000) loss; and $313,000 within the $500,000 excess $500,000 layer. The claims costs reach their total there, where the claim's cost ends. Further, since this loss falls in the band ending in $1,000,000, it is the equivalent of 1.67 losses per the BLEF.

The procedure for estimating the costs in each layer in Table 2B begins with total number of claims ending within each size group, and the total cost of those claims. For each layer below the size group, all the claims in the size group will exhaust the entire length of the layer. So, in each layer below the size group, one computes the aggregate cost of the layer by multiplying the number of claims in the size group by the length of the layer. That properly assigns each group's costs to the lower layers.

However, that does not cover the portion of the size group's claims that fall in the layer matching the size group. Since those claims do not fill their entire layer, their cost in the layer will not equal the full size of the layer, and the cost in the layer will likely vary from claim to claim. Thankfully a fairly simple process is available to compute the cost within that layer. The total aggregate cost for all the claims in this size group in all the layers must still equal the total cost of the claims in the size

group. So, one may simply subtract all the lower layer costs associated with this size group from the overall total for the size group to obtain the costs in the last, highest layer.

Now, all the losses in each layer must be added together to get the total unadjusted losses by layer. As mentioned previously, though, these are affected by policy limits truncation. As a last step, the BLEF is multiplied by the losses in each layer to obtain the final BLEF-adjusted losses. These are the layer losses that will be used in computing the layer factors and consequently computing the ILFs.

Now that the basic information is prepared, the next step is to complete the calculation of the experience-based ILFs using the layer costs in Table 2C. To relate the costs of the various layers to those of the basic limit, and produce the layer factors, one need only divide the total losses in the gray band at the bottom of Table 2B by the corresponding losses at the basic limit[7]. However, these are merely the costs of the individual layers. To provide ILFs, representing all the losses up to a limit, one must sum all the factors up to and including the layer at the top limit (as is done in column (6) below). This chart also includes a basic Miccolis test, as mentioned earlier. Column (5) shows a rate on line (per limit of coverage) for just the losses in each layer. The rates/relative costs must always decrease from lower limit to higher limit. Since this is constructed from actual loss data, it is inevitable that the results pass the test.

[7] If there are limits below the basic limit, one need only tentatively treat the lowest limit as the basic limit. Then, one would divide each tentative ILF by the tentative ILF of the basic limit so as to rebalance the factors. Then, the final basic limit ILF will be unity (1.00) and the ILFs proportional to that.

. **Table 2B: Computation of Loss by Layer and Correction for Policy Limits Truncation**

Part 1- Inputs

(1)	(2)	(3)	(4)	(5)
(1)	(1) -[previous(1)]	Table 1 c.3	Table 2A c.4	Table 0 c.4
Upper Limit	Layer Length	Claims in Layer Ending at Limit	BLEF	Ground-up Cost of Claims Ending in Layer
$250,000	$250,000	200	1.00	$40,000,000
500,000	250,000	40	1.15	15,000,000
1,000,000	500,000	15	1.67	10,500,000
2,000,000	1,000,000	3	5.00	4,200,000
5,000,000	3,000,000	2	15.00	7,000,000
Total		260		$76,700,000

Part 2 - Determination of Truncation-Correction Loses Passing into Each Layer

(6)	(7)	(8)	(9)	(10)	(11)
(1)	$250K*(3)	$250K*(3)	$500K*(3)	$1M*(3)	$3M*(3)
Upper Limit	Basic Limit Equivalent Cost of 250 X0 Layer Claims *	Basic Limit Equivalent Cost of 250 X250 Layer Claims*	Basic Limit Equivalent Cost of 500 X500 Layer Claims*	Basic Limit Equivalent Cost of 1000 X1000 Layer Claims *	Basic Limit Equivalent Cost of 3000 X2000 Layer Claims*
$250,000	$40,000,000	$0	$0	$0	$0
500,000	10,000,000	5,000,000	0	0	0
1,000,000	3,750,000	3,750,000	3,000,000	0	0
2,000,000	750,000	750,000	1,500,000	1,200,000	0
5,000,000	500,000	500,000	1,000,000	2,000,000	3,000,000
Total Pre-Adjust	$55,000,000	$10,000,000	$5,500,000	$3,200,000	$3,000,000
BLEF	1.00	1.15	1.67	5.00	15.00
BLEF Adjusted Losses by Layer					
Total	**$55,000,000**	**$11,538,462**	**$9,166,667**	**$16,000,000**	**$45,000,000**

* The diagonal elements in (7)-(11) were computed as the adjusted (multiplied by item (4)) losses in column (5), minus the sum of the losses in the in the columns to the left within each row.

As one may see, this concludes with the final experience-based ILFs. The next steps involve applying credibility and trending to the results of the above process.

Table 2C: Final Computation of Experience-Based ILFs (with Miccolis Test)

(1)	(2)	(3)	(4)	(5)	(6)
Table 1 c.2	Table 2B c. 2	Totals at Bottom of Table 2B c.(7),(8),(9),(10),(11)	(3)/ [(3) for 250K]	1M *(4)/(2)	Cumulative Sum of (4)
Upper Limit	Layer Length	Basic Limit Equivalent Cost Total Loss	Experience Based Layer Factors	(Miccolis) Relative Cost per Layer Size	Experience Based Increased Limits Factors (ILFs)
$250,000	$250,000	$55,000,000	1.00	4.000	1.00
500,000	250,000	11,538,462	0.21	0.839	1.21
1,000,000	500,000	9,166,667	0.17	0.333	1.38
2,000,000	1,000,000	16,000,000	0.29	0.291	1.67
5,000,000	3,000,000	45,000,000	0.82	0.273	2.49
				(pass)	
Total		$136,705,128			

2.3 CLASSICAL CREDIBILITY BY LAYER (Items 3 and 4 above)

There are two key concerns when classical credibility is used in increased limits ratemaking. As stated earlier, the credibility process should apply to the layer factors, not the ILFs. Also, since classical credibility revolves around claim counts, one should take care to count the number of claims correctly.

The composition of the ILFs, as a mixture of the basic layer costs, the first excess layer costs, the second excess layer costs, etc. is what necessitates credibility weighting layers. Say, for example, that the $500,000 experience-based ILF had a credibility of 60%. Then, in recognition of the lesser claims above $500,000, the $1,000,000 experience-based ILF was assigned a credibility of 40%. While that may superficially seem proper, note that the losses up to $500,000 limit are also part of the data used to compute the $1,000,000 limit. In one situation, they have 60% credibility, but in the second they only have 40%. This logical inconsistency requires a solution. The most straightforward way is to credibility weight the individual layer factors rather than the ILFs. That way, each subject of credibility is treated separately from the others.

The other aspect of credibility is the calculation of the actual credibility values. Due to potential technical complexity, this paper will not deal with how to set a full credibility standard or determine the expected number of claim counts in each layer. Rather it will assume some full credibility standard and will use the actual claim counts rather than expected numbers of claims. That means

that the credibility of each experience-based layer factor, using classical credibility, will be the square root of the result of dividing the number of claims used to compute the factor by the full credibility standard. So, the next step is determining exactly how many claims are present in the data used to compute each layer.

It is tempting to simply use the number of claims in the corresponding size group. But that would ignore the number of claims that simply pass through the layer, contributing cost information as they do. In fact, all the claims from higher layers should be included in the claim counts of the layers below them. For example, consider the $250,000 excess of $250,000 layer. The claims in the $500,000 size group are part of it. But, the claims in the $1,000,000 size group also pass through that lower layer (at $250,000 per claim of loss). Similarly, the claims in the $2,000,000 size group, and all higher size bands, need to have their claims included in the claim count for the $250,000 excess $250,000 layer. As a general principle, the number of claims for credibility of a layer should count those in the matching size group plus all those in higher size groups.

Several tables are needed to illustrate the classical credibility process for layer factors. Since the credibility weighting is performed by layer, the benchmark layer factors used as input are computed in Table 3A.

Table 3A: Calculation of Benchmark Layer Factors -With Miccolis Test

(1)	(2)	(3)	(4)	(5)
Table 1 c.1 (Offset)	Table 1 c.1	Table 1 c.6	(3)-[previous (3)]	1M*(4)/((2)-(1)) (Miccolis)
Bottom of Layer	Top of Layer	Benchmark ILF	Benchmark Layer Factors	Relative Cost per Layer Size in Benchmark
$0	$250,000	1.00	1.00	4.00
250,000	500,000	1.90	0.90	3.60
500,000	1,000,000	3.60	1.70	3.40
1,000,000	2,000,000	4.00	0.40	0.40
2,000,000	5,000,000	5.00	1.00	0.33
				(pass)

Now that the values for the complement are determined, the counts are summed from above in Table 3B to obtain the counts used in computing the credibilities.

Table 3B: Counts Used in Computing Layer Factors for Classical Credibility

(1)	(2)	(3)	(4)
Table 1 c.1 (Offset)	Table 1 c.1	Table 1 c.3	(3)+[(3) Above]
Bottom of Layer	Top of Layer	Claims in Layer Ending at Limit	Number of Claims at or Above Size Layer for Credibility
$0	$250,000	200	260
250,000	500,000	40	60
500,000	1,000,000	15	20
1,000,000	2,000,000	3	5
2,000,000	5,000,000	2	2

At this point, all the input needed to perform the classical credibility process is available. The next step is to simply execute the classical credibility procedure. In this case, a full credibility standard[8] of 683 claims is used. One may follow the calculations in Table 3C, which uses the number of claims used in pricing each layer from Table 3B, the experience-based layer factors from Table 2C, and the benchmark layer factors from Table 3A.

[8] That is not expressed to be optimal or more proper than any other standard. The purpose here is just to illustrate how the calculations flow.

Table 3C: Credibility Weighted ILFs from Classical Credibility

Part 1-Credibility Calculation and Input Data

(1)	(2)	(3)	(4)	(5)	(6)
Table 1 c.1 (Offset) Bottom of Layer	Table 1 c.1 Top of Layer	Table 3B c.4 Claims Passing Into This Layer	((3)/683)^.5 Classical Credibility for Layer	Table 2C c.4 Experience Based Layer Factors	Table 3A c.4 Benchmark Layer Factors
$0	$250,000	260	62%	1.00	1.00
250,000	500,000	60	30%	0.21	0.90
500,000	1,000,000	20	17%	0.17	1.70
1,000,000	2,000,000	5	9%	0.29	0.40
2,000,000	5,000,000	2	5%	0.82	1.00

Part 2 - Result of Classical Credibility and Miccolis Test

(7)	(8)	(9)	(10)	(11)
Table 0 c.1 (Offset) Bottom of Layer	Table 0 c.1 Top of Layer	(4)*(5) +[1.0-(4)]*(6) Credibility Wtd Layer Factor	1M*(9) /((2)-(1)) (Miccolis) Relative Cost per Layer Size	Sum from top of (9) Credibility Adjusted ILF
$0	$250,000	1.00	4.00	1.00
250,000	500,000	0.70	2.78	1.70
500,000	1,000,000	1.44	2.88	3.13
1,000,000	2,000,000	0.39	0.39	3.52
2,000,000	5,000,000	0.99	0.33	4.51
			(fail)-in gray	

Unfortunately, even when the input is two sets of ILFs that pass the Miccolis test, their credibility weighted combination may not pass it. The area in gray in Table 3C illustrates how that can happen. Nevertheless, this example does illustrate the proper application of classical credibility in the ILF

estimation process. A separate process, shown in the next section, is needed to resolve any Miccolis test discrepancies[9] that arise.

2.4 FIXING INCONSISTENCIES BY USING INTERPOLATION ALONG A CURVE (Item 5 above)

As the previous example shows, otherwise actuarially proper calculations of ILFs sometimes give rise to Miccolis test inconsistencies. Therefore, having a method ready to resolve such problems can be helpful. Interpolation along a curve (from Boor 2014) is the core of such a method. A curve can be fit to all the ILFs (not layer factors, the fit to the ILFs is more straightforward) except the problem factors that need to be replaced. Then, that curve is the basis for replacement for the problem point, using interpolation along the curve. Of course, the curve usually will not match all the data points (ILFs) exactly. However, interpolation along the curve[10] alters the curve so that it matches the data points exactly. So, if the value on the curve f at a point c between a and b is f(c)= f(a) + x%[f(b)-f(a)], or f(c) is x% of the way from f(a) to f(b); then the ILF estimate at c would be ILF(a)+x%[ILF(b)-ILF(a)]. The table on the next page shows the approach in practice, starting from the results of Table 3C.

The process involves two stages. First, the increased limits curve produced by a Pareto distribution is fit to all the ILFs but the one associated with the layer factor that failed the Miccolis test. The curve formula is shown, but of course the two parameters of the curve must be chosen. To that end, a least squares error method is used. The squared differences between the values from the curve and the actual values are computed in column (5) of Table 4A, with the total of the column in dark gray. Then, the solver routine in the spreadsheet software identifies the alpha and truncation point[11] that generate the least squared error between the actual and fitted values. Once the curve is fit, both the curve values and the actual ILFs for the layers ending at $500,000 and $2,000,000, with the problem $1,000,000 limit in between are all available. The fitted curve provides a value of 2.407 for the $1,000,000 limit. Since the fit is not exact, using the fitted value can sometimes still create a failure of the Miccolis test or other inconsistency, so interpolation along the curve is used. Following the formula shown at the bottom of the table, the Pareto-based curve values at $500,000, $1,000,000 and $2,000,000 are used in conjunction with the actual values at $500,000 and $2,000,000 to estimate a value at the $1,000,000 limit (2.54) that is consistent with the values around it.

[9] To explicitly discuss the issue, although the calculation of classical credibility using layers may produce results that fail a Miccolis test, it would be expected to do so less often than when credibility is applied to the entire ILF.

[10] An extensive explanation of interpolation along the curve is not included here. For a more detailed discussion one may review the article.

[11] In this case, it was necessary to cap the truncation point at $200,000 to keep it reasonably below the lowest limit ($250,000).

Table 4A: Extrapolation to Replace Inconsistent ILFs from Classical Credibility Analysis

		Solver Setup			
		Alpha	**0.7787**	Lt Gray Values Solved	
		Truncation Pt.	**200,000**	to Minimize Dark Gray	
				$200K is upper restriction]	
(1)	(2)	(3)	(4)	(5)	(6)
Table 1 c.1 (Offset)	Table 1 c.1	Table 3C c.11	***	((3)-(4))^2	(3) and see note **** below
Bottom of Layer	Top of Layer	Classical Credibility Weighted ILFs	Fitted Pareto	Squared Fit Error	Revised ILFs
$0	$250,000	1.00			1.00
250,000	500,000	1.70	1.641	0.00301	1.70
500,000	1,000,000		2.387		2.54
1,000,000	2,000,000	3.52	3.258	0.07072	3.52
2,000,000	5,000,000	4.51	4.634	0.01440	4.51
Total Squared Error				**0.08813**	
$500,000	$1,000,000	****2.54	2.387		

Notes: *** Fitted Pareto values are
{Alpha - [(Truncation/[Column (2)])^(Alpha-1.0)]} / {Alpha - [(Truncation/250,000)^(Alpha-1.0)]}
"**** Value per Interpolation Along the Curve is
1.70 + (3.52-1.70)*(2.387-1.641)/(3.258-1.641)

The interpolated value appears to be reasonable. But, it makes sense to recheck the Miccolis test, per Table 4B following.

Table 4B: Miccolis Test After Problem ILF Corrected by Interpolation

(1)	(2)	(3)	(4)	(5)
Table 1 c.1 (Offset)	Table 1 c.2	Table 4A c.6	(3) -[previous(3)]	$1M*(4)/((2)-(1))
Bottom of Layer	Top of Layer	Revised ILFs	Layer Factors in ILFs	(Miccolis) Relative Cost per Layer Size
$0	$250,000	1.00	1.00	4.00
250,000	500,000	1.70	0.70	2.78
500,000	1,000,000	2.54	0.84	1.69
1,000,000	2,000,000	3.52	0.98	0.98
2,000,000	5,000,000	4.51	0.99	0.33
				(pass)

As one may see, the correction using interpolation along the curve eliminates the Miccolis test inconsistency. One may also note that, although when very disparate sets of ILFs are used in classical credibility, there is a tendency to create Miccolis test failures, Miccolis test failures may occur in many other contexts.

2.5 BEST ESTIMATE CREDIBILITY FOR INCREASED LIMITS (BY LAYER) (Item 6 above)

Along with the presentation of classical credibility for ILFs, it makes sense to introduce an approach to best estimate credibility for the layer factors. So, in addition to often-stable ILF estimates produced by classical credibility, there will be an option for creating estimates that come as close as practicable to the true underlying loss costs in the excess layers.

The basic concept is not very complicated. Per Boor 1992, one need only estimate the expected squared prediction error that each statistic (the empirical layer factor and the benchmark layer factor) makes when estimating the true cost of the layer. Then, per Boor, the credibility of each statistic equals the squared error of the other statistic, divided by the sum of the two squared errors. Therefore, the credibility of each piece of data is proportional to how poorly the other item predicts the losses.

Estimating the expected squared error the benchmark factors make when predicting the layer costs involves a fairly simple calculation. Of course, the benchmark layer factors are not random,

but the true, unknown, layer costs are. Further, this is a process of approximation, given the information that is available. So, the empirical layer factors may be used as a proxy for the true underlying layer factors. So, the squared differences between the benchmark layer factors and empirical factors are used to estimate the expected squared estimation errors made the benchmark layer factors. For another estimate of the estimation errors made by the benchmark layer factors one may also compare the benchmark layer factors to the layer factors that are currently being used[12]. Since the new empirical data may sometimes be very thin in the upper layers, this will yield a more reliable result in some cases, especially where the empirical data is thin.

The calculations are shown in Table 5A. Recognizing that the experience data in the upper layers may be limited, the most logical estimate of the benchmark squared prediction error for this class of data is made using both those indicators.

The expected squared error of the experience-based layer factors requires more work. The key involves the so-called "collective risk" model. As noted in Dean and Mahler 2001, when one draws a random number ($\boldsymbol{n}$) of independent losses($\boldsymbol{S_i}$'s) from a single severity distribution, the variance of that aggregate distribution is Mean($\boldsymbol{n}$) ×Variance($\boldsymbol{S}$)+Variance($\boldsymbol{n}$) ×Mean^2($\boldsymbol{S}$). Then, in the absence of any other evidence, it is usually logical to assume that all the loss occurrences are completely independent. Thus, the counts would follow a Poisson distribution with expected value of, say '$\boldsymbol{N}$'. Under those assumptions, the variance of the aggregate distribution is $\boldsymbol{NE[S^2]}$. That formula will be the basis for determining the process variance of the empirical layer factors.

Just as with the prediction error made by the benchmark, a workable "basic arithmetic" estimate is presented rather than a more complex calculation. The first step is to calculate this "process variance" for each layer when the BLEF correction is excluded from the analysis.

Then, to define a couple of values for a given layer, $\boldsymbol{n}$ will be the number of claims that passed through or stopped in the layer, and $\boldsymbol{k}$ the number of those that passed all the way through (exceeded the limit of the layer). Of course, one must calculate those two values. In the pre-BLEF context of the calculations, the values are computed using the actual, unadjusted by BLEF, counts of the number of claims that passed all the way through the layer, and the count of those that stopped within the layer. Once those are determined, one must compute the average squared loss (within the layer) for each type. It should be clear that every claim that passes all the way through the layer generates a cost in the layer equal to the full length of the layer. Hence, $\boldsymbol{S^2}$ will be the square of the layer length for those claims. In the spirit of estimation, each claim that stopped or ended within the layer may be assumed to be half the layer length. So, for the claims that end in the layer, $\boldsymbol{S^2}$ will be one-fourth the square of the layer length. Thus, for the $\boldsymbol{n}$ historic claims that passed all the way through the layer, and the $\boldsymbol{k}$ claims whose total cost ended in the layer, the variance of the aggregate costs in the layer is estimated by $\boldsymbol{(n + \frac{k}{4})}$× (the square of the layer length). Per the collective risk model, that estimates the process variance associated with the actual claims, as an estimate of the pre-BLEF layer cost.

[12] For a related analysis, one may review Marcus 2010.

Table 5A: Estimation of Squared Errors from Benchmark Layer Factors

Part 1 - Experience, Current and Benchmark Layer Factors

(1)	(2)	(3)	(4)	(5)	(6)
Table 1 c.1 (Offset)	Table 1 c.1	Table 2C c.4	Table 1 c.5	(4) -[previous(4)]	Table 3A c.4
Bottom of Layer	Top of Layer	Experience Based Layer Factors	Current ILF	Current Layer Factors	Benchmark Layer Factors
$0	$250,000	1.00	1.00	1.00	1.00
250,000	500,000	0.21	1.60	0.60	0.90
500,000	1,000,000	0.17	2.50	0.90	1.70
1,000,000	2,000,000	0.29	3.50	1.00	0.40
2,000,000	5,000,000	0.82	4.00	0.50	1.00

Part 2 -Estimation of Squared Error of Benchmark

(7)	(8)	(9)	(10)	(11)
Table 1 c.1 (Offset)	Table 1 c.1	[(3)-(6)]^2	[(5)-(6)]^2	(9),(10) selection
Bottom of Layer	Top of Layer	Squared Benchmark Errors vs. Experience Factors	Squared Benchmark Errors vs. Current Factors	Selected Benchmark Error Parameter
$0	$250,000	-	-	-
250,000	500,000	0.476	0.090	0.100
500,000	1,000,000	2.351	0.640	0.800
1,000,000	2,000,000	0.012	0.360	0.200
2,000,000	5,000,000	0.033	0.250	0.200

The first step is to compute the number of historical claims that passed all the way through or "exceeded" each layer, and the number of actual claims that ended up (or whose total cost lay) in each layer. Both are done in Table 5C. Of course, the number of claims ending in the layer is simply the counts from Table 1.

The next step is to compute the process variance of the actual data using the formula $(\boldsymbol{k}+\frac{\boldsymbol{k}}{\boldsymbol{4}})\times$ (the square of the layer length). Those calculations are performed in Table 5B. As a final step, the impact of the BLEF that was part of the experience-based layer factor calculation in Table 2B, but has not been considered so far, is applied. One may recall that because of policy limits truncation, the raw claim counts that underlie the calculation differ from the BLEF-adjusted counts used in the final computation of each layer factor. However, since the BLEF is simply a multiplier that applies to the layer[13], one need only multiply the process variance of the actual data discussed above by the $BLEF^2$, per the standard variance formula involving a constant multiplier to a random variable. That produces an estimate of the process variance of the full, inclusive of the BLEF, experience-based estimates of the layer factors. The column 11 in Table 5B shows the final resulting process variance values associated with the experience-based layer factors.

[13] Recall that this calculation is done in the spirit of approximation. A possibly more exact estimate of the mix between losses exceeding and ending up in various layers might be obtained with a different calculation. However in this case the variance calculation seems to flow somewhat smoothly with the BLEF correction as the last step.

Table 5B: Estimation of Process Variance in Experience-Based Layer Factors

Part 1 - Inputs for Poisson Collective Risk Model

(1)	(2)	(3)	(4)	(5)	(6)
Table 1 c.1 (Offset)	Table 1 c.1	Table 1 c.3	Sum of (3) Above Limit	[(2)-(1)]^2	{[(2)-(1)]^2} /(4)
Bottom of Layer	Top of Layer	Actual Claims Ending In Layer	Actual Claims Exceeding Layer	Squared Layer Length	Squared Half-Length
$0	$250,000	200	60	6.25E+10	1.56E+10
250,000	500,000	40	20	6.25E+10	1.56E+10
500,000	1,000,000	15	5	2.50E+11	6.25E+10
1,000,000	2,000,000	3	2	1.00E+12	2.50E+11
2,000,000	5,000,000	2		9.00E+12	2.25E+12

Part 2 - Final Process Variance Estimate with All Corrections

(7)	(8)	(9)	(10)	(11)
Table 1 c.1 (Offset)	Table 1 c.1	(3)*(5) +(4)*(6)	(9)*BLEF^2 {from Table 2A c.4}	(10)/[(Basic Limits Loss)^2] {from Table 2C c.3}
Bottom of Layer	Top of Layer	Raw Data Process Variance Estimate per Collective Risk	Process Variance of Total Losses in Experience Layer	Process Variance of Layer Factor Estimate
$0	$250,000	6.88E+12	6.88E+12	0.002
250,000	500,000	1.88E+12	2.50E+12	0.001
500,000	1,000,000	2.19E+12	6.08E+12	0.002
1,000,000	2,000,000	2.75E+12	6.88E+13	0.023
2,000,000	5,000,000	4.50E+12	1.01E+15	0.335

Now, the process variance and the expected squared error of the benchmark are available. Per the Boor paper the best estimate credibility of the experience data will be the squared error of the benchmark divided by the sum of the squared benchmark error and the process variance. The results and the corresponding layer factors are shown in Table 5C. A Miccolis test is included as well.

Table 5C: ILFs per Best Estimate Credibility

Part 1 - Credibilities and Inputs

(1)	(2)	(3)	(4)	(5)	(6)
Table 1 c.1	Table 1 c.1 (Offset)	Table 2C c.4	Table 5A c.11/ [Table 5A c.11+ Table 5B c.11]	Table 3A c.4	1.0-(4)
Bottom of Layer	Top of Layer	Experience Based Layer Factors	Credibility of Experience	Benchmark Layer Factors	Complement of Credibility
$0	$250,000	1.00	100%	1.00	0%
250,000	500,000	0.21	99%	0.90	1%
500,000	1,000,000	0.17	100%	1.70	0%
1,000,000	2,000,000	0.29	90%	0.40	10%
2,000,000	5,000,000	0.82	37%	1.00	63%

Part 2 - Final Credibility Weighted Best Estimates

(7)	(8)	(9)	(10)	(11)
Table 1 c.1	Table 1 c.1 (Offset)	(3)*(4) +(5)*(6)	1M*(9)/[(2)-(1)]	[Cumulative Sum of (9)]
Bottom of Layer	Top of Layer	Best Estimate of Layer	(Miccolis) Relative Cost per Layer Size of Z-Wtd ILF	Best Estimate ILFs
$0	$250,000	1.00	4.00	1.00
250,000	500,000	0.22	0.86	1.22
500,000	1,000,000	0.17	0.34	1.39
1,000,000	2,000,000	0.30	0.3020	1.69
2,000,000	5,000,000	0.93	0.3107	2.62
			(very slight fail)	

In this case, the benchmark does not get much weight, primarily because it is so different from both the raw experience-based layer factors and the current layer factors. However, in the upper layers where there are smaller numbers of claims, it has substantial credibility. One may notice that this process avoids radical shifts in the layer factors when the benchmark is changed. So, as long as at

least one of the current and experience-based layer factors may be thought of as fairly representative of the underlying severity distribution, this produces truly optimal [14]estimates of the ILFs.

Also note that in this case there was only a minor third decimal place inconsistency in the Miccolis test, for which the correction [15] is not included here. By assigning less weight to benchmarks with greatly different layer factors, this approach reduces, but may not eliminate, the possibility of a Miccolis test failure. In general, this method has much to commend it.

3. SUMMARY

The previous sections show how a layer-by-layer approach to computing ILFs leads to much more appropriate calculations, a more logical approach to classical credibility, and even a best estimate approach to credibility for ILF calculations. The reader is encouraged to use this for more reliable, logically consistent calculations.

4. REFERENCES

1) Boor, Joseph A., **'Credibility Based on Accuracy',** *Proceedings of the Casualty Actuarial Society,* Casualty Actuarial Society, Arlington, Virginia 1992: *Vol. LXXIX pp. 166-185*
2) Boor, Joseph A., **'Interpolation Along a Curve',** *Variance,* Casualty Actuarial Society, Arlington, Virginia 2014: *Vol. 8 Issue 1 pp. 9-22*
3) C. Dean and H. Mahler., **Credibility**. In **Foundations of Casualty Actuarial Science**, Casualty Actuarial Society, Arlington, Virginia, fourth edition, 2001, *pp. 485-660.*
4) Marcus, Lawrence F., '**Credibility for Experience Rating, a Minimum Variance Approach'**, *Casualty Actuarial Society E-Forum,* Casualty Actuarial Society, Arlington, Virginia 2010: *Summer, pp. 1-18*
5) Miccolis, Robert S., **'On the Theory of Increased Limits and Excess Loss Ratemaking',** *Proceedings of the Casualty Actuarial Society,* Casualty Actuarial Society, Arlington, Virginia 1977: *Vol. LXIV pp. 27-59*

[14] Note that this does require that at least the current layer factors, hopefully also the experience-based layer factors, be reasonable estimates of the true underlying costs.

[15] The correction would follow extrapolation along the curve per Boor 2014, fitting the curve and multiplying all the extrapolated values by a factor so the result matches the closest actual data point used (in this case, 1.68 at $2,000,000).

An Actuarial Approach to Behavioral Ratemaking: How Fair Rates Will Encourage Safer (and Slower) Driving

Michael C. Dubin, FCAS, FSA, MAAA, FCA

Abstract: Many people regularly drive above the posted speed limit. This type of behavior is risky and the cause of much loss, including loss of life.[1] The World Health Organization has identified speeding[2] as a global health issue.[3] The insurance industry can reduce this loss by implementing a new approach to ratemaking, behavioral ratemaking.[4] The use of current driving speed data (and other telematics[5] data) to adjust insurance pricing on a real-time basis can be used to encourage safer driving behavior and a safer society. In other words, in this model a driver would pay real time for how they drive as they drive. Hereinafter "behavioral ratemaking" is used to denote insurance rates that change in real time. This article discusses what behavioral ratemaking is and how it would operate in this context. It discusses how behavioral rates could be developed, the advantages they present and the logistical, technological and regulatory obstacles preventing their implementation.

WHAT IS BEHAVIORAL RATEMAKING?

Anyone who has taught a child to drive knows that the most important way to reduce the chance of an accident is through safe driving behavior. Since the insurance industry pays for the financial consequences of accidents and other insured events, it would seem they would and should be a promoter of safety as well. "Hazard reduction incentives" are a consideration in designing any insurance risk classification system.[6] However, traditional auto insurance ratemaking uses classification systems that strive to place drivers into classes with homogenous risks based on factors such as age, sex and marital status that do not directly measure risk and do not utilize driving behavior. When behavioral risk is considered in traditional ratemaking, such as in claims or violations history, past rather than current behavior is measured. Walters states, "One of the reasons for classifying is the impossibility of knowing the risks true expected loss or accident likelihood."[7] This is no longer as clear as it was in 1981 as recent technology rapidly advances the potentials of ratemaking. With the

[1] While fatal accidents do not represent the majority of auto insurance claim costs, it is assumed throughout this paper that behaviors reducing fatal accidents also reduce other types of accidents.
[2] Both excess speeding (exceeding posted speed limit) and inappropriate speeding (driving at a speed unsuitable for the prevailing road conditions).
[3] http://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/speed_en.pdf. World report on road traffic injury protection, World Health Organization, 2004.
[4] In this paper, behavioral ratemaking is applied to auto insurance. It can also be applied to other lines of insurance including life and health.
[5] According to Wikipedia, "Telematics" is an interdisciplinary field that encompasses telecommunications, vehicular technologies, road transportation, road safety, electrical engineering (sensors, instrumentation, wireless communications, etc.), and computer science (multimedia, Internet, etc.).
[6] American Academy of Actuaries Committee on Risk Classification. Risk Classification Statement of Principles. 2014
[7] Walters, Michael, Risk Classification Standards, 1981

introduction of telematics data on driving behaviors, actuaries can now, in a way that was impossible previously, transform ratemaking to utilize information that directly impacts risk. Behavioral ratemaking adjusts premium based on controllable driving behavior immediately. Behavioral ratemaking recognizes behavioral influence on the accident likelihood, and the potential severity of the accident, at each moment of actual driving. The overall number of claims would not change - except for the significant impact this measurement should have on actual behavior.

There are many ways to implement these rate adjustments – each with practical issues to be resolved. In any case, they would be based on behaviors in real time. This is not the same as using historical behaviors of the driver to adjust the rate. Behavioral ratemaking provides the insured with immediate premium savings for continuous behavioral improvements.

HOW IS TECHNOLOGY EXPECTED TO TRANSFORM INSURANCE?

With advances in technology, futurists project many industries to be disrupted by innovation. Insurance is no different. Insurtech refers to the use of technology innovations designed to squeeze out savings and efficiency from the current insurance industry model, including using new streams of data from internet-enabled devices to dynamically price premiums according to observed behavior.[8] It has been over a decade since the invention of a telematics device to provide real time input to insurers.[9] Insurtech ideas potentially impacting ratemaking include: increased use of predictive modelling, using telematics or internet data to create improved ways to classify drivers, and mileage based insurance. When the Insurtech sector first developed, many in Insurtech with little insurance expertise believed that new technologies would be able to quickly disrupt the industry and allow for new companies to quickly begin taking significant market share from the established ones.

Such disruption in the insurance market has not transpired. Currently, experts in Insurtech generally agree that there is no standout disruptive technology that will significantly impact market shares of the largest insurers any time soon and many insurtech startups still require help from the major insurers.[10] Industry executives have proclaimed that there is no technology on the horizon that will cause major disruptions in insurance company market shares in the near term.

Behavioral Ratemaking using real time telematics data will change this though. With the increased use of artificial intelligence, smart cars and driving algorithms, insurance ratemaking will need to keep up. Despite the slow start, it is clear that as technology advances new ideas are needed to align insurance

[8] https://www.investopedia.com/terms/i/insurtech.asp
[9] http://www.freepatentsonline.com/6931309.html?highlight=6064970. United States Patent 6931309. Motor vehicle operating data collection and analysis. 2004. Innosurance, Inc.
[10] https://www.investopedia.com/terms/i/insurtech.asp

better with the future of transportation and regulation.

WHY IS BEHAVIORAL RATEMAKING BENEFICIAL?

Behavior ratemaking has many benefits. Benefits to customers include immediate financial rewards for driving safer; provides proven methods to drive safer; and allows individuals and fleet managers to better manage driving risk.

Benefits to insurers using behavioral ratemaking include improved ratemaking which ties premiums charged to actual behavior and risk associated with that behavior. Higher identified risks are charged more, thereby generating increased revenue for high risk behaviors. There will be reduced insurer losses to the extent safer driving practices caused by the application of behavioral rating process are implemented. This leads to more accurate pricing as customers pay an amount more closely aligned to driving risk.

Behavioral ratemaking benefits to society include reduced accident frequency and severity to the extent some drivers adopt safer behaviors. Data collected over time showing how compliance with the posted speed limits impacts losses will have the potential to assist with better, safer programming of self-driving cars.

HOW IS BEHAVIORAL RATEMAKING DIFFERENT FROM PREDICTIVE MODELLING?

It is well known in statistics that correlation does not imply causation. It is preferable if rating variables are based on characteristics that are causal in nature.[11] Predictive modelling relies on finding attributes that are correlated with accidents to make predictions, while behavioral ratemaking relies on attributes that have been shown to cause or increase severity of accidents. Many companies, old and new, use predictive modelling to find better and more complex rating variables and classification systems that improve actuarial soundness. Predictive modelling is similar to traditional ratemaking in that historical information is relied upon to determine current rates. While this does lead to lower rates for safer drivers, the process takes time to design new pricing mechanisms and prove they work better. With predictive modelling safer insureds are asked to trust the insurer that they will eventually be charged lower premium for their safer driving.

There is a necessary delay between when the insurer confirms the safe driving and can reduce premium for the insured. Also, it is not necessarily intuitive which new rating variables or classification systems correlate with lower future costs, so it would be too risky for an insurance company to implement

[11] Modlin and Woerner. Basic Ratemaking, Fifth Edition. 2016, p. 157

changes based on predictive modeling in conjunction with telematics data without adequate proof that the new rates are better. Combined with a pre-existing distrust of insurance companies, this delay in recognition of premium savings resulting from safer driving reduces the ability of predictive modelling based safe driving incentives to take hold. These companies hope that safer drivers will have enough confidence in the possibility of future lower safe driver rates to choose the company before the new rates are fully implemented.

Also, without clear correlations, predictive modeling alone may not find opportunities to improve ratemaking as quickly as with the addition of behavioral ratemaking. This can be shown in the following simplified example with realistic assumptions. Let's assume older drivers are more risky than younger drivers and that older drivers tend to drive slower than younger drivers. In this example, slow driving would be correlated with higher risk when we look at the population as a whole. However, if we look at either subgroup individually, we will likely find that slower driving is actually correlated with lower risk. And for any individual in either group, risk can be reduced by driving slower. This is the most important aspect that represents behavioral ratemaking's untapped potential to improve fairness.

Behavioral ratemaking is different in that drivers see immediate financial rewards for safe driving behavior, in addition to additional benefits for continued improvement in driving behavior. Behavioral ratemaking uses telematics data to make intuitive adjustments to traditional ratemaking techniques. Speeding is but one example of a behavioral characteristic which may impact safety. For example, a company would implement a large discount for drivers who agree to abide by the speed limit. In addition to driving speed, the company would rely on telematics mapping data for location of insured vehicles and corresponding speed limit. A surcharge would be assessed on each mile driven at a certain number of miles per hour over the posted speed limit. An additional discount can also be immediately provided for driving within a certain range of the speed limit. Important assumptions are that safer drivers will be drawn to a rating system that rewards them for safer driving and that they will drive more safely when rewarded. Since the starting point is traditional rates and rating plans, the use of new intuitive rating variables will improve upon overall actuarial soundness. Traditional ratemaking techniques can then be used to adjust rates and adjustments as new data comes in for the population as a whole.

WHY WILL BEHAVIORAL PRICING BE DISRUPTIVE?

Once behavioral pricing takes off (with even a subset of insurance companies) adverse selection may create difficulties for the remaining more traditional insurance companies to co-exist without behavioral pricing. The effect could be similar to the introduction of nonsmoker/smoker pricing in

the life insurance market. Once nonsmoker discounts were introduced by one company, they, practically, needed to be introduced by all for similar reasons. As safer drivers self-select discounts for their own safer driving, insurers using traditional pricing exclusively will be left with less safe drivers, and higher accident frequency and claims costs. Drivers who do not modify driving behavior will self-select the increasing costs of traditional insurance. There is also less risk to insurance companies using behavioral pricing because riskier driving behavior will result in immediate rate surcharges and therefore, increased revenue.

Regulations have always required fair rates by disallowing unfair discrimination. Regulators rely upon actuaries to certify that rates are not unfairly discriminatory. The rating systems that developed in the twentieth century, based primarily on uncontrollable factors such as location, age, gender and marital status, were the fairest possible at the time. Once regulators and actuaries become comfortable with rating factors more directly linked with hazard, it will become apparent that traditional rating plans alone unfairly discriminate against safe drivers.

It is important to note that the business of insurance requires cross-subsidies. No rating mechanism can accurately predict the exact cost of each insured. Actuarially sound rating reduces cross-subsidies. There may also be an ethical limit as to how much cross-subsidies can be reduced. For example, in health insurance it is unacceptable to classify risks based on pre-existing conditions.

Changing driving behavior will be disruptive to more than just insurance. Americans spend billions of hours per year driving. As safety becomes more prominent in the mindset when getting behind the wheel, many other industries are potentially disrupted by this potential shift (such as automobile manufacturing, advertising, infrastructure design, law enforcement, etc.)[12]

HOW CAN BEHAVIORAL PRICING TAKE HOLD?

In order for ratemaking changes to take hold in the automobile insurance industry, there are three requirements which need to be addressed. These have not been adequately addressed by Insurtech thus far, which is the reason for the slow start to disruption.

1. From a customer perspective new changes need to be associated with an immediate monetary incentive. In other words, it needs to be cheaper for at least the safer half of drivers. Otherwise, customers will not move to the new system in a large scale. Would Uber have been able to overcome regulatory challenges if it weren't cheaper than traditional taxis?
2. From an actuarial perspective, telematics confirmation will be needed for the assumption that customers who do sign on will exhibit safer driving behavior. The safer behavior will be due

[12] Sharpin, Banerjee, Adriazola and Welle. The Need for (Safe) Speed: 4 Surprising Ways Slower Driving Creates Better Cities, May 09, 2017, https://www.wri.org/blog/2017/05/need-safe-speed-4-surprising-ways-slower-driving-creates-better-cities

to both attracting safer drivers to begin with, and all drivers driving more safely after they sign up.

3. Investors in behavioral pricing need assurances that customers signing on will have lower loss costs and that rate adjustments can be quickly implemented.

Insurance in the US is regulated on a state by state basis. While statutory guidelines for rates are similar among states, each state is responsible for determining and enforcing what is acceptable for its own residents. Behavioral pricing should lead to rates that are more actuarially sound than traditional rates. In order for behavioral pricing to take hold, insurance companies wishing to spearhead implementation would need to collaborate with individual state regulators. Three advantages to behavioral pricing over traditional pricing that should be important to regulators are: the incentivizing of safety, reduced likelihood of unfair discrimination, and more accurate rating.

IMPROVING SAFETY

One way to enable meeting all three of the aforementioned requirements is to identify, encourage and reward safe behavior. Doing so will reduce rates for policyholders while maintaining or improving profitability for insurance investors and actuarial soundness of rates.

Consideration of insured behavior with respect to safety is an important component of actuarial fairness that has not been adequately addressed in actuarial literature. Although the insurance industry has done much to improve safety in many lines of insurance, safety is not necessarily viewed as having a good financial impact for the insurance industry, either as a whole or by large insurance companies. "You want safer cars. Safer cars mean lower insurance. Safer driving means lower insurance costs", said Warren Buffet[13] making this counterintuitive point. Regulators require actuarial determination that rates are actuarially sound. Actuarial soundness means that the rate is just enough to provide for all costs in the aggregate. Therefore, safer driving should mean lower revenues for the insurance industry as a whole.

Large insurance companies project revenue by considering their own shares of insurance market segments. Therefore, a disruptive drop in revenue for the industry, whether due to safety or anything else, represents a risk to a large insurance company's revenue. Although safety reduces costs for insurance companies, the actuarial soundness requirement for rates implies no long-lasting loss ratio improvements due to decreases in losses. Many large insurance companies had their roots as small insurance companies that were able address to an underserved and safer subset of the market. An example in the life insurance industry is The Phoenix Companies, which began as American Temperance Life Insurance in 1851 and insured only those who abstained in alcohol.[14] An example

[13] in an interview with Yahoo Finance on May 2, 2018
[14] From Wikipedia entry for The Phoenix Companies

in the property casualty insurance industry is GuideOne, which began in 1951 as Preferred Risk Mutual Insurance Company, with the idea that non-drinkers would be in fewer accidents than those that did drink.[15] As in the past, the opportunity presents itself today for a startup or small insurer to focus on safer than average individuals. By using behavioral ratemaking, this company would also create incentives for insureds to become safer.

With respect to improved safety, the insurance industry currently seems to be primarily concerned about the impact of driverless cars. However, there is little evidence that driverless cars will be safer than human drivers in the near term.[16] In addition, the focus on the safety of driverless cars removes energy from how safety can be improved through safer human driving behavior.

THE DIFFICULTY OF RELEARNING SAFER DRIVING BEHAVIORS

It is very difficult for an individual to relearn safe behavior.[17] We cannot let that individual difficulty blind us to the safer possibilities for society as a whole. It may be easier to for some individuals to overcome opioid addiction than for others to correct some unsafe driving habits. Even if that is the case, society as a whole can improve safety. For example, cigarette smoking has decreased dramatically over the last fifty years. While it is very difficult for an individual to quit smoking, it was possible for smoking to be reduced in society overall. Similar driving specific examples of safety improvements that are difficult for the individual but possible for society as a whole are the increase in seat belt usage and the decrease in drunk driving over the past few decades.

Seat belt use is a safe driving behavior that reduces mortality and injury severity after an accident.[18] Therefore, seat belt usage reduces insurance losses. It has been widely observed that seat beat use has greatly increased over thirty years. A widespread survey, taken in 19 cities in 1982, observed 11 percent seat belt use for drivers and front-seat passengers.[19] In 2009, seat belt use averaged 88 percent in the 30 States with primary seat belt laws.[20] Though not exactly apples-to-apples, this represents an eight-fold increase, showing that the vast majority of drivers were ready, willing and able to take on this safer driving behavior. While driving behavior can be very difficult for an individual to change, this example provides evidence that the driving public is able to adopt additional safe driving behaviors.

---

[15] From Wikipedia entry for GuideOne Insurance

[16] Gosch, Susanna. Connect Differently: The Evolution of Automobile Technology and the Impact to Insurance. NAMIC presentation, 2018.

[17] James, Leon. University of Hawaii Student Reports on Driving Personality Makeovers.

[18] Cummins, Koval, Cantu, Spratt. 2011. Do seat belts and air bags reduce mortality and injury severity after car accidents? https://www.ncbi.nlm.nih.gov/pubmed/21720604

[19] Williams and Wells, 2004. UNC Highway Safety Research Center, 2011, p. 2-4

[20] Chen, Y. Y. & Ye, T. J. (2010, May). Seat belt use in 2009 – use rates in the states and territories. Traffic Safety Facts. (Report No. DOT HS 811 324). Washington DC: National Highway Traffic Safety Administration. Available at http://wwwnrd.nhtsa.dot.gov/Pubs/811324.pdf.

Despite the empirical evidence that human driving behavior can become safer as a whole, it may still be difficult to envision improved safety on a wide scale due to improving human driving behavior alone. We do know change for safety is possible, and although it may be unprofitable for large insurance companies that maintain the status quo, it is profitable for a new model of insurance company. Improved safety is good for society.

INSURANCE PRICING WOULD INFLUENCE DRIVING BEHAVIOR

The question is not whether driving behavior can be improved, but whether insurance pricing can encourage safe behavior. In order for all the benefits of behavioral pricing to be realized, it must be true that some drivers can and will change their driving behaviors in response to their insurance price. In the past, common actuarial wisdom was that it was not possible for an insurance pricing system to encourage safe behavior as noted by Michael Walters, "Few drivers wear seat belts despite the life-saving evidence, so the prospect of saving a few dollars of insurance surcharge certainly will not induce a modification of driving behavior."[21] Coincidentally, not too long after that paper was written, most drivers began to consistently use seat belts. According to a Canadian survey, the majority of drivers believe doubling speeding fines would reduce speeding.[22] Immediate insurance surcharges that are directly attributable to speeding are very similar financially to fines. This supports that increasing insurance costs for speeding could reduce speeding.

The advent of telematics has enabled insurance pricing to induce the driving public to drive more safely. In 1981, there was no way to reliably determine whether drivers used seat belts or to monitor other driving behaviors, such as speed. This lack of reliable determination virtually eliminated the possibility of insurers reflecting driving behavior in pricing. Telematics data is now available so that the insurance company can determine driving behavior with great accuracy. Because of the availability of reliably correct telematics data, the behavioral price differences can be substantial. Behavioral pricing combined with the availability of telematics data can now provide the driver with minute by minute updates on insurance pricing as compared with the annual updates of the past. By providing continuous behavioral feedback impacting premium, drivers are enabled to consider premium when choosing a driving behavior.

In order for insurance pricing to influence driving behavior, the pricing difference needs to be significant to the insured. While driving slower saves fuel costs, the resulting savings do not seem to be great enough to significantly influence driving speeds.

---

[21] Walters, Michael, Risk Classification Standards, 1981

[22] EKOS Research Associates. Driver attitude to speed and speed management: a quantitative and qualitative study — final report. Transport Canada, Report No. TP 14756 E (2005)https://www.tc.gc.ca/media/documents/roadsafety/TP14756E.pdf

In order to show that insurance pricing can encourage safe behavior, it is noted that a large part of driving risk is during the daily commute to work. For many people, there are many commuting cost options, including fuel efficiency, parking and use of public transportation. A daily difference in insurance cost would likely impact commuting cost benefit analysis and influence driving behavior to recognize a reduced insurance cost each day.

REVIEW OF SPEED AND OTHER TELEMATICS ATTRIBUTES

"Newtonian relationships between the fourth power of small increases or reductions in speed and large increases or reductions in deaths state the case for speed control."[23] The best choice of driving attribute to be used for behavioral ratemaking is speed. As opposed to other attributes, such as cornering, braking and acceleration, speed has several advantages including that it relates to the hazard. According to Walters, attributes "should reasonably relate to the potential for, or hazard of, loss."[24] Compared to the other attributes: speed is more commonly a direct cause of accidents[25] and speed is likely correlated with other aggressive and risky driving behaviors such as assuring safe following distance.[26] A slower driver would be less likely to be tempted into a risky maneuver to pass an even slower moving vehicle. Regardless of the cause of the accident, virtually every accident would have a reduced cost if the initial speed were reduced and a better (slower) speed score would always be associated with reduced hazard. Similarly, a worse (faster) speed score would almost always be associated with increased hazard. A better cornering score is not always correlated with decreased risk as crossing a yellow line at an intersection could improve the score but increase accident potential.

Some attributes for which it may seem reasonable to adjust the rate based on historical behaviors would not be feasible for behavioral ratemaking. While "hard braking" can be used as part of an overall analysis of safe driving, it does not directly relate to cost of risk. If a driver frequently brakes hard, the driver may be exhibiting unsafe behaviors prior to the hard-braking. While a hard-braking surcharge may reduce some unsafe behaviors, the hard-braking attribute does not work for behavioral ratemaking. The hard braking itself is used by the driver for the purpose of reducing hazard and it doesn't make sense to charge the driver for the hard braking in the seconds before an accident that reduced the cost, or to discourage the driver from hard braking to avoid an accident. Compared with good speeding scores, good braking scores are not as clearly associated with safe driving and can be associated increased accident probability. For example, rolling through rather than completely stopping at a stop sign could improve braking score while increasing the chance of an accident.

[23] Speed, road injury, and public health. Richter ED, Berman T, Friedman L, Ben-David G. Annu Rev Public Health. 2006;27:125-52. https://www.ncbi.nlm.nih.gov/pubmed/16533112
[24] Walters, Michael, Risk Classification Standards, 1981
[25] Quick Facts 2016, NHTSA, October 2017 (Updated February 2018)
[26] https://www.nhtsa.gov/risky-driving/speeding

Conversely, a bad braking score could be the result of successfully avoiding an accident or making a complete stop for a pedestrian in a crosswalk. Using a hard-braking attribute could increase risk if the braking surcharge discourages drivers from hard braking when necessary to avoid an accident. The braking attribute just does not make intuitive sense when used on a real time telematics data since hard braking may be the result of trying to avoid or reduce the cost of an accident. Also, it wouldn't make sense to charge a driver for braking hard one second before an accident. What would make sense is charging the driver for going too fast before the hard braking that led to the need for the hard-braking evasive action in the first place.

Speed meets another criteria better than other attributes such as braking or cornering: it is easier to measure. The attribute "should be susceptible to measurement by actual experience data."[27] Drivers already understand that speed relates to risk and are trained to objectively measure speed. The other attributes would require additional training to show drivers how behavior impacts their score.

Other groups concerned with safety, such as law enforcement and the medical community, have determined that slower speeds are safer. There has been no such determination for cornering or braking. The public already understands that speeding causes insurance losses. According to a Canadian study, about ninety percent of drivers believe driving over the speed limit increases the chance of accidents, injuries and getting killed.[28] While there are certainly other behavioral factors which may impact accident risk, the insurance industry should focus on speed as the first attribute to use with behavioral ratemaking.

Data shows that speed increases costs of risk

Since the beginning of the automotive age, it has been known that increasing speed increases the cost of driving risks. According to NHTSA, "For more than two decades, speeding has been involved in approximately one-third of all motor vehicle fatalities."[29] According to the NHTSA and NTSB, speeding causes as many deaths as drunk driving.[30] Considering this statistic only includes accidents where speed was actually recorded as the cause, speeding fatalities may be understated. Other accidents where the initial speed exceeded the speed limit are not included. There is no way to determine how many fatalities in these accidents could have been avoided had the initial speed not been excessive.

[27] Walters, Michael, Risk Classification Standards, 1981
[28] EKOS Research Associates. Driver attitude to speed and speed management: a quantitative and qualitative study — final report. Transport Canada, Report No. TP 14756 E (2005)https://www.tc.gc.ca/media/documents/roadsafety/TP14756E.pdf
[29] https://www.nhtsa.gov/risky-driving/speeding
[30] "Speeding kills about as many people each year as drunken driving, NTSB warns", USA Today, July 25, 2017

Slower speeds reduce accident probability

"At lower speeds, drivers have a wider field of vision and are more likely to notice other road-users."[31] Before an accident occurs, something unexpected must happen within the minimum distance (this could be defined as the distance travelled in two seconds, for example) needed by the driver to make normal driving adjustments in speed and direction. When this happens, the driver will undertake evasive action to reduce the probability of the accident and potential severity of the accident. The smaller this distance is, the less likely it is for an unexpected event to occur within that distance. If the initial speed is reduced, the minimum distance is proportionally smaller, so it is less likely for an event requiring evasive action to occur. Therefore, a decrease in initial speed decreases accident frequency at least proportionally.

According to Nilsson, speed has a greater than proportional impact on accident frequency. $A_2 = A_1\left(\frac{v_2}{v_1}\right)^2$ or, the number of injury accidents after the change in speed (A2) equals the number of accidents before the change (A1) multiplied by the new average speed (v2) divided by the former average speed (v1), raised to the square power.32

Slower speeds reduce accident severity

Since kinetic energy is proportional to the square of velocity, it can be hypothesized that the cost of damage caused by an accident is proportional to the square of speed at impact. This hypothesis is borne out by studies.[33] While ethical experimental confirmation of how bodily injury costs relate to speed of impact is not possible, it can also be hypothesized that bodily injury costs are also proportional to the square of the speed.

How reduced speed impacts expected cost of accidents

Since total costs are frequency times severity, an X% reduction in speed may cause approximately 2X% to 3X% reduction in accident costs. This calculation does not consider how other safe driving behaviors are likely correlated with slower driving, so more analysis is needed to conclude this relationship. While there is a range of driving speeds, it is not uncommon for the average speed on a highway segment to be 20% greater than the speed limit. In these cases, for example, a 20% reduction in speed could cause a 20% decrease in probability of an accident and a 36%[34] reduction in severity

[31] Sharpin, Banerjee, Adriazola and Welle. The Need for (Safe) Speed: 4 Surprising Ways Slower Driving Creates Better Cities, May 09, 2017, https://www.wri.org/blog/2017/05/need-safe-speed-4-surprising-ways-slower-driving-creates-better-cities

[32] Nilsson, G. (2004) Traffic safety dimensions and the power model to describe the effect of speed on safety. Bulletin 221, Lund Institute of Technology, Lund. https://ec.europa.eu/transport/road_safety/specialist/knowledge/speed/speed_is_a_central_issue_in_road_safety/speed_and_accident_risk_en

[33] Richards. 2010. Relationship between Speed and Risk of Fatal Injury: Pedestrians and Car Occupants

[34] 100% – 80% ^ 2

yielding a 49%[35] reduction in costs.

DRIVING ALGORITHMS: PROGRAMMING HUMANS VERSUS CARS

Programmers will need assistance from the actuarial profession to consider safety within the automated driving algorithm. It would be a mistake to assume that automated driving algorithms will reduce losses so significantly that actuarial pricing would not be needed. As with any new insurance product, actuaries need to understand it to price and underwrite the insurance accurately. Accurate insurance pricing will encourage safety in the design. Perhaps actuarial pricing programs can be written to apply self-driving algorithms in model driving situations to assess how well adapted it is to avoid and reduce severity of accidents.

"In the future, the actuary will be in the car."[36] With respect to driverless cars, programmers strive to create driving algorithms that are at least as safe as a human driver. Automated algorithms will certainly reduce some types of accidents such as distracted driving. As long as the driverless car is at least as safe as a human driver, implementation will improve safety. Currently, incentive and responsibility to significantly improve safety beyond human driving is lacking. There may be minimum requirements to obtain and possess an "automated" driving license, but the best incentive for programmers to produce safer algorithms would be to reduce insurance costs through behavioral ratemaking. With the incentive of saving on the costs of insurance risk it would be possible to experiment with possible behaviors to improve telematics attributes and safety.

Human drivers, too, are not primarily concerned with safety when deciding how they wish to drive. As with any automation, programmers should be expected to program automated vehicles to drive the way a human driver would drive. This is similar to an individual having the responsibility to decide how to drive. In either case, it is the responsibility of the insurance industry to determine how much to charge for insurance using the chosen driving behavior as an input. The difference with an automated driving algorithm is that there are explicit decisions with respect to risk and safety.

There are clearly cases where humans are better than automation. Humans appear to be better at determining where in the lane to drive[37] and better at driving in bad weather.[38] Futurists believe that the insurance rating formula should be determinable based on the algorithm and placed within the program to determine the insurance charges based on the algorithm and other factors such as time,

[35] 100% - (100% – 20%) x (100% – 36%)
[36] Quote from a programmer on the topic of how autonomous driving will impact insurance pricing at the second annual smart driving car summit in Princeton, NJ in May of 2018.
[37] Based on research by Insurance Institute for Highway Safety on vehicles with lane departure warning presented at CAS Crash Course in Vehicle Technology and Driverless Cars. JULY 19, 2018
[38] Bloomberg Business Week. Self-Driving Cars Can Handle Neither Rain nor Sleet nor Snow. September 17, 2018.

location and mileage of operation. In order to encourage safer and less risky driving algorithms, the insurance rating formula should consider driving behaviors of the algorithm. The programmers can then consider adjustments to the driving algorithm in consideration of the insurance costs.

Individual human drivers also have driving algorithms. Their driving behaviors could theoretically be reduced to a set of procedures to apply in all situations. Unlike automated driving algorithms, human driving algorithms are unknowable. While human driving algorithms may be able to be closely approximated based on observed driving behaviors in a great number (probably billions of miles would be needed) of situations, they cannot be used directly to determine insurance costs. Due to this complexity, the actuarial field may be a long way off from being able to create an insurance pricing formula based on an automated driving algorithm, but in the meantime behavioral ratemaking is the bridge to getting to that point. In addition to using behavioral ratemaking for human drivers, it can also be used for automated vehicles as they become more mainstream. Either way, behavioral ratemaking differentiates among various driving behaviors and safety characteristics. Actuarial expertise is needed now to connect driving behaviors with risk and in even more so in the future.

While many seem to have an initial expectation that automated driving may reduce insurance losses to near zero, automated driving will have losses for the foreseeable future. It may be many decades before fully automated vehicles are on the road.[39] In the meantime, there needs to be responsibility for understanding the risk consequences. Actuaries are the best profession to ensure that the automated driving algorithms of the future adequately consider insurance risks.

INFLUENCE ON TRAFFIC SAFETY AND LAW ENFORCEMENT

Since the beginning of the automotive age, society has created rules for the purpose of safety to reduce the risks of driving. These rules include obeying traffic signals, speed limits, stop signs, and lane markings. It is common knowledge that following driving rules reduces driving hazards. Traditionally, traffic enforcement has been an important means of improving traffic safety. Many studies have provided evidence of connections between the level of police enforcement and both driving behavior and the number of traffic accidents.[40] Since insurance companies are largely impacted by these financial costs, history shows insurers as being strong advocates of safe driving. Historically, insurance companies had no way to determine how well drivers mind driving rules. Other than consideration of traffic citations, there was no way to factor rule-following into the rating process. Most breaking of driving rules does not result in a traffic citation. Reliable determination of rule breaking is now possible

[39] https://www.nhtsa.gov/automated-vehicles/vision-safety. lynn.greenbauer.ctr@dot.gov (11 September 2017). "A Vision for Safety: Advancing Automation for Safer Roads ". NHTSA

[40] Stanojevic P, Jovanovic D, Lajunen T. Influence of traffic enforcement on the attitudes and behavior of drivers Accident; analysis and prevention. 2013. Mar;52:29–38.

with telematics data.

The general public has all seen drivers use devices to elude traffic cops such as radar detectors. In our society, many view traffic cops as bad and that speeding should be accepted and tolerated. An important role of government is to enable safe travel. The government sets driving rules such as speed limits and should enforce those rules. It is possible that behavioral ratemaking will be better at encouraging safe driving than traditional public services messages and law enforcement. Traffic regulators may need to work with actuaries and other experts in insurance risk to determine the best way to moderate insurance risk.

There are hundreds of thousands of traffic officers and other individuals dedicated to improving safety through speed limit enforcement in this country. There are hundreds of millions of drivers who seem to be more concerned about evading law enforcement than safety. There are only a few thousand actuaries who can determine how driving behaviors should be considered when addressing actuarial fairness to regulators.

HOW WILL BEHAVIORAL RATEMAKING ENABLE COMPANIES TO IMPROVE FLEET SAFETY?

Businesses that use highways have exposure to driving risks that need to be carefully managed. OSHA has published guidelines to help employers manage these risks.[41] According to the Royal Society for the Prevention of Accidents, "One of the most significant risks . . . is driving or riding at inappropriate speeds on work-related journeys."[42] Because driving behavior is difficult to change for any driver, attempting to manage another driver's behavior is difficult and could be offensive. We may have no choice but to trust the driver to be safe. As an example, plan to politely ask your next cab driver to drive within the speed limit. While this would be a perfectly reasonable request to manage our own risk of bodily harm, you may find it to be a difficult discussion. Commercial vehicles taking various levels of risk can be frequently observed. This risk directly translates to financial risk of the drivers' employers. In the past, many employers had limited ability to address this risk until the driver was involved in an accident and then, the only recourse may have been termination of the driver. Drivers spent their workday out of sight of their employer, and, for example, there may be a temptation to attend to non-work-related matters and to catch up on their deliveries by speeding.

Telematics is now increasing the ability of fleets to manage driving behavior. As there are many business reasons other than insurance cost (better service to customers, risk to reputation, etc.) to

[41] https://www.osha.gov/Publications/motor_vehicle_guide.html

[42] Driving for Work: Safer Speeds. https://www.rospa.com/rospaweb/docs/advice-services/road-safety/employers/work-safer-speed.pdf

reduce driving hazards, companies can use telematics to better manage driving risk. In addition, large self-insured companies can reduce insurance costs by making sure their drivers are driving safely.

For companies too small to self-insure, monetization of driving behavior improvement is extremely uncertain in timing and amount. Behavioral ratemaking can create immediate savings for smaller fleet managers if they encourage safe driving.

There is also the possibility that fleets that are successful in improving safety can bring other companies drivers or even individual drivers into their program to pass on insurance savings.

POSSIBLE METHODS TO INSTANTANEOUSLY ADJUST RATES

Throughout this paper we talk about instantaneously adjusting insurance rates. However, it is not entirely intuitive how this might take place since it has never been attempted with respect to US auto insurance which is highly regulated. There may be current laws or regulations in some states that would prohibit behavioral ratemaking, requiring changes to enable it. In other states, the introduction of a behavioral ratemaking might stimulate new laws and regulations to better control and regulate it. Similar with other uses of telematics data, may be privacy concerns.[43] This concern is reduced for behavioral ratemaking because many states already allow the use of telematics data for insurance pricing. Depending mainly on acceptability to regulators, and how to guarantee payment of surcharges, some possibilities include:

- Include surcharges as part of a normal rate filing. As a somewhat simplistic example, certain policies could have a $0.10 surcharge for every mile driven between 10 and 14 mph greater than the speed limit.
- For assessable mutual insurance policies, include surcharges as assessments.
- Create a relationship between the insured and a non-insurance company risk bearing entity that could change surcharges and take some financial responsibility for encourage safe driving behaviors. This concept would not be dissimilar to professional employer organizations taking some of the risk of their clients' workers compensation and employee health insurance benefits.

CONCLUSION

Speed has long been known to be one of the very most important driving safety factors and may be the best behavioral ratemaking risk factor. An insurance scheme with increased rewards for driving slower and more safely, that encourages implementation of safer driving practices, would be both beneficial and disruptive.

In the last few years, Insurtech has spawned many ideas to transform insurance. Although there are

[43] Insurance Journal. April 10, 2017. Driver Privacy at Risk when Telematics Data Stored in the Cloud: Researchers

many Insurtech initiatives to transform the auto insurance industry, most do not appear to be disruptive any time soon. This new approach to ratemaking, Behavioral ratemaking, is different and would be expected to cause disruption in the near term. The disruptions would be to not only the auto insurance industry, but the impact would also affect traffic enforcement policies, road infrastructure and car programming. Behavioral ratemaking will encourage safer driving and ultimately lead to safer roads.

Behavioral ratemaking is intended to put the driving population on the path to continuous and conscious relearning of safer driving skills. Complete transformation could be a long and difficult process, but significant benefits would be expected almost immediately. Regardless of whether transformation of driving occurs, behavioral ratemaking is an opportunity to create a successful insurance enterprise built upon safety conscious drivers. Behavioral ratemaking will also assist fleet management.

To move ahead with implementation, the industry needs to understand what is needed for an Insurtech idea to transform ratemaking and how safety can be aligned with insurance company financial goals. When insureds are encouraged to behave more safely, with improved behavior confirmed through telematics data, this transformation will benefit society. Examples show that insurance pricing can impact behavior. Actuarial ratemaking needs to be considered as part of automated driving algorithm creation processes.

In order to implement behavioral ratemaking, a new method to modify insurance premium instantaneously for driving behaviors must receive regulatory acceptance. Many insurance professionals witness the gory details of death and serious injury every day. Although their witness may only be through insurance claim files, it is otherwise similar to first responders and medical personnel. Spirits speak from the grave to focus on safety to give meaning to unnecessary deaths.

Applying Maximum Entropy Distributions To Determine Actuarial Models

Jonathan Evans

Abstract

Maximum information entropy distributions are a powerful and versatile tool for determining actuarial models from limited information, not necessarily from sample data, while not introducing unnecessary assumptions. These distributions, and potential applications, were presented in a 1967 paper in the Proceedings of the Casualty Actuarial Society, with a discussion following in 1968, but afterward effectively forgotten by the CAS community. The abandonment was likely primarily due to limited computational resources at the time. A relationship between maximum entropy and maximum likelihood is explained, along with an invariance property of the maximum entropy distributions under certain coordinate transforms of random variables. Applications of maximum entropy distributions to determine actuarial models for several practical problems are demonstrated. Some examples demonstrated include determining distributions consistent with the California Workers Compensation Rating Bureau's Tables M and L, and LER tables, determining distributional information sufficient for Bayesian or Credibility calculations, multivariate predictive models naturally adapted to special constraints and automatically including credibility adjustments that are difficult to incorporate in GLMs.

Keywords: Bayesian Models, Credibility, Information Entropy, Loss Models, Maximum Entropy Distributions, Maximum Likelihood Estimation, Predictive Models, Generalized Linear Models

1. BACKGROUND INTRODUCTION

Information entropy, a central part of Information Theory introduced in 1948 by Claude Shannon ([16]), is a scalar measure of the uncertainty, or lack of information, in a probability distribution. The entropy of a Uniform Distribution on a finite set of points increases with the number of points. A deterministic 100% probability for a single point has the lowest entropy of any distribution on any discrete set of points, finite or infinite. It also happens to be that the entropy of a Normal Distribution increases with its standard deviation and is independent of its mean.

Maximum information entropy distributions are a powerful and versatile tool for determining actuarial models, particularly with respect to the objective of parsimony, when information is limited. In cases where the specified constraints, implied by what information is available, are not sufficient to uniquely determine a probability distribution, entropy maximization can often be used to determine a distribution that satisfies the constraints, but otherwise assumes the least additional information. It is important to bear in mind that these constraints do not have to derive from a sample of data observations. They may come from any source of information, such as expert opinion, knowledge about the underlying data generating mechanism, generic assumptions, etc. Even if the constraints are derived from a data sample, they are sufficient statistics for a maximum entropy distribution, and it is not necessary to have the details of the sample itself. It is also not necessary to specify any

underlying statistical framework, Frequentist or Bayesian, of hypothesized models. Maximum entropy distributions may have significant predictive value, but there is no intrinsic need for prediction performance fitting or testing procedure.

Example 1.1 Selecting The Highest Entropy Model That Satisfies A Basic Constraint

Losses are known to be non-negative with mean 10,000 but no other information is known about the distribution of losses and no sample data is given. There are many distributions satisfying these constraints, including:

Some Distributions Meeting Constraints	Density Function	Information Entropy
Wide Uniform	$\frac{1}{20000} \quad x \in [0, 20000]$	9.90349
Narrow Uniform	$\frac{1}{2000} \quad x \in [9000, 11000]$	7.6009
Lognormal ($\sigma = 1$)	$\frac{exp\left(-\left(\frac{1}{2}\right)\left(\frac{1}{2} - log(10000) + \log(x)\right)^2\right)}{x\sqrt{2\pi}}$	10.1293
Exponential	$0.0001\ exp(-0.0001\ x)$	10.2103
Pareto (min loss =100)	$\frac{10000 \times 10^{2/99}}{99} x^{-199/99}$	6.58512

It makes sense to use the distribution with the highest entropy, in this case the Exponential Distribution, to minimize unnecessary implicit assumptions. In fact, the Exponential Distribution has

the maximum entropy of any distribution that fits the given constraints. Note, maximum entropy as used here for model selection is distinct from information criteria such as AIC, BIC, etc. (see Appendix B), which could not be used in this situation because no sample data is given.

Maximum information entropy distributions, and potential applications, were presented in a 1967 paper, with the very appropriate name "A Discipline for the Avoidance of Unnecessary Assumptions," in the Proceedings of the Casualty Actuarial Society (PCAS), (Roberts [14] with a discussion following in 1968 by Hurley [9]). The Roberts [14] and Hurley [9] papers are excellent references, and readers are strongly encouraged to become familiar with these papers as background for understanding this paper, including a more thorough treatment of the meaning of entropy. At the time of this writing, the Wikipedia article "Maximum Entropy Probability Distribution," (see [16]) is also a very useful additional source of information. In this paper we focus on areas of actuarial application, with many examples, along with some important general properties of maximum entropy distributions.

In the half century following [14] and [9], knowledge of maximum entropy distributions was effectively forgotten by the CAS community. Some of the very rare exceptions with some mention of information entropy were: the use of entropy for constructing automobile rating territories (see Conger [7]) in 1987, a proposed unified approach to pricing risk (see Kull [10]) in 2003, an application to jump diffusion processes in 2013 (see McKean [13]), and cross entropy applied with machine learning (see Chalk and McMurtrie [6]) in 2016. An interesting non-actuarial application of maximum entropy to financial risk management in 2015 is Geman, Geman, and Taleb [8], showing what effect that constraints on the probability of ruin and the expected shortfall conditional on ruin will have on the returns of an investment portfolio.

The general abandonment of maximum entropy applications by the CAS was most likely a consequence of limitations in computing power available to actuaries in the 1960s when [14] and [9] appeared in the PCAS and in subsequent decades. A secondary reason may have been the focus on directly data driven and less computationally intense statistical methods, such as Generalized Linear Models (GLMs) and Credibility Models. Even in the absence of computing power, it is unfortunate that CAS actuaries have not been generally aware of the maximum entropy derivation of common distributional forms. Most of these common forms (Exponential, Normal, Lognormal, Gamma, etc.) correspond to maximum entropy distributions given certain constraints (see table of common distributions by maximum entropy constraints in [17]). This knowledge can be useful in selecting which common forms to apply in a given application. For example, if the constraints in Example 1.1 had been stated in terms of the first two moments of the logarithm of the loss, instead of the first

moment of the loss itself, a fitted Lognormal Distribution would have had the highest entropy.

This paper demonstrates the application of maximum entropy distributions to determine actuarial models in types of problems that often occur in actuarial practice. Section 2 is a basic introduction to the mathematical definition of information entropy. Section 3 covers the general format for the distributional density and generalized moment constraint equations that can be solved to determine a maximum entropy distribution given a particular limited amount of information. Section 4 demonstrates that fitting maximum entropy distributions in the format shown in Section 3 is equivalent to solving for maximum likelihood, for the special case when a sample of observations is given whose sample generalized moments have the same values as those specified in the constraints from the Section 3 format. Section 5 shows that maximizing entropy before or after certain coordinate transformations of a random variable are applied, as long as the constraint equations are consistently transformed. Note, the material in Sections 4 and 5 is well known outside of the actuarial community. It is presented here for the benefit of the actuarial readership and not claimed as original results. Section 6 consists of further useful examples for problems that are common in actuarial practice. Section 7 introduces a general framework for applying maximum entropy distributions to determine multivariate predictive models that can also naturally include special constraints and/or Bayesian/Credibility type adjustments that are difficulty to include in Generalized Linear Models (GLMs). Appendix A contains a brief discussion of computational and software coding challenges. Appendix B clarifies of some confusion reviewers of an earlier draft of this paper had, identifying maximum entropy distributions with several other very different things. Importantly, applying maximum entropy distributions is quite distinct from applying the information criteria (AIC, BIC, etc.) that were first introduced in the 1970s (see [1], [2], [3], [4], and [15]) for selection between different hypothesized models given sample data. It is also distinct from the ordinary technique of matching moments or statistical techniques where use of the exponential family of distributions is central, such as Generalized Linear Models (GLMs) and exact credibility.

2. INFORMATION ENTROPY

The *information entropy* of a probability distribution is a measure of the extent of lack of information. For probabilities p_i on a finite set of points $\{x_1, \ldots, x_n\}$ the information entropy is defined as:

$$S = -\sum_{i=1}^{n} p_i \, log(p_i) \tag{1}$$

with the convention that if $p_i = 0$, then $p_i \, log(p_i) = \lim_{p \to 0^+} p \log(p) = 0$.

The maximum possible entropy for n points is $log(n)$ for the uniform distribution $p_i = \frac{1}{n}$, and the minimum possible entropy is 0 for a single point mass distribution $p_k = 1, \; p_{i \neq k} = 0$. Information gain is equivalent to the loss of entropy. A realized outcome in this example lowers the entropy (increases information) by $log(n)$. Changing from natural logarithmic base e to another logarithmic base $b > 1$ would only have the effect of multiplying S by $\log(b) > 0$, which would not change the relative order of different distributions as ranked by S. This definition for information entropy can easily be generalized, with some care to measure theoretic issues, to infinite discrete sets and continuous probability distributions, with the integral expression for S in the continuous setting:

$$S = -\int p(x) \, log\left(p(x)\right) dx \tag{2}$$

For continuous distributions the entropy *S* can be negative. Thus, entropy for continuous distributions, though serving the same purpose and generally having the same properties, is not directly comparable to entropy for discrete distributions. Sometimes entropy for continuous distributions is referred to by a term other than "Information Entropy," such as "Differential Entropy." We will mostly focus on continuous distributions and the integral expression (2), but results will typically also be valid in the discrete setting.

Example 2.1 Entropy of A Normal Distribution Is An Unbounded Increasing Function Of Standard Deviation Independent Of Mean

For a Normal Distribution the information entropy will only depend on the standard deviation parameter σ since S as defined in (2) is invariant under translation of x by an additive constant. The entropy is $S = \frac{1}{2}\log(2\pi e) + \log(\sigma)$ an increasing function of σ, such that $\lim_{\sigma \to 0} S = -\infty$ and $\lim_{\sigma \to \infty} S = +\infty$.

Example 2.2 Incomparability of Discrete And Continuous Entropy

A discrete single point mass at $x = 1$ has entropy 0, but a continuous uniform distribution for $x \in [0.9, 1.1]$ has a lower entropy of -1.60944, although it clearly contains less information than the point mass.

Information entropy is not necessarily correlative with typical concepts or measures of quantitative risk. It is sensitive to the distribution of probability among different possible outcomes, but it is insensitive to the relative magnitudes of these outcomes.

Example 2.3 Incommensurability Of Information Entropy And Quantitative Risk

A discrete uniform distribution on two possible outcomes has the same information entropy, $\log(2)$, whether the possible outcome set is $\{1, 2\}$ or $\{1, 10^9\}$. Similarly, the information entropy for a uniform continuous distribution on the interval $[0, 1]$ is the same value 0 as for a continuous uniform distribution on the union $[0, 0.5] \cup [10^9, 10^9 + 0.5]$ of two far apart intervals.

Although we will mostly present examples and properties for a scalar random variable, it is usually possible to extend the examples and properties that follow to vector, or multi-variate, random variables. Section 7 will focus on the multivariate context.

3. MAXIMUM ENTROPY PROBABILITY DISTRIBUTIONS

Suppose the following form for the density function, whether the random variable X is discrete or continuous, for measurable functions $\{g_1(x), \dots, g_m(x)\}$, which will be called *generalized moment functions*:

$$p(x) = exp\left(-a_0 - a_1 g_1(x) - \ \dots \ - a_m g_m(x)\right) \tag{3}$$

Define m *generalized moments* of this distribution as:

$$E[g_i(X)] = \int g_i(x)\, exp\left(-a_0 - a_1 g_1(x) - \ \dots \ - a_m g_m(x)\right) dx = c_i \tag{4}$$

It can be shown though Lagrange Multipliers (see [14] and further references listed there to papers by Jaynes and Tribus) that $p(x)$ has the highest entropy of any distribution having these specific values c_i for these m generalized moments. Often the given constraints on a probability distribution, in a specific application problem, can be expressed in the form equations $E[g_i(X)] = c_i$ together with the formal normalization constraint $E[1] = 1$, for a set of functions $g_i(x)$ and a set of constants c_i. Therefore, if it is possible to solve for the parameter values $\{a_0, a_1, \dots, a_m\}$ so that these constraints equations are satisfied by $p(x)$, then $p(x)$ has the highest entropy of any distribution satisfying these constraints.

Equation (3) is a very general form in that an arbitrary density may be stated in many ways that fit this form, although some or all of the parameters of the original density may be embedded in the functions $g_i(x)$ rather than corresponding to $\{a_0, a_1, \dots, a_m\}$. A trivial form for any density is:

$$p(x) = exp(-a_0 - a_1 \log(p(x)) \qquad a_0 = 0 \qquad a_1 = 1 \tag{5}$$

For convenience a_0 can be stated in terms of the normalization formula (6), leaving only the m other constraint equations that need to be solved such that (6) has a finite value so that the normalization constraint is automatically satisfied.

$$a_0(a_1, \dots, a_m) = Log\left(\int exp\left(-a_1 g_1(x) - \ \dots \ - a_m g_m(x)\right) dx\right) \tag{6}$$

Then the generalized moments can be expressed as:

$$E[g_i(x)] = -\frac{\partial a_0}{\partial a_i} \tag{7}$$

The following formula for the variance of the moment functions is also potentially useful:

$$Var[g_i(x)] = \frac{\partial^2 a_0}{\partial {a_i}^2} \tag{8}$$

Example 3.1 Maximum Entropy Distribution Determined with Very Limited Information About Losses

A reinsurer has only the following very limited information about the losses for individual claims but needs to completely determine the per claim loss distribution to price excess layers.

- 90% of claims are under 100,000
- The mean of the unlimited layer excess of 10 million is 1 million

Let the moment functions be:

$$g_1(x) = 1 \;\; if\; x \in [0, 10^5)$$

$$= 0 \;\; if\; x \in [10^5, +\infty)$$

$$g_2(x) = Max(0, x - 10^7)$$

Then the maximum entropy density form is:

$$p(x) = exp(-a_0 - a_1) \;\; if x \in [0, 10^5]$$

$$= \;\; exp(-a_0) \;\; if \;\; x \in (10^5, 10^7)$$

$$= exp(-a_0 - a_2(x - 10^7)) \;\; if x \in (10^7, +\infty)$$

The normalization parameter is:

$$a_0 = log\left(10^5 exp(-a_1) + (10^7 - 10^5) + \frac{1}{a_2}\right)$$

The constraint equations are:

$$\frac{10^5 exp(-a_1)}{10^5 exp(-a_1) + (10^7 - 10^5) + \frac{1}{a_2}} = 0.9$$

$$\frac{(1/a_2)^2}{10^5 exp(-a_1) + (10^7 - 10^5) + \frac{1}{a_2}} = 10^6$$

Numerical rooting finding leads to:

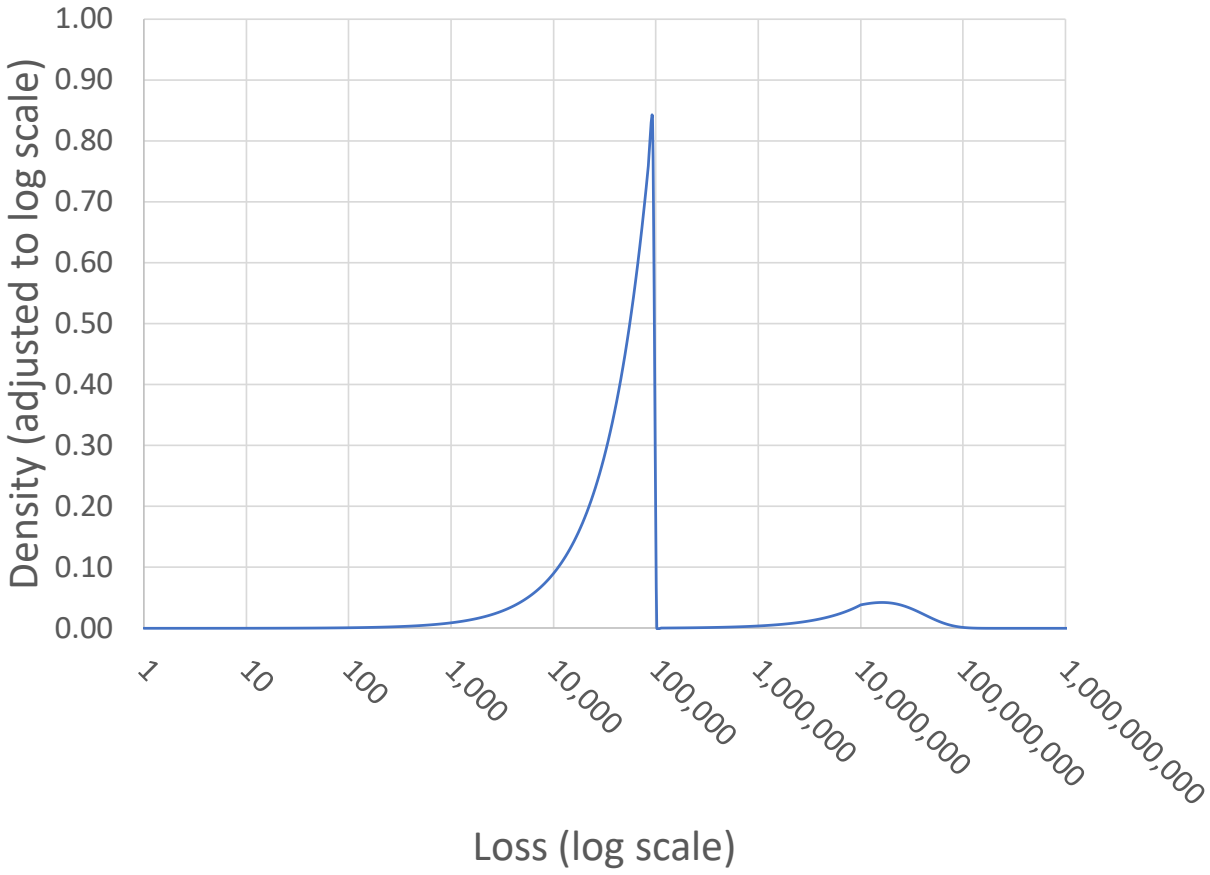
$$a_0 = 19.3776 \qquad a_1 = -7.75927 \qquad a_2 = 6.19750 \times 10^{-8}$$

Using $p(x)$, the ground up loss has mean 1.86 million and standard deviation 7.5 million, for a 404% coefficient of variation. The expected losses for excess layers of interest can be calculated:

Table 3.1 Some Layer Calculations For Maximum Entropy Solution In Example 3.1

Attachment	Limit	Expected Loss	Probability a Loss Hits Layer
0	100,000	55,000	100.0%
100,000	400,000	39,693	10.0%
500,000	500,000	48,752	9.8%
1,000,000	4,000,000	355,446	9.7%
5,000,000	5,000,000	357,887	8.1%
10,000,000	10,000,000	461,922	6.2%
20,000,000	30,000,000	454,253	3.3%
50,000,000	50,000,000	80,046	0.5%
100,000,000	Infinity	3,781	0.02%

Figure 3.1 Density Of Maximum Entropy Solution In Example 3.1



However, often even for mathematically consistent constraints there is no maximum entropy distribution.

Example 3.2 Some Constraints Where No Maximum Entropy Distribution Exists

A non-negative random variable has 90% probability of being less than 1000. These constraints are satisfied by the family of densities:

$$p(x) = 0.0009 \;\; if \;\; x \in [0, 1000)$$

$$= \;\; \frac{0.1}{L} \;\; if \;\; x \in [1000, 1000 + L)$$

$$= \;\; 0 \;\; if \;\; x \geq 1000 + L$$

A maximum entropy distribution cannot exist, because the entropy of a member of this family is an increasing function of L with no upper bound:

$$S(L) = \; -0.9 \log(0.0009) + 0.1 \log (10 \; L)$$

Example 3.3 Maximum Entropy Distribution for A Bounded Number Of Claim Counts

A certain type of insurance policy is limited to a maximum of 5 claims per year, and historically these policies have averaged 0.7 claims per year. The maximum entropy distribution for the annual number of claims can be found as:

$$a_0 = log\left(\sum_{k=0}^{5} \exp\,(-a_1 k)\right)$$

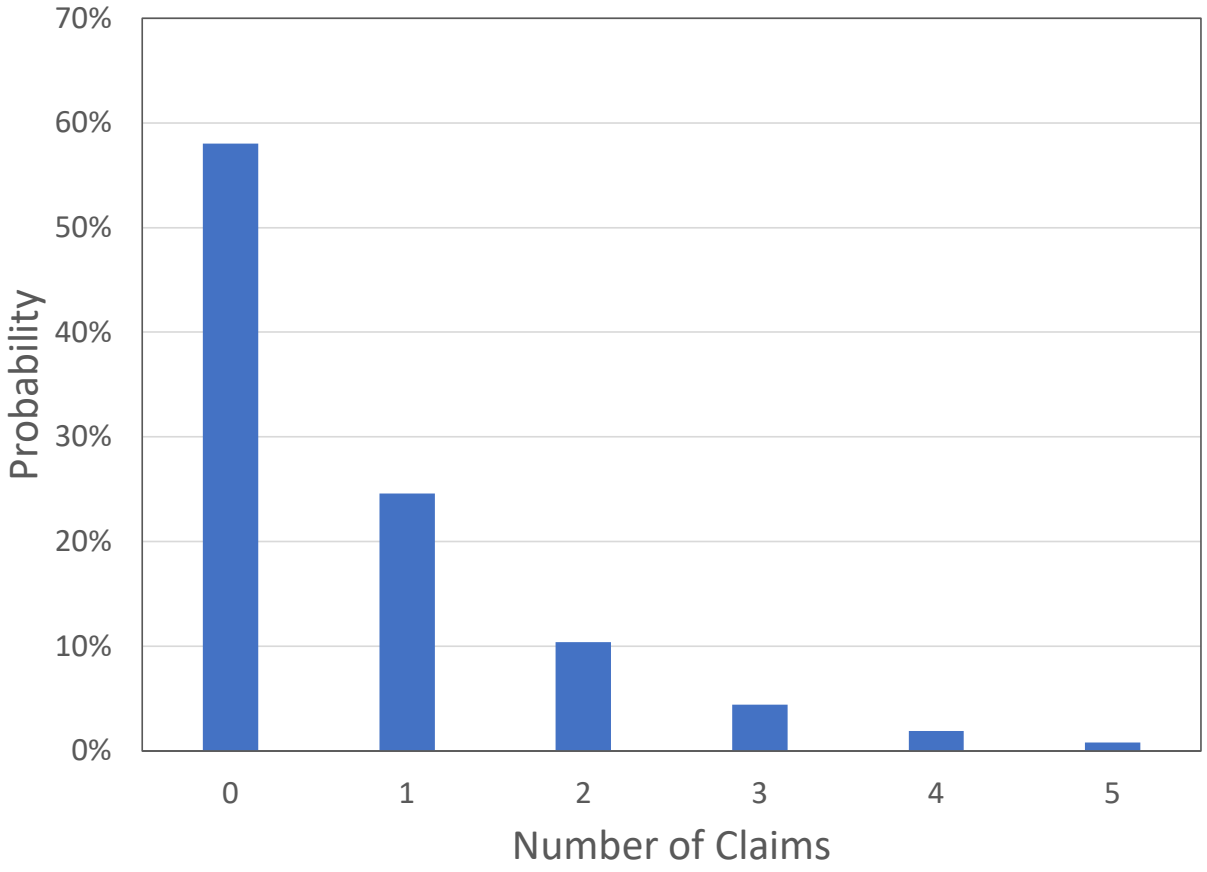
$$-\frac{\partial a_0}{\partial a_1} = \frac{\sum_{k=0}^{5} \mathrm{k} \exp\,(-a_1 k)}{\sum_{k=0}^{5} \exp\,(-a_1 k)} = 0.7$$

$$a_0 = 0.545133 \qquad\qquad a_1 = 0.859003$$

Table 3.2 Density Of Maximum Entropy Solution In Example 3.3

Number of Claims	Probability
0	58.0%
1	24.6%
2	10.4%
3	4.4%
4	1.9%
5	0.8%

Figure 3.2 Density of Maximum Entropy Solution In Example 3.3



4. A RELATIONSHIP BETWEEN MAXIMUM ENTROPY AND MAXIMUM LIKELIHOOD

The maximum entropy form (3) may be determined for the given constraints, without any sample of data. However, there is a general relationship between maximum likelihood estimation (MLE) for a density of the form (3) on a sample of observations $\{x_1, \dots, x_n\}$ and maximizing entropy such that the generalized moments (4) of the density are equal to the sample values $\frac{1}{n}\sum_{j=1,..,n} g_i(x_j)$. This makes sense as form (3) is a subset of the exponential family with the generalized moment functions $\{g_1(x), \dots, g_m(x)\}$ fitting in the position of the *sufficient statistics functions*. For a fixed parametric distributional form, such as form (3), the *sufficient statistics*, that is sample averages for these functions, include all information about determining the parameters that can be obtained from a given sample. Put another way, often an MLE fit is – even if unknowingly to the practitioner - a maximum entropy distribution for constraints based on sufficient statistics implicit in a distribution from the exponential family and their values when applied to the sample data.

Given a sample of observations $\{x_1, \dots, x_n\}$ and specific moment functions $\{g_1(x), \dots, g_m(x)\}$ the log-likelihood function for the distributional form given in (3) is:

$$\log\left(L(a_1, \dots, a_m)\right) = \sum_{j=1,..,n} \left(-a_0(a_1, \dots, a_m) - a_1 g_1(x_j) - \ \dots \ - a_m g_m(x_j)\right) \tag{9}$$

If $(a_1, \dots, a_m)^*$ is a maximum likelihood solution for (9) then:

$$\frac{\partial \log\left(L(a_1, \dots, a_m)\right)}{\partial a_i}\Big|_{(a_1,\dots,a_m)^*} = -n \left.\frac{\partial a_0}{\partial a_i}\right|_{(a_1,\dots,a_m)^*} - \sum_{j=1,..,n} g_i(x_j) = 0 \tag{10}$$

Consequently:

$$E[g_i(X)] = -\frac{\partial a_0}{\partial a_i}\Big|_{(a_1,\ldots,a_m)^*} = \frac{1}{n}\sum_{j=1,..,n} g_i(x_j) \tag{11}$$

So, in addition to $(a_1, \ldots, a_m)^*$ maximizing likelihood for the distributional form (3) given the sample observations, the resulting distribution is also the maximum entropy distribution constrained to have the same values for generalized moments $E[g_i(x)]$ as the sample averages for these generalized moments $\frac{1}{n}\sum_{j=1,..,n} g_i(x_j)$. That is to say that maximizing the likelihood for a distributional form like (3) on a sample, is the same as finding the maximum entropy distribution whose generalized moments corresponding to the functions $\{g_1(x), \ldots, g_m(x)\}$ are matched to the sample averages of the functions.

Alternately, if $(a_1, \ldots, a_m)^*$ satisfies (11), and hence (3) will be the density of the maximum entropy distribution for the constraints (11), then $(a_1, \ldots, a_m)^*$ will automatically be a critical point of the loglikelihood function in (9). The elements of the Hessian matrix of the loglikelihood in (9) can be shown to be:

$$H_{i,j} = -\frac{\partial^2 a_0}{\partial a_i \partial a_j} = -E\big[g_i(X)g_j(X)\big] + E[g_i(X)]E[g_j(X)] = -Cov[g_i(X)g_j(X)] \tag{12}$$

The determinant of the covariance matrix of a set of linearly independent random variables (none of which is a trivial point mass) will be positive since it is similar to the diagonal matrix of the variances. Consequently, the determinant of this Hessian must be negative for all points $(a_1, \ldots, a_m)$ that correspond to a legitimate density. So, the critical point is also a global maximum. (Note: If the

random variables $g_i(X)$ are linearly dependent then the original set of generalized moment functions $g_i(x)$, and their corresponding constraint equations, can be reduced through a linear transformation into a smaller linear independent set. If any of the $g_i(X)$ are point masses, these can be split out with their constraint equations automatically yielding point mass probabilities. Therefore, the original maximum entropy form and constraint equations can be restated to eliminate any linearly dependent and/or point mass generalized moments.)

Example 4.1 Maximizing Likelihood for a Normal Distribution Is Equivalent to Maximizing Entropy Given the Mean and Standard Deviation

A Normal Distribution with mean μ and standard deviation σ has density:

$$p(x) = \frac{exp\left(-\frac{(x-\mu)^2}{2\,\sigma^2}\right)}{\sigma\sqrt{2\pi}} = exp\left(-\left(\frac{\frac{\mu^2}{\sigma^2} + \log\left(\sigma\sqrt{2\pi}\right)}{2}\right) + \left(\frac{\mu}{\sigma^2}\right)x - \left(\frac{1}{2\sigma^2}\right)x^2\right)$$

This is clearly the maximum entropy form for $g_1(x) = x$ and $g_2(x) = x^2$ with moments $E[g_1(x)] = \mu$ and $E[g_2(x)] = \sigma^2 + \mu^2$. The maximum likelihood estimators for a sample $\{x_1, \ldots, x_n\}$ are given by the familiar formulas: $\hat{\mu} = \frac{1}{n}\sum_{i=1,\ldots,n} x_i$ and $\hat{\sigma}^2 = \left(\frac{1}{n}\sum_{i=1,\ldots,n} {x_i}^2\right) - \hat{\mu}^2$

When these estimators are used for the parameters, the moments of the distribution are set equal to the sample moments: $E[g_1(x)] = \frac{1}{n}\sum_{i=1,\ldots,n} x_i$ and $E[g_2(x)] = \frac{1}{n}\sum_{i=1,\ldots,n} {x_i}^2$ and this maximum likelihood solution for the Normal Distribution is also the maximum entropy distribution for a real valued random variable with these specified moments.

5. AN INVARIANCE PROPERTY OF MAXIMUM ENTROPY DISTRIBUTIONS UNDER CERTAIN COORDINATE TRANSFORMATIONS

Some coordinate transformations, that is certain smooth invertible functions of a continuous variable X, along with the correspondingly transformed generalized moment functions will result in the same maximum entropy distribution as if the maximum entropy distribution is determined before the coordinate transformation and then transformed. Note however, the value of the information entropy itself may change under these coordinate transformations.

Suppose $X = f(Y)$, where $f(Y)$ is differentiable and invertible. Then the equivalent transformed density of the maximum entropy form of $p(x)$ from (3) is:

$$q(y) = exp\big(-a_0 - a_1 g_1(f(y)) - \ \ldots \ - a_m g_m(f(y))\big)|f'(y)| \tag{13}$$

The transformed generalized moment equations (4) will be:

$$E[h_i(Y)] = c_i \qquad h_i(Y) = g_i(f(Y)) \tag{14}$$

These equations will still be satisfied because:

$$\int g_i(x)\ exp\big(-a_0 - a_1 g_1(x) - \ \ldots \ - a_m g_m(x)\big)\ dx =$$

$$= \int g_i(f(y))\ exp\big(-a_0 - a_1 g_1(f(y)) - \ \ldots \ - a_m g_m(f(y))\big)\ |f'(y)|\ dy \tag{15}$$

$$= \int h_i(y)\ exp\big(-a_0 - a_1 h_1(y) - \ \ldots \ - a_m h_m(y)\big)\ |f'(y)|\ dy$$

Furthermore, if $|f'(y)|$ can be expressed in the form:

$$|f'(y)| = exp\big(-b_0 - b_1 h_1(y) - \ \ldots \ - b_m h_m(y)\big) \tag{16}$$

then:

$$q(y) = exp\big(-(a_0 + b_0) - (a_1 + b_1)h_1(y) - \ \ldots \ - (a_m + b_m)h_m(y)\big) \tag{17}$$

is the maximum entropy distribution for the transformed constraints $E[h_i(y)] = E[g_i(f(x))] = c_i$. Therefore, as long as the generalized moment functions are consistently transformed, and $|f'(y)|$ can also be expressed in the standard maximum entropy form in the transformed space, it does not matter if the maximum entropy distribution is solved before or after the coordinate transform.

Example 5.1 Maximum Entropy Equivalence Between Normal Distribution And Lognormal Distribution

Suppose $X = \log(Y)$, the support of X is $(-\infty, +\infty)$, the support of Y is $(0, +\infty)$, and the given constraints are $E[X] = 0$ and $E[X^2] = 1$, then the maximum entropy distribution is the Normal Distribution with density:

$p(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$, which would transform to $q(y) = \frac{\exp(-log(y)^2/2)}{y\sqrt{2\pi}}$, the density of a Lognormal Distribution that is the maximum entropy distribution for the constraints $E[log(Y)] = 0$ and $E[log(Y)^2] = 1$.

Example 5.2 Counterexample - Maximum Entropy Non-Equivalence Under Transformation

If we repeat Example 5.1 using only the second constraint, $E[X^2] = 1$ then the maximum entropy distribution is still the Normal Distribution with density:

$p(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$, which would also still transform to $q(y) = \frac{\exp(-log(y)^2/2)}{y\sqrt{2\pi}}$, the density of a Lognormal Distribution that is the maximum entropy distribution for the constraints $E[log(Y)] = 0$ and $E[log(Y)^2] = 1$. However, the maximum entropy distribution under only the relevant

transformed constraint $E[log(Y)^2] = 1$ would be $r(y) = \sqrt{\frac{a}{\pi}} \exp(-a\ \log(y)^2 - a/4)$ with $a = \frac{1}{4}(1+\sqrt{5}) = 0.8090169943749475\ldots$. The first transformed restraint, which we discarded, is not met by $r(y)$, since under $r(y)$, $E[log(Y)] = 0.714863 \neq 0$. Also, $r(y)$ has entropy 1.79637, which is higher than the entropy 1.41894 of $q(y)$.

6. FURTHER EXAMPLES

6.1 Determining A Distribution Consistent With Excess Ratios In Tables M And L

The California Workers Compensation Insurance Rating Bureau (WCIRB) produces tables of per risk expected loss excess ratios ("insurance charges" in this context) by entry ratio (loss amount/mean loss), (see [5]). These tables are organized in columns corresponding to Expected Loss Groups (ELGs) that are ranges of expected loss per risk. The Table L varieties include adjustment for various per accident limits and Table M is unlimited.

Example 6.1.1 Excerpt from WCIRB's 2019 Table L

Below is an excerpt of values from WCIRB's 2019 Table L for loss limit $100,000 for ELG 50, corresponding to expected per risk loss in the range from $165,605 through $181,201. The Excel spreadsheet available online at [5] has many digits of precision, but often only 4-digit precision numbers are available in printed material.

Table 6.1.1 Sample from WCIRB's 2019 Table L for loss limit $100,000 for ELG 50

Entry Ratio	Rounded Excess Ratio	Unrounded Excess Ratio
0.00	1.0000	1.000000000000000
0.50	0.6719	0.671935231318322
1.00	0.5000	0.500000000000000
2.00	0.3938	0.393813297572041
5.00	0.3734	0.373364646661730
10.00	0.3695	0.369524681712078

A common actuarial problem is to determine the probability distribution underlying these tables for various practical applications. It can be very challenging to fit a typical functional form probability distribution, or even a mixture of typical forms, and such a fit may make unnecessary implicit assumptions. An alternative approach is to take the negative finite differences of the excess ratios to approximate the cumulative probability distribution, but this approach is very sensitive to numerical rounding errors and other aspects of discrete tabular representation. It often produces inconsistencies where the cumulative distribution decreases or remains unchanged as the entry ratio increases. However, there is a straightforward maximum entropy distribution for this situation.

Example 6.1.2 Maximum Entropy Distribution for WCIRB's 2019 Table L for loss limit $100,000 for ELG 50

From the Table L column underlying Example 6.1.1, we select for fitting purposes the following sample values, spaced out in terms of sequential differences in excess ratios, but including the highest available entry ratio of 10.00:

Table 6.1.2 Selected Values For Fitting From WCIRB's 2019 Table L For Loss Limit $100,000 For ELG 50

Entry Ratio	Excess Ratio
0.00	1.000000000
0.03	0.973293029
0.07	0.940656486
0.11	0.909512502
0.15	0.879861331
0.20	0.844934352
0.24	0.818432243
0.29	0.786763533
0.35	0.751193714
0.40	0.723319956
0.46	0.691745050
0.53	0.657705024
0.60	0.626617188
0.67	0.598501845
0.76	0.566402702
0.86	0.535579836
0.99	0.502275425
1.14	0.471953286
1.35	0.441000272
1.69	0.409781057
2.66	0.378248119
10.00	0.369524682

The generalized moment functions can be defined as:

$$g_i(x) = Max(0, x - x_i), \quad x_1 = 0.00, \ x_2 = 0.03, \ \ldots, \ x_{22} = 10.00$$

with density function:

$$p(x) = exp(-a_0 - a_1 x - a_2 Max(0, x - 0.03) - \cdots - a_{22} Max(0, x - 10.00))$$

and 23 constraint equations, including normalization, in integral form:

$$\int_0^\infty p(x)\, dx = 1$$

$$\int_0^\infty x\, p(x)\, dx = 1$$

$$\int_0^\infty Max(0, x - 0.03)\, p(x)\, dx = 0.973293029$$

……

$$\int_0^\infty Max(0, x - 10.00)\, p(x)\, dx = 0.369524682$$

The integrals can be broken down into piecewise calculations of means of exponential distributions over a sequence of intervals and simplified, although into very lengthy expressions in terms of exponential functions and algebraic operations. For example:

$$\int_0^\infty x\, p(x)\, dx = \exp(-a_0)\left(\frac{-a_1 0.03 \exp(-a_1 0.03) - \exp(-a_1 0.03) + 1}{{a_1}^2} + \cdots\right)$$

After a significant amount of calculus, numerical root finding can be applied to solve for the parameters. In practice, the author found it was necessary to do so sequentially. $\{a_0, a_1\}$ was solved first, while zeroing out $\{a_2, \ldots, a_{22}\}$ and ignoring the equations for $\{g_2(x), \ldots, g_{22}(x)\}$. Then, this solution was used as an initial search point to solve for $\{a_0, a_1, a_1\}$ while zeroing out $\{a_3, \ldots, a_{22}\}$ and ignoring the constraint equations for $\{g_3(x), \ldots, g_{22}(x)\}$. Continuing in the same manner eventually a solution for $\{a_0, \ldots, a_{22}\}$ under all the constraint equations was found:

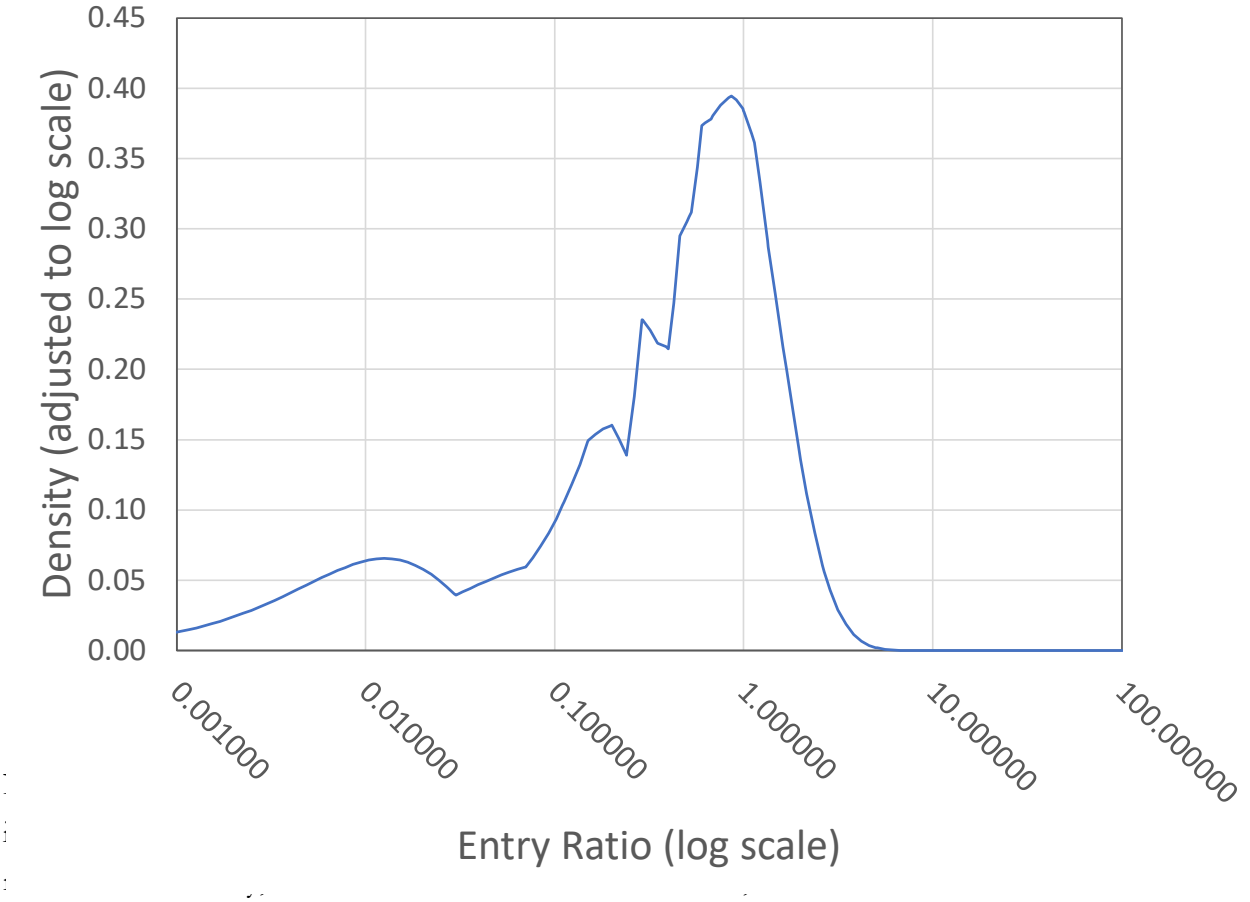
$a_0 =$ -2.64110659609105
$a_1 =$ 78.84798657065390
$a_2 =$ -67.84628757037710
$a_3 =$ -13.53961416054840
$a_4 =$ 1.05093302568349
$a_5 =$ 5.85233052253736
$a_6 =$ 3.77603379643773
$a_7 =$ -14.91579228797020
$a_8 =$ 11.13262310317780
$a_9 =$ -1.31419292343747
$a_{10} =$ -6.01333535467965
$a_{11} =$ 4.20493185505049
$a_{12} =$ -2.03880133396856
$a_{13} =$ 2.19601290431692
$a_{14} =$ -0.28563054188110
$a_{15} =$ -0.03229457248671
$a_{16} =$ 0.19073261292819
$a_{17} =$ 0.10266905342995
$a_{18} =$ 0.53399856270167
$a_{19} =$ -0.16195956428452
$a_{20} =$ 0.02984942588180
$a_{21} =$ -0.08113782038222
$a_{22} =$ -1.68856945802222

Here are some excess ratios and cumulative distribution values for the fitted entry ratios, the entry ratios from the original excerpt from Example 6.1.1, and some extrapolated entry ratios.

Table 6.1.3 Some Excess Ratios And Cumulative Distribution Values From The Maximum Entropy Solution In Example 6.1.2

Entry Ratio	Actual Excess	Fit Excess	Actual – Fit Excess	Fit Cumulative Probability
0.00	1.000000000	1.000000000	0.000000000	0.00000%
0.03	0.973293029	0.973293029	0.000000000	16.12129%
0.07	0.940656486	0.940656486	0.000000000	20.38433%
0.11	0.909512502	0.909512502	0.000000000	23.95617%
0.15	0.879861331	0.879861331	0.000000000	27.82634%
0.20	0.844934352	0.844934352	0.000000000	32.30302%
0.24	0.818432243	0.818432243	0.000000000	35.03824%
0.29	0.786763533	0.786763533	0.000000000	38.48090%
0.35	0.751193714	0.751193714	0.000000000	42.76647%
0.40	0.723319956	0.723319956	0.000000000	45.66500%
0.46	0.691745050	0.691745050	0.000000000	49.18993%
0.50	0.671935231	0.671925770	0.000009461	51.69306%
0.53	0.657705024	0.657705024	0.000000000	53.49086%
0.60	0.626617188	0.626617188	0.000000000	57.72643%
0.67	0.598501845	0.598501845	0.000000000	61.87659%
0.76	0.566402702	0.566402702	0.000000000	66.71170%
0.86	0.535579836	0.535579836	0.000000000	71.55581%
0.99	0.502275425	0.502275425	0.000000000	77.05588%
1.00	0.500000000	0.500000386	(0.000000386)	77.44246%
1.14	0.471953286	0.471953286	0.000000000	82.33417%
1.35	0.441000272	0.441000272	0.000000000	87.82233%
1.69	0.409781057	0.409781057	0.000000000	93.27846%
2.00	0.393813298	0.393690968	0.000122330	96.08540%
2.66	0.378248119	0.378248119	0.000000000	98.73112%
5.00	0.373364647	0.370564313	0.002800334	99.95802%
10.00	0.369524682	0.369524682	0.000000000	99.98205%
50.00	NA	0.362412747	NA	99.98239%
100.00	NA	0.353715001	NA	99.98281%
1,000.00	NA	0.228430321	NA	99.98890%
10,000.00	NA	0.002882415	NA	99.99986%
100,000.00	NA	0.000000000	NA	100.00000%

Figure 6.1.1 Density Of Maximum Entropy Solution In Example 6.1.2



reasonable extrapolation for an empirical model given extra information about the tail, it is a reasonable extrapolation given the pattern in the Table L values available.

6.2 Determining A Distribution Consistent with Excess Ratios in Loss Elimination Ratio Tables

The WCIRB also produces tables of Loss Elimination ratios (LERs), that are excess ratios on a per accident basis in terms of the dollar amount of the limit (see [13]). Although the WCIRB releases some details of the underlying probability distribution, which is fairly complicated, recovering a maximum entropy distribution from the table of LERs illustrates a different approach from the Tables M and L example in the previous section, since in that case the overall mean was known to be 1.00

due to the normalization to produce entry ratios. Additionally, the final tables of LERs contain excess ratios rounded to only 3 digits, contributing to the difficulty of recovering the underlying distribution.

Example 6.2.1 Maximum Entropy Distribution for WCIRB's 2019 Overall LERs

Below is WCIRB's 2019 table of overall (all Hazard Groups combined) LERs. The Excel spreadsheet available online has only 3 digits of precision. The values in **Bold** have been selected for the specified constraints to fit.

Table 6.2.1 Selected Values For Fitting From WCIRB's 2019 Loss Elimination Ratios (Overall, All Hazard Groups)

Limit	Excess Ratio	Constraint Index
0	**1.000**	**1**
25,000	**0.689**	**2**
35,000	0.617	
50,000	**0.533**	**3**
75,000	0.434	
100,000	**0.368**	**4**
150,000	0.290	
200,000	0.247	
250,000	**0.219**	**5**
300,000	0.199	
400,000	0.172	
500,000	**0.154**	**6**
600,000	0.141	
700,000	0.131	
800,000	0.122	
900,000	0.115	
1,000,000	**0.109**	**7**
2,000,000	**0.072**	**8**
3,000,000	0.053	
4,000,000	0.040	
5,000,000	**0.031**	**9**
6,000,000	0.024	
7,000,000	0.019	
8,000,000	0.015	
9,000,000	0.012	
10,000,000	**0.010**	**10**
15,000,000	**0.004**	**11**
20,000,000	**0.001**	**12**

The generalized moment functions, corresponding to constraint indexed rows in the prior table, can be defined as:

$$g_i(x) = LER_i x - Max(0, x - x_i), \quad i = 1, ...,12$$

Note, $g_1(x) = 0$ for all x, so we can set $a_1 = 0$ and eliminate $g_1(x)$ from the density function:

$$p(x) = exp(-a_0 - a_2\,(0.689x - Max(0, x - 25{,}000)) \quad - \cdots - a_{12}\,(0.001x - Max(0, x - 20{,}000{,}000)))$$

and 12 relevant constraint equations, including normalization, in integral form are:

$$\int_0^\infty p(x)\,dx = 1$$

$$\int_0^\infty (0.689x - Max(0, x - 25{,}000))p(x)\,dx = 0$$

……

$$\int_0^\infty (0.001x - Max(0, x - 20{,}000{,}000))p(x)\,dx = 0$$

Some calculus and numerical root finding, similar to what was done for Table L in Example 6.1.2, is required. This includes sequentially solving for small subsets of the parameters and constraints, to be used as initial search points for the next larger subsets, as described before. This process leads to:

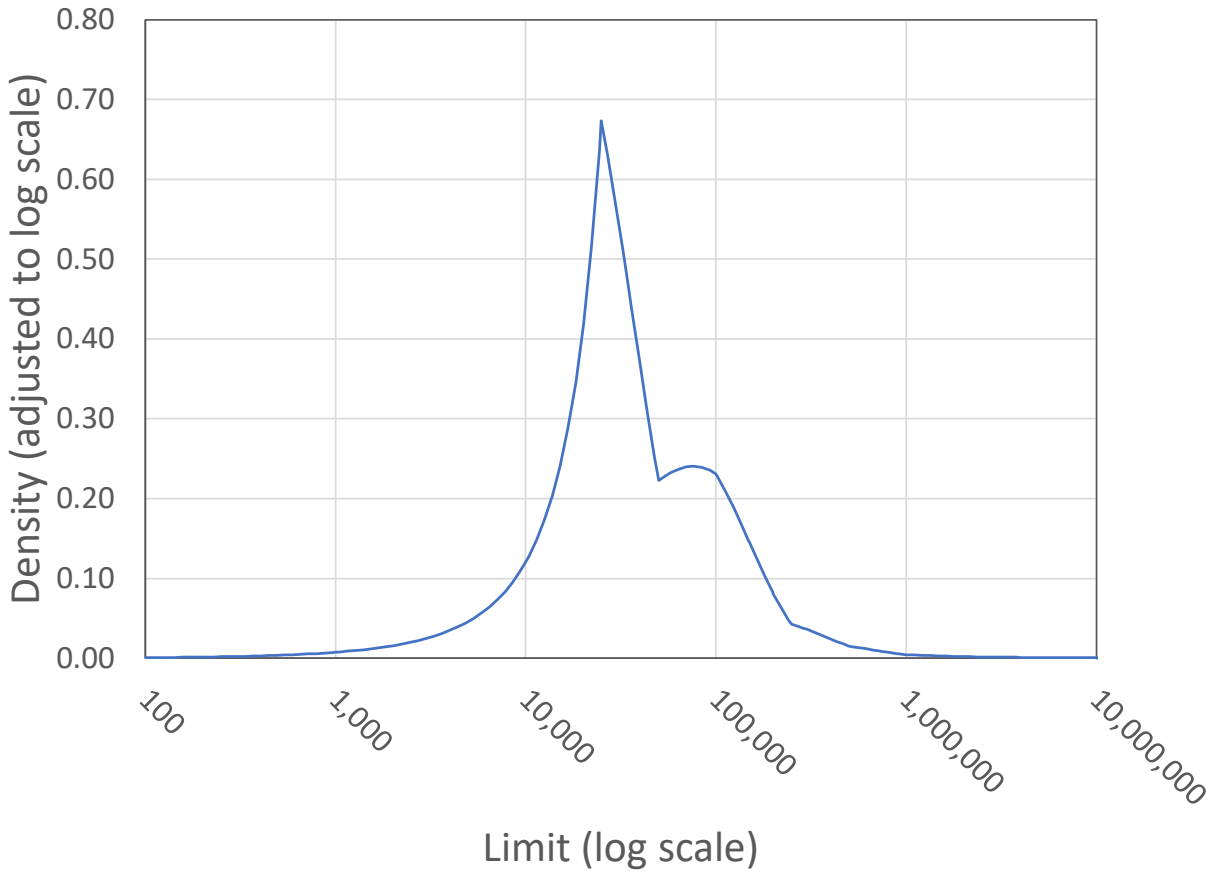
$a_0 =$	11.864788254009900000000
$a_2 =$	-0.000125730464385769000
$a_3 =$	0.000058923444849539900
$a_4 =$	-0.000004271540735520050
$a_5 =$	0.000010447441490470400
$a_6 =$	0.000002998956261260340
$a_7 =$	0.000002383315118547030
$a_8 =$	0.000000899163635770769
$a_9 =$	0.000000396715370202262
$a_{10} =$	-0.000000178908277119939
$a_{11} =$	0.000000541076550263008
$a_{12} =$	-0.000000464839849564974

The mean of the fitted maximum entropy distribution is $68,730 with standard deviation $272,939, and corresponding coefficient of variation 397%. Below are the actual and fitted LERs, including some extrapolated limits.

Table 6.2.2 Some Excess Ratios And Cumulative Distribution Values From The Solution In Example 6.2.1

Limit	Actual LER	Fit LER	Actual – Fit LER	Fit Cumulative Probability
0	**1.000**	**1.0000000000000**	**0.0000000000000**	**0.000000%**
25,000	**0.689**	**0.6890000000001**	**-0.0000000000001**	**37.053989%**
35,000	0.617	0.6130435581319	0.0039564418681	56.252159%
50,000	**0.533**	**0.5330000000001**	**-0.0000000000001**	**68.272954%**
75,000	0.434	0.4357874659584	-0.0017874659584	77.759218%
100,000	**0.368**	**0.3680000000002**	**-0.0000000000002**	**84.596264%**
150,000	0.290	0.2880494554155	0.0019505445845	92.319659%
200,000	0.247	0.2456477754649	0.0013522245351	95.560066%
250,000	**0.219**	**0.2190000000002**	**-0.0000000000002**	**96.919603%**
300,000	0.199	0.1993658363524	-0.0003658363524	97.641066%
400,000	0.172	0.1721086529617	-0.0001086529617	98.512431%
500,000	**0.154**	**0.1540000000002**	**-0.0000000000002**	**98.948454%**
600,000	0.141	0.1406382524283	0.0003617475717	99.198536%
700,000	0.131	0.1302860472939	0.0007139527061	99.367439%
800,000	0.122	0.1219664467532	0.0000335532468	99.481514%
900,000	0.115	0.1150196400910	-0.0000196400910	99.558558%
1,000,000	**0.109**	**0.1090000000002**	**-0.0000000000002**	**99.610593%**
2,000,000	**0.072**	**0.0720000000002**	**-0.0000000000002**	**99.827239%**
3,000,000	0.053	0.0522656417306	0.0007343582694	99.894359%
4,000,000	0.040	0.0397372191010	0.0002627808990	99.929672%
5,000,000	**0.031**	**0.0310000000002**	**-0.0000000000002**	**99.948252%**
6,000,000	0.024	0.0243601864887	-0.0003601864887	99.959998%
7,000,000	0.019	0.0192358385274	-0.0002358385274	99.969187%
8,000,000	0.015	0.0152970520606	-0.0002970520606	99.976376%
9,000,000	0.012	0.0122857403603	-0.0002857403603	99.982000%
10,000,000	**0.010**	**0.0100000000002**	**-0.0000000000002**	**99.986400%**
15,000,000	**0.004**	**0.0040000000001**	**-0.0000000000001**	**99.994451%**
20,000,000	**0.001**	**0.0010000000001**	**-0.0000000000001**	**99.997607%**
25,000,000	NA	0.0001753644633	NA	99.999580%
50,000,000	NA	0.0000000290837	NA	~100%
100,000,000	NA	0.0000000000000	NA	~100%

Figure 6.2.1 Density Of Maximum Entropy Solution In Example 6.2.1



6.3 Fitting a Distribution to Match Higher Moments

The maximum entropy distribution to match a specified set of m positive integer moments $\{E[X^{k_1}], \ldots, E[X^{k_m}]\}$, if it exists, has a very straight forward form:

$$p(x) = exp(-a_0 - a_1 x^{k_1} - \ \ldots \ - a_m x^{k_m}) \tag{18}$$

There is a closed form solution for the density of the maximum entropy distribution, if it exists, for a non-negative random variable with a single higher positive integer moment specified.

Example 6.3.1 Maximum Entropy Distribution For A Single Higher Moment

A non-negative random variable is known to have a mathematically consistent k[th] moment equal b.

$$a_0 = log\left(\int_0^\infty exp(-a_1 x^k)\, dx\right) = \log\left(\Gamma\left(1+\frac{1}{k}\right) a_1^{-\frac{1}{k}}\right)$$

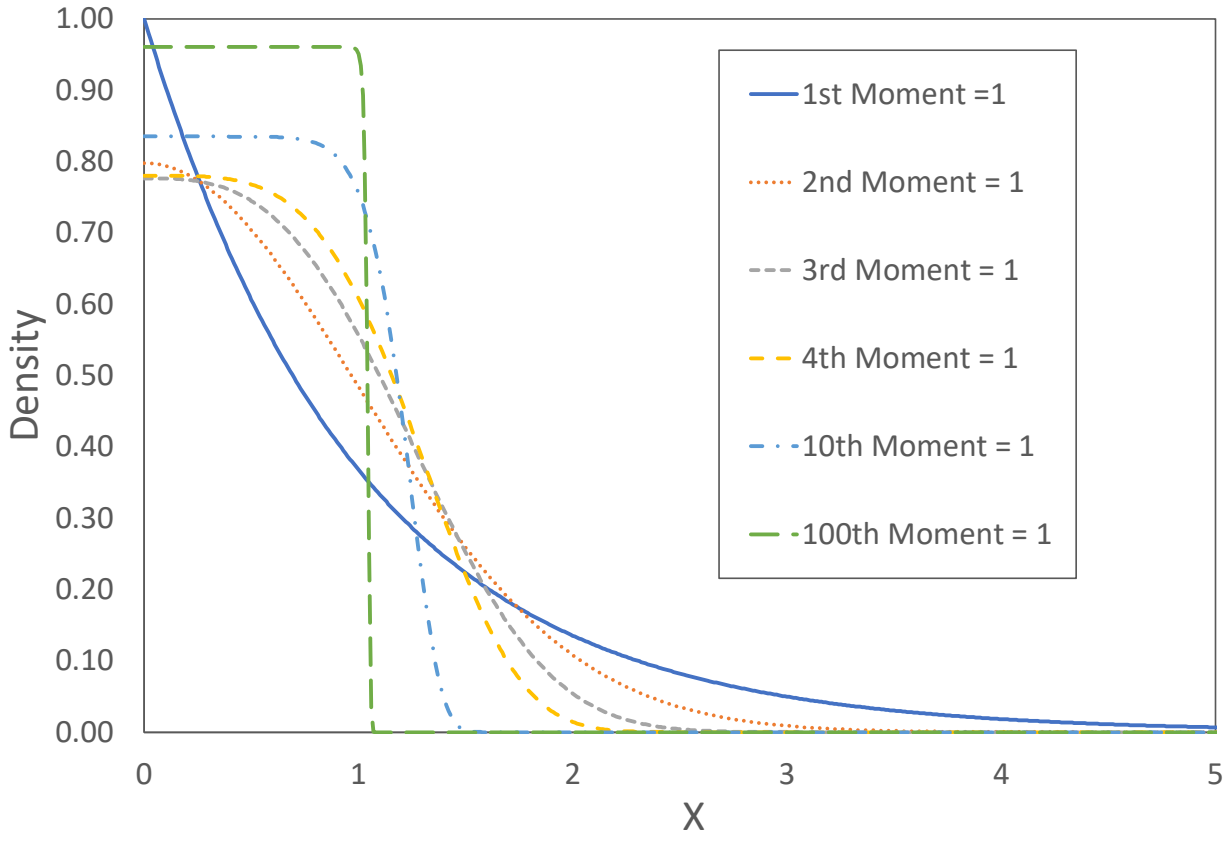
$$= log\left(\Gamma\left(1+\frac{1}{k}\right)\right) - \frac{1}{k}\log(a_1)$$

$$-\frac{\partial a_0}{\partial a_1} = \frac{1}{k\, a_1} = b \Longrightarrow \ a_1 = \frac{1}{k\, b}$$

$$p(x) = \frac{exp\left(-\frac{1}{k\, b} x^k\right)}{\Gamma\left(1+\frac{1}{k}\right)(k\, b)^{1/k}}$$

Multiple higher moments can be a challenge to numerically solve. (For a treatment of this general problem aimed at applications in physics see [12].) As of this writing, the author has not yet found a generally effective and satisfactory way, even using the sequential parameter/constraint subset process that worked very well for the excess ratio problems described in Examples 6.1.2 and 6.2.1, to reliably solve for a significant set (4, 5, or more) of the higher moments. A practical way of doing this would be particularly useful in many applications.

Figure 6.3.1 Density Of Maximum Entropy Solution In Example 6.3.1



Example 6.3.2 Maximum Entropy Distribution For 1st And 3rd Moment

A non-negative random variable is known to have mean 15 and 3rd moment 5,000.

$$g_1(x) = x$$

$$g_2(x) = x^3$$

$$\mathrm{p(x)} = \exp\left(-a_0 - a_1 x - a_2 x^3\right)$$

$$a_0 = log\left(\int_0^\infty \exp\left(-a_1 x - a_2 x^3\right) dx\right)$$

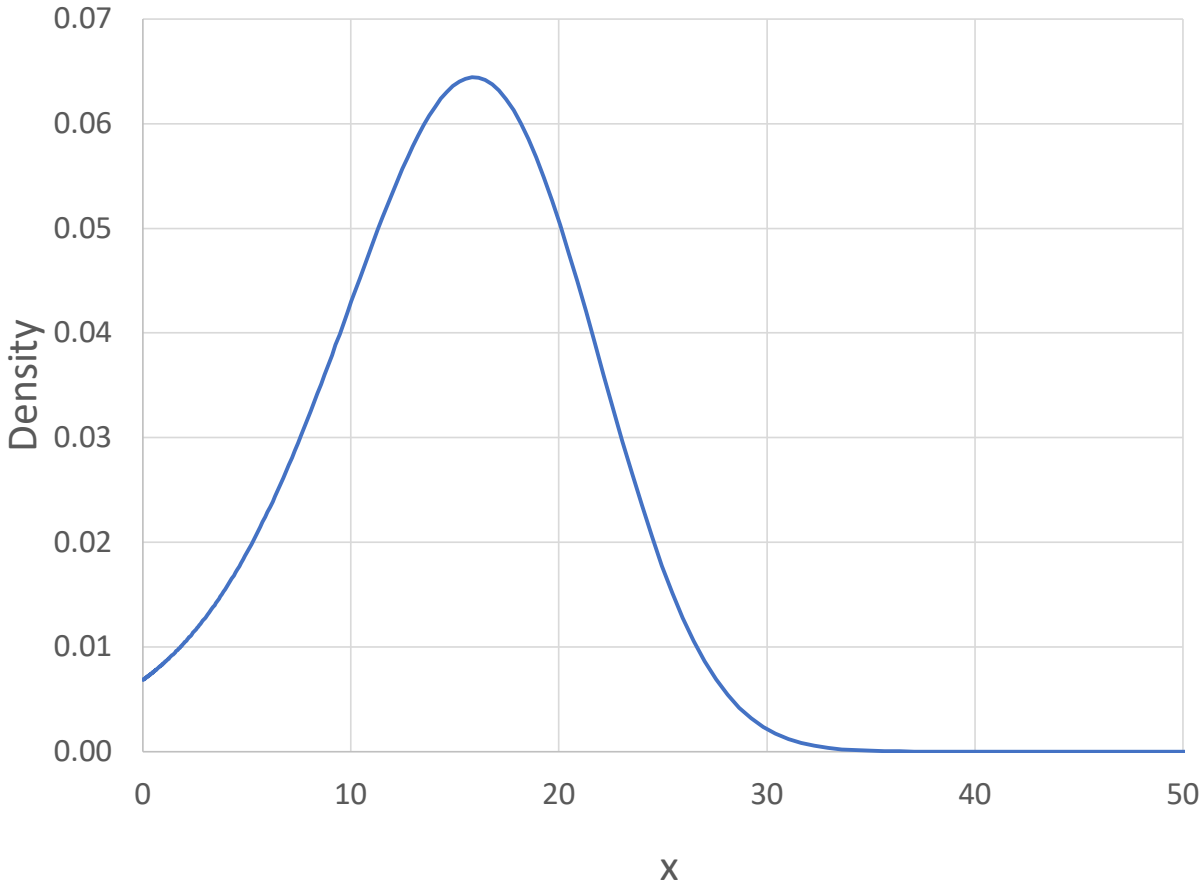
$$-\frac{\partial a_0}{\partial a_1} = \frac{\int_0^\infty x \exp(-a_1 x - a_2 x^3)\, dx}{\int_0^\infty \exp(-a_1 x - a_2 x^3)\, dx} = 15$$

$$-\frac{\partial a_0}{\partial a_2} = \frac{\int_0^\infty x^3 \exp(-a_1 x - a_2 x^3)\, dx}{\int_0^\infty \exp(-a_1 x - a_2 x^3)\, dx} = 5{,}000$$

A numerical search leads to:

$$a_0 = 4.98497 \qquad a_1 = -0.211337 \qquad a_2 = 0.000278004$$

Figure 6.3.2 Density Of Maximum Entropy Solution In Example 6.3.2



6.4 Implicit Aggregate Loss Models

In practice, sometimes only limited information is available about the distribution of aggregate losses for a portfolio of risks, but a maximum entropy distribution can be determined.

Example 6.4.1 Maximum Entropy Distribution for TVAR And Mean

A primary insurance company estimates the 99% Tail Value at Risk (TVAR) of its aggregate losses is \$1 billion and has a current booked ultimate aggregate loss of \$100 million. If we interpret the booked ultimate as an expected value and let the 99th percentile be an unknown value L, generalized moment functions may be set up as follows:

$$g_1(x) = x$$

$$g_2(x) = 0 \; if \; x < L$$

$$= \; 100x \; if \; x \geq L$$

$$g_3(x) = 0 \; if \; x < L$$

$$= \; 1 \; \; if \; x \geq L$$

Then the constraint equations, though quite complicated, may be set up as:

$$a_0 = log\left(\int_0^\infty \exp\left(-a_1 x - a_2 g_2(x) - a_3 g_3(x)\right) dx\right)$$

$$= log\left(\frac{\exp(-a_1 L) - 1}{-a_1} + \frac{-\exp(-a_3 + (-a_1 - 100\, a_2)L)}{-a_1 - 100\, a_2}\right)$$

$$-\frac{\partial a_0}{\partial a_1}$$

$$= -\frac{-\frac{\exp\left(-a_3 - (\, a_1 + 100 a_2)L\right)}{(a_1 + 100 a_2)^2} - \frac{1 - exp(-a_1 L)}{{a_1}^2} + \frac{L\; exp(-a_1 L)}{a_1} - \frac{L \exp\left(-a_3 - (\, a_1 + 100 a_2)L\right)}{a_1 + 100 a_2}}{\frac{\exp\left(-a_3 - (\, a_1 + 100 a_2)L\right)}{a_1 + 100 a_2} + \frac{1 - \exp(-a_1 L)}{a_1}}$$

$$= \$100{,}000{,}000$$

$$-\frac{\partial a_0}{\partial a_2} = -\frac{-\frac{100\exp(-a_3-(a_1+100a_2)L)}{(a_1+100a_2)^2} - \frac{100\,L\exp(-a_3-(a_1+100a_2)L)}{a_1+100a_2}}{\frac{\exp(-a_3-(a_1+100a_2)L)}{a_1+100a_2} + \frac{1-exp(-a_1L)}{a_1}}$$

$$= \$1{,}000{,}000{,}000$$

$$-\frac{\partial a_0}{\partial a_3} = -\frac{-\exp(-a_3-(a_1+100a_2)L)}{(a_1+100a_2)\left(\frac{\exp(-a_3-(a_1+100a_2)L)}{a_1+100a_2} + \frac{1-exp(-a_1L)}{a_1}\right)}$$

$$= 0.01$$

A numerical root finding search found plausible solutions for L from around \$92 million through around \$980 million, with the entropy seeming to peak around L = \$566 million. Here are some properties of the solution at L = \$566 million:

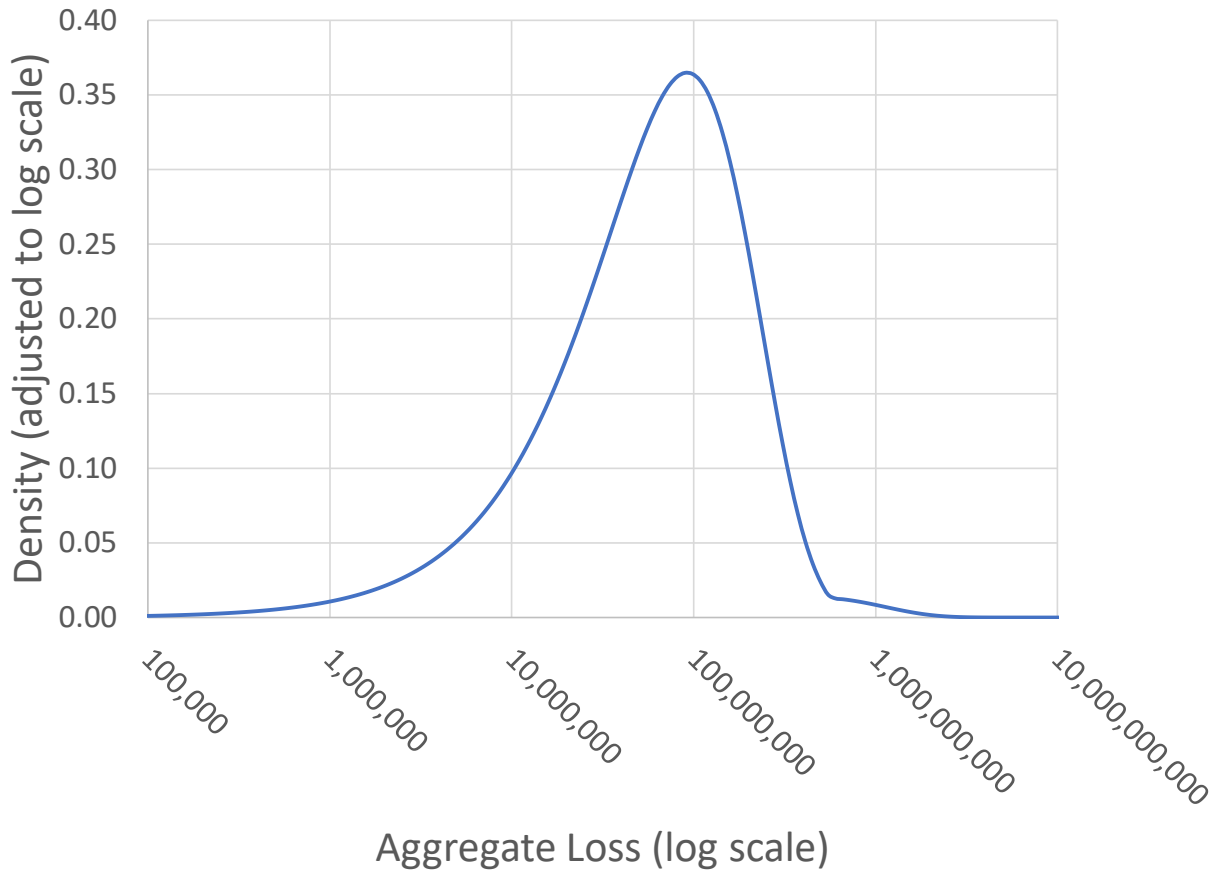
$$a_0 = 18.3466 \qquad a_1 = 1.08546 \times 10^{-8} \qquad a_2 = -8.55045 \times 10^{-11} \qquad a_3 = 4.843$$

The standard deviation is \$133.4 million, for a coefficient of variation of 133.4%. Some interesting percentiles and corresponding unlimited expected excess loss amounts are:

Table 6.4.1 Some Expected Excess Losses And Cumulative Distribution Values From The Solution In Example 6.4.1

Attachment	Percentile	Unlimited Expected Excess
\$0	0%	\$100 million
\$10 million	10.2%	\$90.5 million
\$50 million	41.6%	\$61.3 million
\$100 million	65.7%	\$38.7 million
\$200 million	87.9%	\$17.5 million
\$500 million	98.8%	\$5.1 million
\$1 billion	>99.6%	\$1.6 million

Figure 6.4.1 Density Of Maximum Entropy Solution In Example 6.4.1



Example 6.4.1 is structurally similar to "*4.1. Case A: Constraining the Global Mean"* from [8], except that in latter the risk-taker's wealth, analogous to L in Example 6.4.1, is specified rather than solved as part of maximizing entropy given the other constraints.

6.5 Bayesian or Credibility Estimation

Bayesian estimation generally requires model assumptions that completely specify both the prior distribution of parameters and the conditional density, or likelihood, of observations. Credibility estimation generally does not require complete distributional specifications but does require model assumptions that specify certain distributional variances. Maximum entropy distributions can be utilized in many cases to formulate these model assumptions where the available information would not otherwise completely specify them. Examples 6.5.1 and 6.5.2 apply maximum entropy distributions to conventional Bayesian and Credibility approaches. In Section 7 we will present a much more general multivariate maximum entropy framework that can automatically implement an implicit Bayesian/Credibility type adjustment for multivariate predictive models.

Example 6.5.1 Maximum Entropy Distributions For Bayesian Prior And Likelihood

Detailed data in not available, but it is known that in prior experience individual drivers have averaged 0.1 claims per year. What is the posterior distribution for expected number of claims after an individual driver has experienced $k \in \{0, 1, 2 \ldots\}$ claims in a single year?

Since the average number of claims is non-negative and we only know the mean is 0.1, the maximum entropy prior is simply a continuous Exponential Distribution with density function $q(m) = 10 \exp(-10\, m)$. The maximum entropy density on the discrete numbers $\{0, 1, 2 \ldots\}$ given the conditional mean is also of the Exponential form $p(k|m) = \exp(-a_0(m) - a_1(m)\, k)$, where the parameters solve to $a_0(m) = Log(m+1)$ and $a_1(m) = Log\left(\frac{m+1}{m}\right)$. Therefore $p(k|m) = m^k(m+1)^{-k-1}$ and the posterior density is $q(m|k) = \frac{10 \exp(-10\, m)\, m^k(m+1)^{-k-1}}{\int_0^\infty 10 \exp(-10\, m)\, m^k(m+1)^{-k-1}\, dm}$. So, the numerical results for several values of k are (up through $k = 5$):

Table 6.5.1 Bayesian Posterior Results from Maximum Entropy Solution in Example 6.5.1

k	Posterior Density $q(m\|k)$	Posterior $\hat{m} = E[m\|k]$
0	$\frac{10.921 \exp(-10m)}{m+1}$	0.09214
1	$\frac{138.95\, m \exp(-10m)}{(m+1)^2}$	0.17230
2	$\frac{1003.3\, m^2 \exp(-10m)}{(m+1)^3}$	0.24409
3	$\frac{5383.1\, m^3 \exp(-10m)}{(m+1)^4}$	0.30962
4	$\frac{23823\, m^4 \exp(-10m)}{(m+1)^5}$	0.37026
5	$\frac{91815\, m^5 \exp(-10m)}{(m+1)^6}$	0.42695

Example 6.5.2 Maximum Entropy Distributions to Determine Process and Parameter Variances for Credibility

What would the credibility estimates be for Example 6.4.1?

The Variance of the Hypothetical Means (VHM) = 0.01, that is the variance of the continuous Exponential Distribution with 0.1 mean. The process variance for the conditional density $p(k|m) = m^k(m+1)^{-k-1}$ is $m(m+1)$. So, the Expected value of the Process Variance (EPV) $= \int_0^\infty m(m+1)\, 10 \exp(-10\, m)\, dm$ = 0.12. Consequently, the credibility constant is $K = \frac{EPV}{VHM} = 12$ and since we only have one observation $Z = \frac{1}{1+K} = \frac{1}{13}$.

Table 6.5.2 Credibility Results From Maximum Entropy Solution In Example 6.5.2

k	Credibility $\hat{m} = \left(\frac{1}{13}\right) k + \left(\frac{12}{13}\right) 0.1$
0	0.09231
1	0.16923
2	0.24615
3	0.32308
4	0.40000
5	0.47692

7. MAXIMUM ENTROPY PREDICTIVE OR EXPLANATORY MODELS

Actuarial models often involve predicting or explaining the distribution, or at least the expected value, of one random response variable Y, scalar or vector, given the outcome of another random variable X, scalar or vector. For example, Generalized Linear Models (GLMs), though usually from a fixed effects standpoint, are commonly used for this purpose. This can be described in a very general framework in terms of a single vector valued random variable $X = \{Y_1, \dots, Y_m, X_1, \dots, X_n\}$ consisting of both response components $X_{resp} = \{Y_1, \dots, Y_m\}$ and explanatory components $X_{expl} = \{X_1, \dots, X_n\}$. Fixed effects can also be included in the generalized moment functions $g_i(x)$ and/or the specified generalized moments c_i. If the complete joint density $p(y_1, \dots, y_m, x_1, \dots, x_n)$ is known then the density of the response components $x_{resp} = \{y_1, \dots, y_m\}$ conditioned on the realized values of the conditioned on the realized values of the explanatory components $x_{expl} = \{x_1, \dots, x_n\}$ through the Bayesian calculation:

$$p(y_1, \dots, y_m | x_1, \dots, x_n) = \frac{p(y_1, \dots, y_m, x_1, \dots, x_n)}{\int \dots \int p(y_1, \dots, y_m, x_1, \dots, x_n)\, dy_1 \dots dy_m} \tag{19}$$

Example 7.1 Correlated Bivariate Maximum Entropy Distribution

Suppose the random variable $X = \{Y_1, X_1\}$ is known to have the following properties:

- Y_1 has mean 2,000 and standard deviation 2,000
- X_1 has mean 3,000 and standard deviation 3,000

- Y_1 and X_1 have a correlation coefficient of 30%

The basic linear regression model is:

$$Y_1 = m\,X_1 + b + \varepsilon(0,\sigma)$$

$m = 30\% \left(\frac{2{,}000}{3{,}000}\right) = 0.2 \qquad b = 2{,}000 - 0.2(3{,}000) = 1{,}400$

$\varepsilon(0,\sigma)$ is a normally distributed random variable, independent of Y_1 and X_1, with mean 0 and standard deviation $\sigma = \sqrt{(2{,}000)^2 - (0.2 \times 3{,}000)^2} = 1{,}908$.

The same result can be obtained by solving for the maximum entropy distribution for Y_1 and X_1, both assumed to be real values, with the following generalized moment constraints:

$g_1(X) = Y_1 \quad E[g_1(X)] = 2{,}000 \qquad g_2(X) = {Y_1}^2 \quad E[g_2(X)] = 8{,}000{,}000$

$g_3(X) = X_1 \quad E[g_3(X)] = 3{,}000 \qquad g_4(X) = {X_1}^2 \quad E[g_4(X)] = 18{,}000{,}000$

$g_5(X) = Y_1 X_1 \qquad E[g_5(X)] = 7{,}800{,}000$

The maximum entropy distribution would be the same as the Bivariate Normal Distribution, since it can match the given constraints and can be stated in the standard form in (3). The Bayesian calculation in (16) would then result in the same linear regression model.

However, suppose we also know that $Y_1 \geq 0$. Now, the normality assumption for Y_1 underlying the linear regression model clearly is a poor choice. However, the maximum entropy distribution can still be numerically solved with this extra piece of information by setting up the same moment constraints equations above, but changing the region of integration for calculating the underlying integrals:

$$a_0 = log\left(\int_0^{+\infty}\int_{-\infty}^{+\infty} \exp\left(-a_1y_1 - a_2{y_1}^2 - a_3x_1 - a_4{x_1}^2 - a_5y_1x_1\right) dx_1dy_1\right)$$

$$a_i = -\frac{\partial a_0}{\partial a_i} = \frac{\int_0^{+\infty}\int_{-\infty}^{+\infty} g_i(x_1) \exp\left(-a_1y_1 - a_2{y_1}^2 - a_3x_1 - a_4{x_1}^2 - a_5y_1x_1\right) dx_1dy_1}{\int_0^{+\infty}\int_{-\infty}^{+\infty} \exp\left(-a_1xy_1 - a_2{y_1}^2 - a_3x_1 - a_4{x_1}^2 - a_5y_1x_1\right) dx_1dy_1}$$

Numerical root finding leads to:

$a_0 = 16.748$ $a_1 = 0.000615231$ $a_2 = 1.23815 \times 10^{-8}$ $a_3 = -0.000256411$

$a_4 = 6.105 \times 10^{-8}$ $a_5 = -5.49445 \times 10^{-8}$

Figure 7.1 Expected Value of Response Variable Conditional On Explanatory Variable In Example 7.1

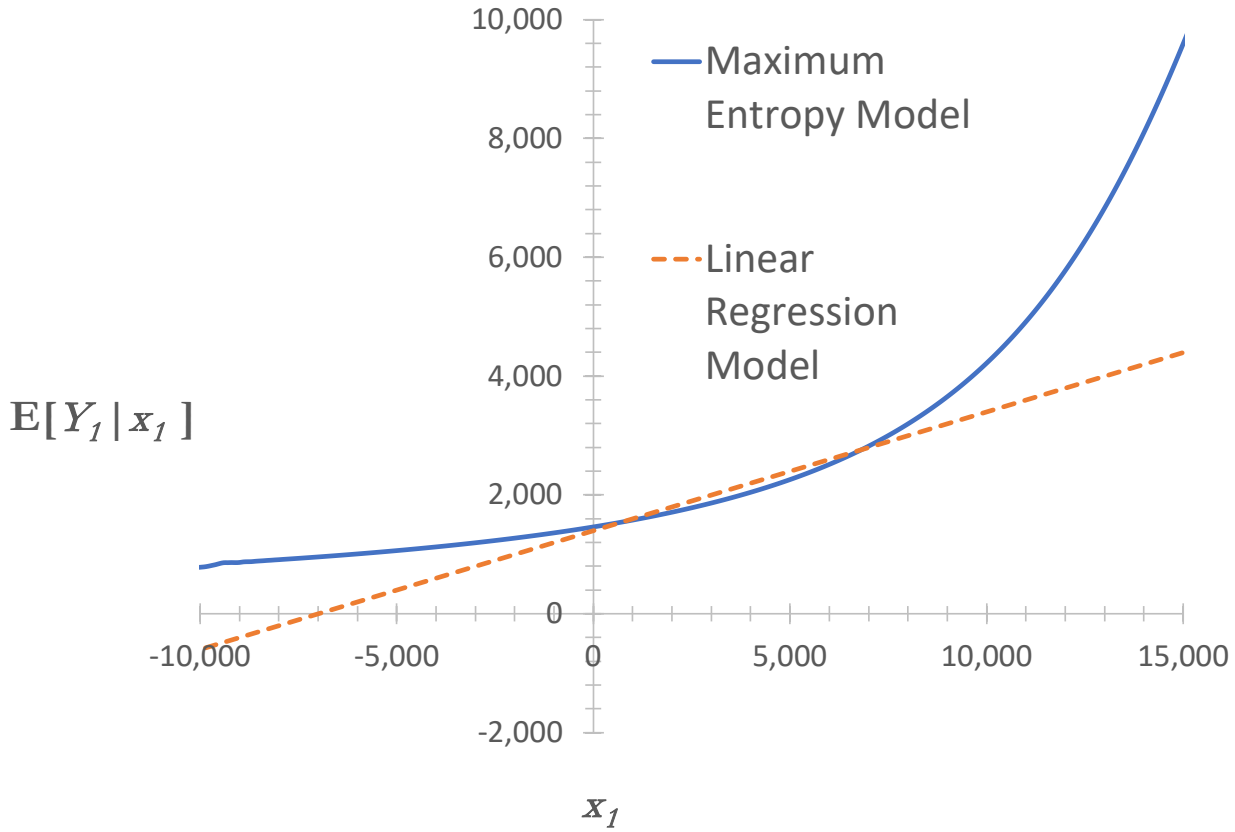


Figure 7.1 shows how the values for $\mathrm{E}[Y_1|x_1]$, for the two different models, diverge in both the left and right tails of X_1. The maximum entropy model naturally captures effects of the restriction $Y_1 > 0$ but the linear regression model does not. Figure 7.2 shows that for the conditional density $\mathrm{p}(\mathrm{y}_1|\mathrm{x}_1 = -7{,}000)$ in the left tail of X_1 the linear regression model incorrectly shows that Y_1 is equally likely to be positive or negative. Figure 7.3 shows that for the conditional density $\mathrm{p}(y_1|x_1 = 12{,}000)$ in the right tail of X_1 the linear regression model gives almost no probability that $Y_1 \geq 10{,}000$, but the maximum entropy model gives 16% probability that $Y_1 \in [10{,}000\ ,\ 20{,}000]$.

Figure 7.2 Density Of Response Variable Conditioned On Explanatory Variable = -7,000 In Example 7.1

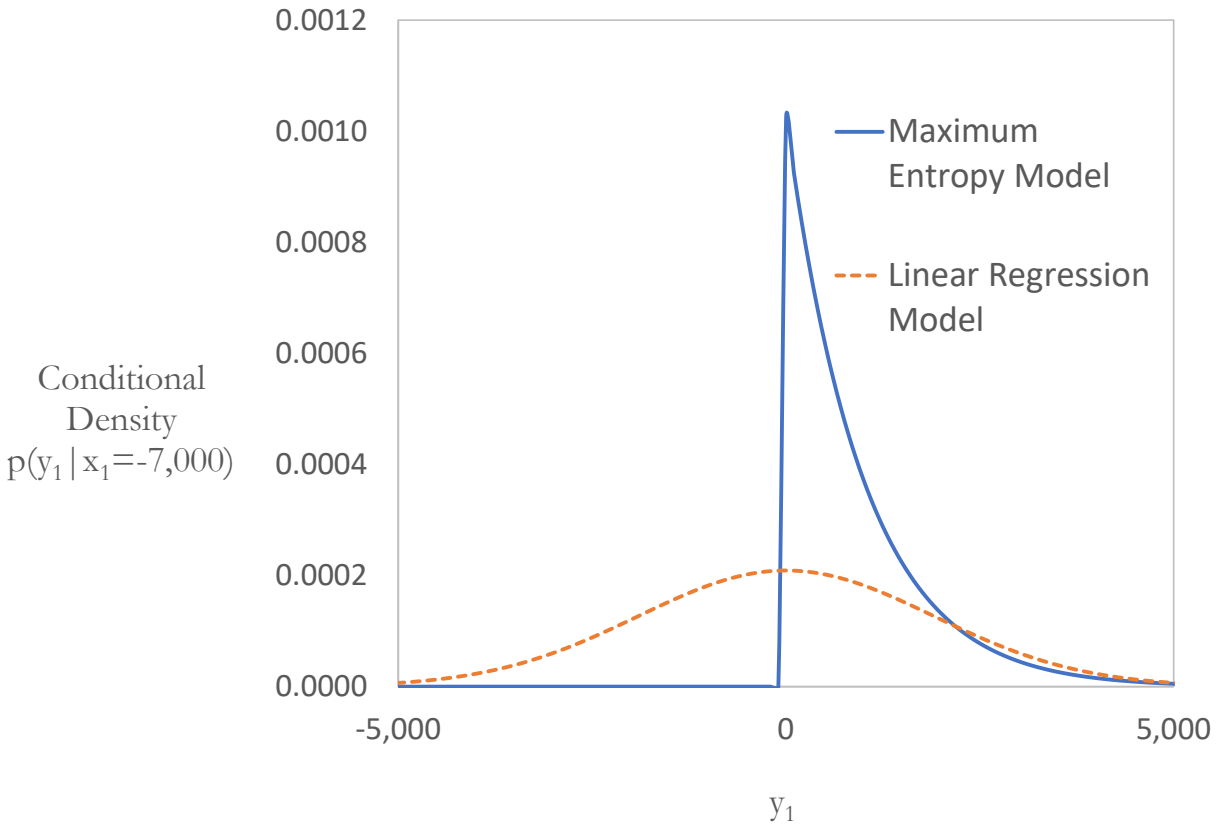
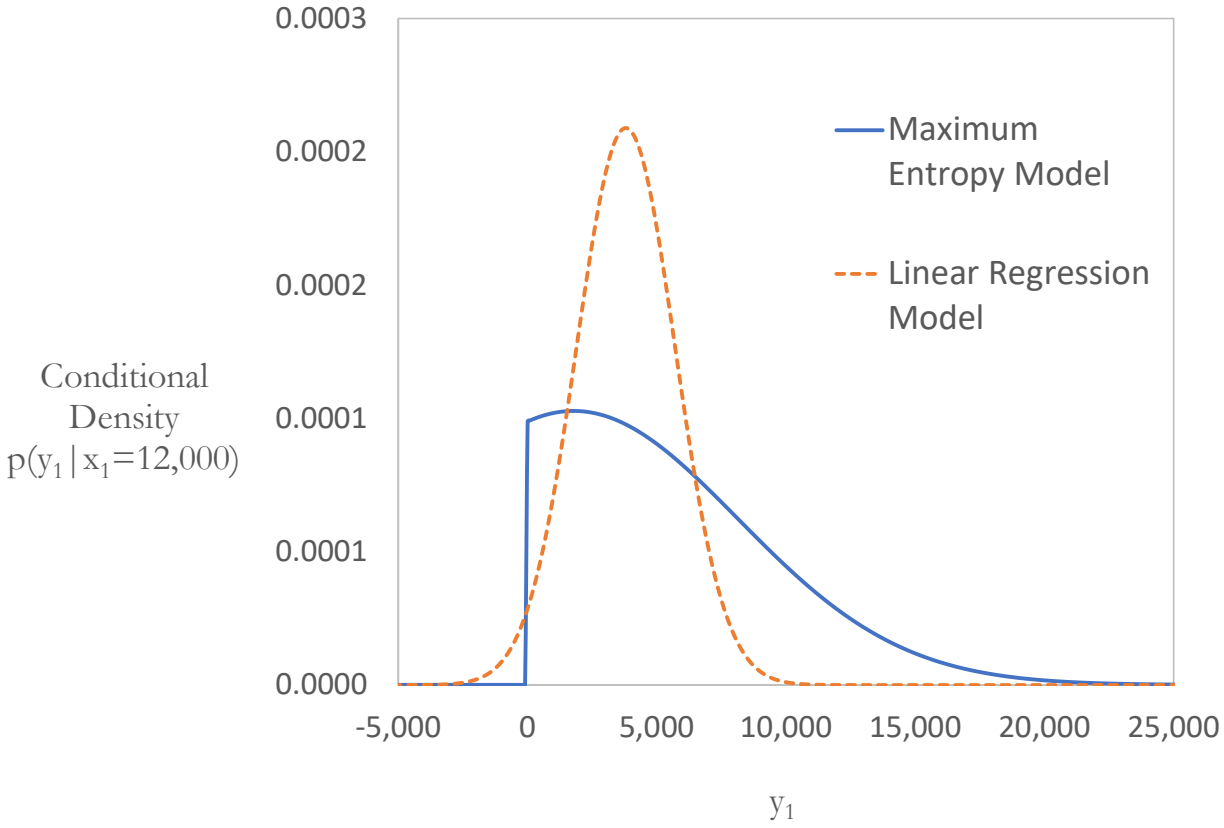


Figure 7.3 Density of Response Variable Conditioned on Explanatory Variable = 12,000 in Example 7.1



GLMs require the specification of a design matrix for the explanatory variables, a link function that connects them to the expected value for the response variables, and a conditional distribution for the response variables. When GLMs are fit for maximum likelihood they can be very vulnerable to low volume erratic observations in levels for certain factors, and incorporating credibility adjustments into GLMs (random effects, Gibbs sampling, etc.) can be a very awkward and difficult process.

In contrast it can be very straightforward to simultaneously fit a multi-factor model and incorporate credibility type adjustments when fitting a maximum entropy distribution.

Example 7.2 Maximum Entropy Multivariate Model With Automatic Bayesian/Credibility Adjustment

Suppose the following pure loss ratio experience is available for workers compensation insurance:

Setting	Business Type	Experience Pure Loss Ratio	Volume of Experience
Urban	Manufacturing	500%	?
Urban	Service	60%	?
Rural	Service	0%	?

Although the volume of experience is not known, the following information is given:

- There is thought to be no aggregate off balance, so that the overall expected pure loss ratio is 100%.
- Broader longtime experience has shown that the mean squared error between actual loss ratios for categories like these and a very good relativity estimate is 1.

A log-Poisson GLM, which has a conditional variance of 1 when the conditional expected value of the response variable is 1, fairly consistent with the bullets above, produces multiplicative relativity indications:

Setting	GLM Relativity Indication	Business Type	GLM Relativity Indication
Urban	2.000	Manufacturing	1.786
Rural	0.000	Service	0.214

This GLM has likely been fooled by randomness, as these values do not seem very realistic. Hopefully, when final full premium rates are implemented Rural policies will be charged more than $0.

Alternatively, this situation can be approached as a maximum entropy problem as follows. Let Y_1 be the actual outcome losses, X_1 and X_2 be random effects corresponding to good estimates for multiplicative relativities for Setting and Business Type, respectively.

The constraints will be:

$$g_1(X) = Y_1 \quad E[g_1(X)] = 1 \qquad g_2(X) = X_1 \quad E[g_2(X)] = 1$$

$$g_3(X) = X_2 \quad E[g_3(X)] = 1 \qquad g_4(X) = (Y_1 - X_1X_2)^2 \qquad E[g_4(X)] = 1$$

Setting up these equations involves integration in 3 dimensions:

$$a_0 = log\left(\int_0^{+\infty}\int_0^{+\infty}\int_0^{+\infty} \exp\left(-a_1y_1 - a_2x_1 - a_3x_2 - a_4(y_1 - x_1x_2)^2\right) dy_1 dx_1\, dx_2\right)$$

A numerical solution is:

$$a_0 = 0.235246 \qquad a_1 = 0.717116 \qquad a_2 = 0.856358 \qquad a_3 = 0.856358$$

$$a_4 = 0.213261$$

$$p(y_1, x_1, x_2) = exp\,(-a_0 - a_1y_1 - a_2x_1 - a_3x_2 - a_4(y_1 - x_1x_2)^2)$$

So, this gives the joint density of the observed loss ratio Y_1 and good estimates for the relativities X_1 and X_2. In the data table we need to estimate 4 relativities $\{X_{1U}, X_{1R}, X_{2M}, X_{2S}\}$ based on 3 observations. The posterior joint density of these relativities conditioned on the observations is:

$$q(x_{1U}, x_{1R}, x_{2M}, x_{2S}) =$$

$$= \frac{p(5, x_{1U}, x_{2M})\; p(0.6, x_{1U}, x_{2S}) p(0, x_{1R}, x_{2S})}{\int_0^{+\infty}\int_0^{+\infty}\int_0^{+\infty}\int_0^{+\infty} p(5, x_{1U}, x_{2M})\; p(0.6, x_{1U}, x_{2S}) p(0, x_{1R}, x_{2S})\, dx_{1U}\, dx_{1R}\, dx_{2M}\, dx_{2S}}$$

The overall mean values for the relativities using this joint density demonstrate a Bayesian/Credibility type of shrinkage in the relativity indications, and are clearly more realistic:

Setting	Max Entropy Relativity Indication	Business Type	Max Entropy Relativity Indication
Urban	1.261	Manufacturing	2.644
Rural	0.996	Service	0.438

The GLM relativities predict a 0% pure loss ratio for Rural Service policies.

		Pure Loss Ratio		
Setting	Business Type	Experience	GLM	Max Entropy
Urban	Manufacturing	500%	357%	333%
Urban	Service	60%	43%	55%
Rural	Manufacturing	NA	0%	263%
Rural	Service	0%	0%	44%

Although this example did not include any volume of experience, that could be used for weights, the GLM would have still given a 0.000 relativity indication if weights had been available and included in the GLM run. Some sort of credibility adjustment could have been implemented with the GLM, but it would have been somewhat awkward and ambiguous to set up given the limited amount of data. In contrast the Maximum Entropy model was very natural and unambiguous to set up with a built in Bayesian/Credibility type adjustment.

Figures 7.4 and 7.5 show the marginal densities for the Setting and Business Type relativities, respectively. The maximum entropy distribution naturally yields the parameter uncertainty of the fit.

Figure 7.4 Marginal Densities Of Setting Relativities From Maximum Entropy Approach In Example 7.2

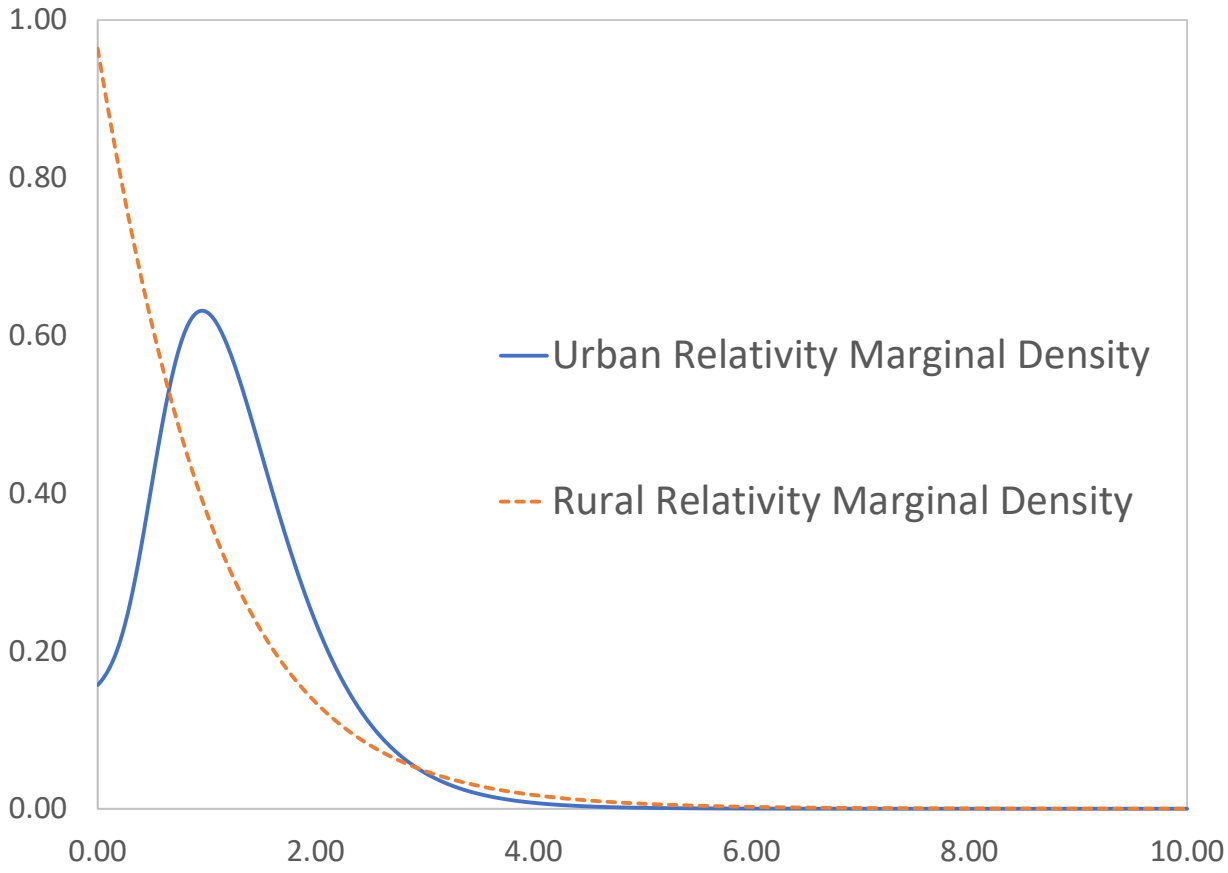
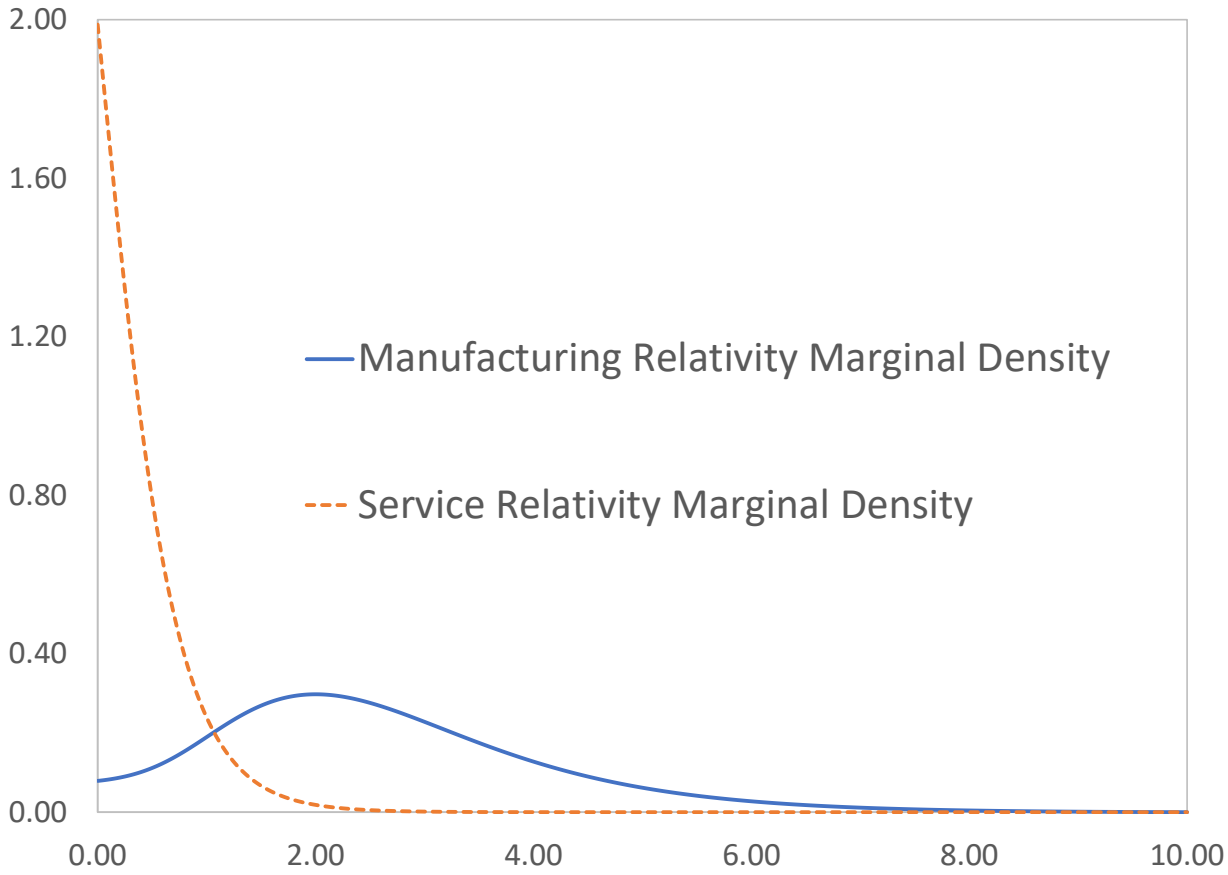


Figure 7.5 Marginal Densities Of Business Type Relativities From Maximum Entropy Approach In Example 7.2



Acknowledgment

The author thanks Curtis Gary Dean, Louise Francis and Steve Mildenhall for providing valuable discussion and feedback during the process of developing the material underlying this paper.

Appendix A - Computational and Software Coding Challenges

Although much more can be done with modern computing power than was possible in the past, fitting maximum entropy models is still often very challenging because it usually entails solving a set of highly nonlinear equations. These non-linear equations sometimes contain very lengthy expressions and usually require integration (or summation) and sometimes differentiation to set up. Numerical root finders are readily available in many software environments, such as Excel (Solver), R, Python, MATLAB, etc. Some numerical optimizers, like Google's Tensorflow, are designed to utilize powerful Graphics Processing Unit (GPU) hardware. Additionally, symbolic manipulation of complex expressions, including integration and differentiation, is available in software environments such as MATHEMATICA and Maple.

Even with modern software and hardware resources, converging on a numerical solution is often an arduous process involving restatement of the coding of the problem and reselecting initial search points, even when the problem has a similar form to a previous problem. Relying on symbolic manipulation is also undesirable for practical applications. Unfortunately, much of the coding involved for the examples in this paper is rather messy, complicated and not really standardized to general classes of problems. So, at present, no code samples are provided with this paper. Nevertheless, developing software code, preferably for commonly available environments such as Excel and R, that reliably solves broad classes of maximum entropy problems would provide a very valuable resource for practicing actuaries. Hopefully, this paper will encourage others to do so and the author may also pursue developing such standardized software tools.

Nevertheless, here are some tips that were useful in solving for the numerical examples in this paper.

It is often helpful to solve for the parameters $\{a_0, a_1, \dots, a_m\}$ sequentially stepwise. $\{a_0, a_1\}$ can be solved first, while zeroing out $\{a_2, \dots, a_m\}$ and ignoring the constraint equations for $\{g_2(x), \dots, g_m(x)\}$. Then, this solution can be used as an initial search point to solve for $\{a_0, a_1, a_1\}$ while zeroing out $\{a_3, \dots, a_m\}$ and ignoring the constraint equations for $\{g_3(x), \dots, g_m(x)\}$, and so on.

One potentially problematic issue is that numerical root finding software typically uses inaccurate finite differencing approximations for derivatives, as part of a Newton-Raphson iteration. It is possible to replace these finite difference calculations with more accurate numerical integrations. We can restate the problem of solving for the maximum entropy distribution in vector form.

Given:

$$C = (1, c_1, \dots\ , c_m) \text{ and } G(x) = (1, g_1\ (x), \dots\ , g_m(x))$$

Find:

$$A = (a_0, a_1, \dots\ , a_m\ ) \text{ such that } \int G(x)\ \exp\left(-A \cdot G(x)\right)\ dx\ = C.$$

This is equivalent to finding a root $\hat{A}$ for the vector valued function:

$$V(A) = \int G(x) \exp\left(-A \cdot G(x)\right)\ dx\ - C$$

Newton-Raphson Iteration can be done by first picking a starting point $\hat{A}_0$ and then iterating $\hat{A}_{n+1} = \hat{A}_n - \left(\nabla_A V(A)|_{\hat{A}_n}\right)^{-1} V\left(\hat{A}_n\right)$. The practical problem comes in when root finding software attempts to approximate $\nabla_A V(A)|_{\hat{A}_n}$ through small numerical differences.

However, a more accurate approach is to note that:

$$\left(\nabla_A V(A)\right)_{i,j} = \partial_{a_j}\left(\int g_i(x) \exp\left(-A \cdot G(x)\right)\ dx\ - C\right)$$

By differentiating under the integral sign:

$$= -\int g_i(x)\, g_j(x) \exp\left(-A \cdot G(x)\right)\ dx$$

It is generally much easier and more accurate to numerically estimate these integrals. If the limits of integration are unbounded there may be problems with these integrals numerically diverging for some values of $\hat{A}_n$ even if a solution exists. So, it may be useful to either limit the bounds of integration (that is the domain of possible outcomes for the random variable) or limit x to a finite number of values so that the integral may be replaced with a finite sum.

In some cases, it may be helpful to perform a transformation on A. For example, if $G(x) = (1, x, \ldots\ , x^m)$ and $x \in (0, +\infty)$ then substituting $a_m = \exp(b)$ and solving in terms of $B = (a_0, a_1, \ldots\ , a_{m-1}, b)$ will keep the integrals above from diverging. Note that after a substitution like this, due to the chain rule, the integrals corresponding to differentiation with respect to b will need to be multiplied by $\exp(b)$, specifically:

$$\left(\nabla_B V(B)\right)_{i,m} = -\exp(b) \int_0^{+\infty} g_i(x)\, g_m(x) \exp\left(A \cdot G(x)\right)\ dx$$

Appendix B – Clarification of Some Confusions of the Maximum Entropy Distribution Technique With Several Other Distinct Things

Some reviewers of an earlier draft of this paper confused maximum entropy distributions with several other very different things that actuaries have remained conscious of, and utilized, over the decades following the 1960s. We will clarify the differences below. It is worth noting that in practice these other things generally required much lower computational burdens than maximum entropy distributions, and hence were more practically tractable during this time.

Ordinary Method of Matching Moments

An ordinary method of moments fit of a distribution is not necessarily a maximum entropy distribution because the selected parametric form to be fit may not be the appropriate maximum entropy form.

Example B.1

Matching a first moment of 10,000 with the family of Uniform Distributions of with density $1/a$ for $x \in [0, a]$ and 0, results in $a = 20{,}000$ and entropy 9.90349, as was shown in Example 1.1 However, the maximum entropy distribution for a non-negative random variable with first moment of 10,000 is an exponential distribution and has entropy 10.2103

Furthermore, a maximum entropy distribution is not necessarily an example of ordinary matching moments since the generalized moment functions $g_i(x)$ are in fact very general functions, and certainly not constrained to be of the form x^k for some integer k. More general moment functions appear in many examples throughout this paper, such as Example 3.1

Exponential Families

It is noted in Section 4 that the maximum entropy form (3) presented in Section 3 is a subset of the exponential family and the generalized moment functions $g_i(x)$ play the role of sufficient statistics for form (3) when sample data is given. However, the constraints for maximum entropy distributions may come from any source, such as expert opinion, a priori hypothesis, etc.; not necessarily sample data.

Many actuaries have encountered the exponential family in the context of Generalized Linear Models (GLMs) or Exact Credibility, where the greatest accuracy credibility estimate equals the Bayesian posterior estimate. These contexts all require sample data and parametric family assumptions about underlying data generating processes, neither of which are required by maximum entropy distributions. It is also worth noting that the use of exponential families for GLMs and Exact Credibility, starting in the 1970s, was highly motivated by reduction of computational burdens in both cases. However, there was no apparent comparable technique to reduce the computational burdens of maximum entropy distributions to tractability at that time.

Model Selection Through Information Criteria

In 1974 the Akaike Information Criterion (AIC) was introduced as an estimator of relative quality among statistical models fit to sample data ([2]). There is also an important small sample adjusted version (AICc) ([3]) In 1978 a similar criterion, the Bayesian Information Criterion (BIC) was introduced ([15]). These criteria are useful for selecting among competing models hypothesized to underly sample data. We will first recount the definitions of these criteria. Then we will demonstrate how they differ from maximum entropy distributions with an explicit example. Finally, we will briefly discuss how the foundations behind how these information criteria were derived differs from maximum entropy distributions. The derivations of these criteria are very mathematically and statistically sophisticated. Consequently, we will not attempt to even approach the detail presented in the original references but will attempt to convey a meaningful general concept of what is going on.

Suppose a sample of data observations $\{x_1, \dots, x_n\}$ is given, as usual assumed to be independent and arising from the same underly model. Also, $\{M_1, \dots, M_q\}$ is a set of parametric probability distribution models hypothesized to potentially be the true model M underly the data with $\{k_1, \dots, k_q\}$ number of parameters, respectively. Let $\{\hat{L}_1, \dots, \hat{L}_q\}$ be the likelihood function values for the maximum likelihood estimates of the respective $\{k_1, \dots, k_q\}$ parameters of each of the models $\{M_1, \dots, M_q\}$ fit to $\{x_1, \dots, x_n\}$. The definitions of the information criteria mentioned previously are:

$$AIC_i = 2\,k_i - 2\,Log(\hat{L}_i)$$

$$AICc_i = 2\,k_i - 2\,Log(\hat{L}_i) + \frac{2{k_i}^2 + 2k_i}{n - k_i - 1}$$

$$BIC_i = Log(n)\,k_i - 2\,Log(\hat{L}_i)$$

For each of these criteria the lower the value the better the hypothesized model.

Example B.2

In Example 1.1 no sample data was given, and the competing hypothesized models were fit using moment matching (1st moment only) with no sample data available. We will now revisit this example for two different data samples, each having the target moment of mean 10,000. Samples 1 and 2 were simulated from Wide Uniform Distribution and the Lognormal Distribution, respectively, as given in Example 1.1 and then each renormalized to have sample mean 10,000. In Table B.1 the distributional forms from Example 1.1, aside from the Narrow Uniform, are shown with new MLE parameters for Sample 1 and 2, respectively.

Sample 1

2593	8679	14482	6228	18704	8007	7201	15804	19077	16323
13049	2496	10996	14183	11122	4822	7586	1190	14617	13185
11883	12326	10106	5147	10931	2028	6930	10831	19905	14175
3183	13533	18013	7577	7260	122	6944	13917	12260	12801
10785	10353	303	15271	7110	5782	20793	13601	15338	2310
13442	10969	2098	19147	13876	10363	19360	3819	4039	17333
3443	5289	13781	941	2447	1463	6538	5956	12488	3688
9485	7195	14626	5589	508	4992	3669	8560	13621	18331
14716	14085	8156	7254	15493	18884	4151	18093	19492	10101
4044	5861	18749	4795	13242	19799	6638	8207	16208	3114

Sample 2

8680	8738	9252	31428	130529	27498	3911	4505	27851	3821
5649	3606	8210	35463	1913	4739	1510	2091	3730	990
4952	21982	1287	8385	2894	3199	20020	11880	39584	481
454	6146	2297	2992	327	4872	34422	6901	8257	9864
3019	4126	3078	2017	3026	9434	6625	10319	2991	1592
1420	5975	5445	9908	2454	2006	615	15717	5517	49731
31062	11831	16390	15383	3006	2328	50596	10474	10083	17470
16103	969	1688	3765	1224	2892	15511	9384	6455	2592
952	727	9276	3108	20191	4123	1081	8776	4138	1003
17033	1286	6546	4754	10562	2868	4123	4625	3580	1717

Table B.1 Distributional Forms from Example 1.1 With Parameters Refit to Samples 1 and 2

Density Form	ME Density Function (Mean = 10,000)	MLE1 Density Function (MLE on Sample 1)	MLE2 Density Function (MLE on Sample 2)
Wide Uniform	$\frac{1}{20000} \quad x \in [0, 20000]$	$\frac{1}{20793} \quad x \in [0, 20793]$	$\frac{1}{130529} \quad x \in [0, 130529]$
Lognormal (for ME $\sigma = 1$)	$\frac{exp\left(-\left(\frac{1}{2}\right)\left(\frac{1}{2} - log(10000) + \log(x)\right)^2\right)}{x\sqrt{2\pi}}$	$\frac{0.43666\, exp(-0.599013\,(-8.94021 + \log(x))^2)}{x}$	$\frac{0.0.348916\, exp(-0.382465\,(-8.54211 + \log(x))^2)}{x}$
Exponential	$0.0001\, exp(-0.0001\, x)$	$0.0001\, exp(-0.0001\, x)$	$0.0001\, exp(-0.0001\, x)$
Pareto (for ME min loss =100)	$\frac{10000 \times 10^{2/99}}{99} x^{-199/99}$ Min=100	$0.772353\, x^{-1.24177}$ Min = 122	$2.97849\, x^{-1.36335}$ Min = 327

Table B.2 Information Entropy of Distributions from Table B.1

	Information Entropy			Information Entropy Rank		
Density Form	ME	MLE1	MLE2	ME	MLE1	MLE2
Wide Uniform	9.90349	9.94237	11.7794	3	4	1
Lognormal	10.1293	10.2688	10.095	2	2	4
Exponential	10.2103	10.2103	10.2103	1	3	3
Pareto	6.58512	11.3600	10.5545	4	1	2

The Wide Uniform and Exponential each have one parameter to fit. The Lognormal and Pareto each have two parameters to fit. Table B.2 shows that refitting parameters with MLE results in the form closest to the underlying data process, the Wide Uniform for MLE1 (Sample 1) and the Lognormal for MLE2 (Sample 2), having the lowest entropy, or the most information. This makes sense for this context of fitting to sample data, where the objective is to gain as much information from the data as possible. However, it stands in stark contrast with the criterion of maximum entropy when the objective is to simply match to certain generalized moment constraints.

Table B.3 shows AIC, AICc, and BIC calculated and ranked (lowest to highest) for the original and refit parameter estimates on each data sample. Not surprisingly, all three of the information criteria produce the same rankings in for each combination of sample data and parameter fits. Here again for MLE1 (Sample 1) and MLE2 (Sample 2) the forms closest to the underlying data process always rank 1st. However, it is worth noting that among the ME fits, simply to mean 10,000 without any sample data, the maximum entropy distribution, the Exponential Distribution, ranks 1st for Sample 1 and 2nd for Sample 2. Furthermore, when the sample is mismatched with the MLE fit, as with MLE1 (Sample 2) and MLE1 (Sample 1) the Exponential, which still has the same parameter value being the maximum entropy distribution for the sample mean, ranks 1st. When the MLE is matched to its sample, MLE1 (Sample 1) and MLE 2 (Sample 2), the Exponential ranks 2nd.

Table B.3 Information Criterion Calculated For Distributions From Table B.1 On Samples 1 and 2

Sample 1	AIC			AIC Rank		
Density Form	ME	MLE1	MLE2	ME	MLE1	MLE2
Wide Uniform	Infinity	996.2	1179.9	4	1	3
Lognormal	1034.3	1030.9	1041.3	2	3	2
Exponential	1023.0	1023.0	1023.0	1	2	1
Pareto	1334.9	1140.0	Infinity	3	4	4

Sample 1	AICc			AICc Rank		
Density Form	ME	MLE1	MLE2	ME	MLE1	MLE2
Wide Uniform	Infinity	996.3	1180.0	4	1	3
Lognormal	1034.4	1031.0	1041.4	2	3	2
Exponential	1023.1	1023.1	1023.1	1	2	1
Pareto	1335.0	1140.1	Infinity	3	4	4

Sample 1	BIC			BIC Rank		
Density Form	ME	MLE1	MLE2	ME	MLE1	MLE2
Wide Uniform	Infinity	998.8	1182.5	4	1	3
Lognormal	1039.5	1036.1	1046.5	2	3	2
Exponential	1025.6	1025.6	1025.6	1	2	1
Pareto	1340.1	1145.2	Infinity	3	4	4

Table B.3 Information Criterion Calculated For Distributions From Table B.1 On Samples 1 and 2 (continued)

Sample 2	AIC			AIC Rank		
Density Form	ME	MLE1	MLE2	ME	MLE1	MLE2
Wide Uniform	Infinity	Infinity	1179.9	4	4	4
Lognormal	1016.9	1028.9	1013.5	1	2	1
Exponential	1023.0	1023.0	1023.0	2	1	2
Pareto	1254.9	1090.6	1059.4	3	3	3

Sample 2	AICc			AICc Rank		
Density Form	ME	MLE1	MLE2	ME	MLE1	MLE2
Wide Uniform	Infinity	Infinity	1180.0	4	4	4
Lognormal	1017.0	1029.0	1013.6	1	2	1
Exponential	1023.1	1023.1	1023.1	2	1	2
Pareto	1255.0	1090.7	1059.6	3	3	3

Sample 2	BIC			BIC Rank		
Density Form	ME	MLE1	MLE2	ME	MLE1	MLE2
Wide Uniform	Infinity	Infinity	1182.5	4	4	4
Lognormal	1022.1	1034.1	1018.7	1	2	1
Exponential	1025.6	1025.6	1025.6	2	1	2
Pareto	1260.1	1095.8	1064.7	3	3	3

Example B.2 illustrates the difference between selecting generalized moment constraints, even if sample data is available, and determining the maximum entropy distribution, versus postulating several different parametric forms, MLE fitting the parameters, and then ranking them according to information criteria. Interestingly, the maximum entropy distribution fit, independent of any sample data, to a mean of 10,000 actually ranked very well on these two samples, both having mean 10,000 but otherwise being very different.

AIC, introduced in 1974 ([2]), derives from a Frequentist philosophy utilizing Information Theory. Specifically, AIC derives from an asymptotic (as $n \to \infty$) estimate of the Kullback–Leibler (K-L) divergence (also called relative entropy), between the true underlying distribution for sample data and a hypothesized parametric model. The K-L divergence was introduced in 1951 ([11]) as a type of generalization of information entropy. Akaike had earlier pointed out a relationship between Maximum Likelihood Estimation (MLE) and the K-L divergence ([1]). Among competing hypothesized models, the lower the K-L divergence the better, as it indicates a likely lower information difference between a hypothesized model and the true distribution. AICc is based on the same foundational reasoning, with the addition of a correction term to improve accuracy for small data samples. For a true underlying distribution model P with density $p(x)$ and a hypothesized distribution model P with density $q(x)$, the K-L divergences is defined as:

$$D_{KL}(P||Q) = H(P,Q) - H(P,P)$$

where $H(P,Q)$ is the cross entropy, defined as:

$$H(P,Q) = -\int Log\big(q(x)\big)\, p(x)\, dx$$

For the special case when the distributions P and Q are equal, the cross entropy $H(P,P)$ is the information entropy of a distribution, as used throughout this paper. Akaike derived the asymptotic estimate:

$$D_{KL}(M||M_i) = k_i - Log\big(\hat{L}_i\big) + Constant$$

Dropping the constant and multiplying by two, an arbitrary convention, leads to:

$$AIC_i = 2\, k_i - 2\, Log(\hat{L}_i)$$

BIC was introduced by Schwarz in 1978 ([15]), deriving from a Bayesian framework without utilizing Information Theory. In this framework a number of competing models are assumed to have the same probability, prior to any observed data. BIC is asymptotically related to the logarithm of the Bayes formula updated probabilities for each model M_i, posterior to data being observed, is derived. Schwarz derived the asymptotic estimate:

$$Log(Probability[M_i|data]) = Log(\hat{L}_i) - \frac{k_i}{2} Log(n) + Bounded\ Term$$

Dropping the bounded term and multiplying by -2 leads to the BIC criterion that can be used in the same way as AIC or AICc, as previously given:

$$BIC_i = Log(n)\ k_i - 2\ Log(\hat{L}_i)$$

So, although rooted in a Bayesian framework, BIC is also used to select among competing models in a Frequentist framework. It has also been noted that AIC can be derived in a similar fashion starting with a different prior distribution on the competing models ([4]). Alternatively, competing models could be weighted together in a Bayesian framework, based on posterior probabilities derived from prior probabilities related either to AIC or BIC.

Whereas maximum entropy distributions do not even require sample data or hypothesized parametric models, these information criteria require both. When both sample data and hypothesized parametric models are given, the maximum entropy distribution selected to match the sample value of a selected generalized moment function may be different from the model selected by these information criteria.

REFERENCES

[1] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in Petrov, B. N.; Csáki, F. (eds.), *2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281. Republished in Kotz, S.; Johnson, N. L., eds. (1992), *Breakthroughs in Statistics*, I, Springer-Verlag, pp. 610–624.

[2] Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19 (6): 716–723.

[3] Burnham, K. P.; Anderson, D. R. (2002), *Model Selection and Multimodel Inference: A practical information-theoretic approach* (2nd ed.), Springer-Verlag.

[4] Burnham, K. P.; Anderson, D. R. (2004), "Multimodel inference: understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33: 261–304.

[5] California Workers Compensation Insurance Rating Bureau, "2019 California Retrospective Rating Plan and Tables," https://www.wcirb.com/content/california-retrospective-rating-plan.

[6] Chalk, Alan and McMurtrie, Conan," A Practical Introduction to Machine Learning Concepts for Actuaries," *Casualty Actuarial Society Forum*, Spring, 2016.

[7] Conger, Robert F., "The Construction of Automobile Rating Territories In Massachusetts," *PCAS* LXXIV, 1987.

[8] Geman, D.; Geman, H.; Taleb, N.N., "Tail Risk Constraints and Maximum Entropy," *Entropy* **2015**, *17*, 3724-3737.

[9] Hurley, Robert L., "Discussion of A Discipline for the Avoidance of Unnecessary Assumptions," *PCAS* LV, 1968.

[10] Kull, Andreas., "A Unifying Approach to Pricing Insurance and Financial Risk," *Casualty Actuarial Society Forum*, Winter, 2003.

[11] Kullback, S.; Leibler, R.A. (1951). "On information and sufficiency," *Annals of Mathematical Statistics*. 22 (1): 79–86.

[12] Mead, L.R. and Papanicolaou, N. (1984). "Maximum entropy in the problem of moments," *Journal of Mathematical Physics* 25, 2404-2417.

[13] McKean, Rasa V., "Calibration of A Jump Diffusion," *Casualty Actuarial Society Forum*, Winter, 2013.

[14] Roberts, Lewis H., "A Discipline for the Avoidance of Unnecessary Assumptions," *PCAS* LIV, 1967.

[15] Schwarz, Gideon E. (1978), "Estimating the dimension of a model," Annals of Statistics, 6 (2): 461–464.

[16] Shannon, C.E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, pp. 379–423 & 623–656, July & October 1948.

[17] Wikipedia, "Maximum Entropy Probability Distributions," https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution, retrieved on 1/8/2019.

Biography of the Author

Jonathan Evans, FCAS, FSA, FCA, CERA, MAAA, WCP is a property and casualty consulting actuary with 25 years of experience. He is currently President of Convergent Actuarial Services, Inc.

Evolving Estimation Methodology

Charles Stein startled the statistical world with his 1956 paper "Inadmissibility of the usual estimator for the mean of a multivariate distribution." (Stein (1956)) He showed that when you are estimating more than two elements, shrinking their estimates towards the overall mean always reduces estimation error compared to MLE. The resulting James-Stein estimator is the same as the least-squares credibility estimator. The 1979 Morris and van Slyke CAS paper – also at Morris and Van Slyke (1978) – discussed the connection. A famous example is estimating seasonal batting averages for a number of batters from their July 4 averages.

Any kind of regression or linear model is estimating means for a vector of observations, but it was not so clear how to apply James-Stein/credibility shrinkage without having the variances of the observations – the process variance. Several years later, a form of shrinkage called ridge regression was developed. Instead of minimizing the negative loglikelihood (NLL) it minimized NLL $+\lambda \sum \beta_j^2$, where the β_j are the coefficients. Hoerl and Kennard (1970) proved that there is always some $\lambda > 0$ that makes the estimation error for this less than for straight MLE. This starts with linear transforms of the independent variables to make them each mean zero, variance one, which makes the parameter sizes comparable across the variables. The constant term (which is not included in the parameters being shrunk) and the coefficients adjust to compensate for the linear transforms. Each fitted value is the constant plus a linear combination of mean zero variables, so the constant is the overall mean, and the fitted values are shrunk towards that.

Demoment (1989) introduced lasso, which minimizes NLL $+\lambda \sum |\beta_j|$. This makes some of the coefficients exactly zero (all of them if λ is high enough), so is also a means of variable selection. This makes it popular, and it is largely replacing MLE for linear modeling. The problem is how to pick λ. The preferred method has become cross validation, which comes down to dividing the data into subsets which are left out one at a time and the NLL measured on the model estimated on the remaining points in turn. Adding up these left-out NLL pieces gives a penalized NLL which is used to find the best λ.

These methods are forms of regularization, which is a more general mathematical approach for reducing errors in difficult estimation problems, from Tikhonov (1943). (The rather uninformative name for this was an approximate translation from his original Russian.) Ridge regression started as a method for correlated data, which has some difficulties of its own. There was actually something resembling a ridge of artificial data used in that.

Bayesian shrinkage produces similar results. It starts with giving each parameter a shrinkage prior, which is a mean-zero, mode-zero prior like the standard normal. These pull the posterior

estimates towards zero, just like λ does in lasso and ridge regression. The Bayesian approach has a big advantage in that there are applicable goodness-of-fit measures like loo and WAIC, discussed below. Measures like AIC, etc. don't work because the shrunk parameters don't act as whole parameters, but what fraction to count is not apparent.

Random Effects Ties It All Together

The literature on random effects can be confusing and sometimes inconsistent. I take off from the setup in Klinker (2011). Bayesian and classical shrinkage have a lot in common, but they have a philosophical difference in that in classical statistics parameters are constants, but for Bayesians they have distributions. A link is provided by frequentist random effects. There you have a collection of statistical effects across a population – such as differences in accident frequency from the state mean by territory – that are assumed to average to zero. These look like parameters to Bayesians, but frequentists allow effects, but not parameters, to have distributions. With mean zero, their distribution across the population could be described by a single dispersion parameter, like the normal σ. Forms of this reproduce ridge regression and lasso estimates. (Most papers assume the effects to be normally distributed, but this is not a conceptual limitation.) Random-effects estimation and Bayesian shrinkage do not require all effects to have the same distribution, or to be independent, but I assume those here.

To see how this works, assume that the effects are double exponential (Laplace) distributed in λ. This is a distribution that looks like an exponential for positive values, and its mirror image over the y-axis for negative values. The density for an effect β is

$$f(\beta|\lambda) = 0.5\lambda e^{-\lambda|\beta|}$$

This has variance $= 2/\lambda^2$ and kurtosis $= 6$. Say there are k random effects β_j, plus perhaps other parameters, including λ. One way to simultaneously estimate the parameters and project the effects is to maximize the joint likelihood, which is the likelihood of the data, given the parameters and the random effects, times $\Pi f(\beta_j|\lambda)$, the probability of the effects. The negative of the log of $f(\beta_j|\lambda)$ is just $log(2) - log(\lambda) + \lambda|\beta_j|$. Then maximizing the joint likelihood becomes minimizing

$$NLL + \lambda \sum |\beta_j| - log(\lambda)k$$

For a fixed value of λ, the last term does not affect the minimization. The result is the lasso minimization formula, here used for the projection of the random effects. The value of λ produced by the minimization including the $-log(\lambda)k$ term gives the random-effects estimate of λ as well.

This connects with the Bayesian approach. For data X and parameters β, Bayes Theorem is:

$$p(\beta|X) = \frac{p(X|\beta)p(\beta)}{p(X)}$$

The left side is the posterior distribution of the parameters given the data, and the numerator of the right side is the likelihood times the prior. Here the β are parameters, but this numerator is the same mathematical formula as the joint likelihood in the random effects case, where the β are effects, not parameters. The denominator $p(X)$ is a constant for a given dataset, so maximizing the numerator maximizes the posterior. Thus the random effects solution gives the Bayesian posterior mode, and if the Laplace prior is used for the parameters, it gives classical lasso. This is why the use of the Laplace prior is called Bayesian lasso. A normal distribution for the random effects gives ridge regression.

Bayesian Markov Chain Monte Carlo (MCMC) estimation simulates a numerical sample from the posterior distribution of parameters by sampling from the joint likelihood – the numerator of Bayes Theorem – using sampling methods like the Hastings-Metropolis sampler or the Gibbs sampler. These are efficiency improvements over the original MCMC sampler that generates a sample from the previous sample. It has a candidate generator, and if the candidate sample improves the joint likelihood, it is retained. If not, a test on a random draw is done to keep it or not. It has been proved that after a burn-in period the sample is representative of the posterior distribution.

One detail here is that in the Bayesian case the optimization works as discussed above for a fixed value of λ. If λ is to be estimated as well, it also must be given a prior. If it has a uniform prior $= U$ over some interval, then a term $= -log(U)$ is included in the log of the prior. But since that is a constant, it does not affect the minimization, and so the posterior mode is still at the minimum of $NLL + \lambda \sum |\beta_j| - log(\lambda)k$ from random effects. Other priors might give better estimates of λ, however. Note that as λ increases, the parameters are pushed more towards zero to compensate, but that makes the NLL get higher. At the same time, $-log(\lambda)k$ is decreasing. Thus at some point they all balance at a minimum.

An increasingly popular shrinkage prior is the Cauchy distribution, with $1/p(\beta) = \pi(\lambda^2+\beta^2)/\lambda$ and $-log(p(\beta)) = -log\lambda + log\pi + log(\lambda^2 + \beta^2)$. For a fixed λ, the posterior mode minimizes $NLL + \sum log(\lambda^2 + \beta_j^2)$. This is an alternative to both lasso and ridge regression. The Cauchy prior often yields more parsimonious models than the normal or Laplace priors do. It can have a bit better or bit worse penalized likelihood (see discussion below), but even if slightly worse, the greater parsimony makes it worth considering. It has more weight near zero but is also heavier tailed, which pushes parameter more towards zero, but allows a few larger parameters

when they are called for. It also seems to produce tighter distributions of parameters.

The frequentist approaches (random effects, ridge regression, lasso) are all calculations of the posterior mode, which is a drawback, as there are some advantages of the posterior mean over the posterior mode. The parameters that maximize the posterior probability could be doing so by over-fitting the particular sample. This is sometimes described as fitting the sample vs. fitting the population. Too good a fit for the sample might be responding to particular features of that sample that would not hold for future samples.

The posterior mean averages all parameter sets that provide a plausible explanation of the data. The posterior mean does not optimize a comparison of actual vs. fitted values – in fact any such measure runs the risk of sample bias. It does minimize the squared difference of actual vs. estimated parameters. The posterior mode optimizes what could be called the lottery number measure for parameter error: all deviations from the exactly right parameter set are equally bad. But for parameters, even though not for the lottery, getting closer to the right answer is advantageous, so the mode is less appealing. Below I use the posterior mean, not the posterior mode, for the parameter estimation, so this does not agree exactly with classical lasso, etc. I do test different priors and different approaches for estimating λ.

All in all, random effects gives frequentists the ability to use a Bayesian-like framework without having to recognize parameter distributions. They start with a postulated unconditional distribution of effects, and project the effects from the data. There does not seem to be any reason that they could not also use MCMC to sample from the conditional distribution of the effects given the data, which would let them use the posterior mean instead of the posterior mode, but I haven't seen them actually do that.

Choosing λ and Goodness of Fit

How much shrinkage to do is usually selected using cross-validation: you divide up the data into subsets, fit using all but one subset, compute the NLL on the left out subset, repeat for each subset, and add up the NLLs. If resources are available, the limiting case of making each observation a subset seems to be preferred. This is called leave-one-out (loo) cross validation.

The sum of the individual NLLs from loo is known to be a good way to correct the NLL for sample bias – which is what penalized likelihood measures like AIC, BIC, etc. are trying to do as well. They are trying to estimate what the NLL would be for a new sample from the same population. But penalizing based on parameter counts doesn't work with shrinkage. If the estimation is overfitting, shrinking the parameters will reduce the overfitting but will increase the NLL as well. AIC etc. will not change the penalty in response to the shrinkage, but loo will. Thus it is a goodness-of-fit measure that still works fine with parameter shrinkage. It

stops improving when too much shrinkage deteriorates the NLL on the omitted points. Thus it provides a way to determine how much to shrink. (There is another goodness-of-fit measure called WAIC that also uses MCMC output. The loo measure has some minor technical advantages for some models, but the two measures generally rank models the same.)

The R lasso package glmnet is very fast and might make a grind-out calculation of loo feasible computationally for lasso. For MCMC, Gelfand (1996) developed an approximation for an omitted point's likelihood from an estimation for all the points, using the numerical integration technique importance sampling. This estimates a point's left-out likelihood by a weighted average likelihood across all the samples, with the weight for a sample proportional to the reciprocal of the point's likelihood under that sample. That gives greater weight to the samples that fit that point poorly, and is a good estimate of the likelihood the point would have if it had been left out of the estimation. This estimate turns out to be the harmonic mean over the samples of the point's probability in each sample. Then the MCMC sample of the posterior distribution is enough to calculate the loo goodness-of-fit measure.

This gives good but volatile estimates of the loo loglikelihood. Vehtari, Gelman, and Gabry (2017) addressed that by a method similar to extreme value theory – they fit a Pareto to the probability reciprocals and use the Pareto values instead of the actuals for the largest 20% of the reciprocals. This "Pareto-smoothed importance sampling" has been extensively tested and is becoming widely adopted. Their penalized likelihood measure is labeled $\widehat{elpd}_{loo}$, standing for "expected log pointwise predictive density." Here I call $-\widehat{elpd}_{loo}$ simply loo. There is an R package called loo that does this calculation quickly on MCMC output.

The fact that this is a good estimate of the NLL without sample bias comes with a caveat. The derivation of that assumes that the sample comes from the modeled process. That is a standard assumption but in financial areas, models are often viewed as approximations of more complex processes. Thus a new sample might not come from the process as modeled. Practitioners sometimes respond to this by using slightly under-fit models – that is more parsimonious models with a bit worse fit than the measure finds optimal. BIC was designed for this kind of situations as well, and it also leads to more parsimonious models.

Some analysts choose the variance of the shrinkage prior – like the double exponential prior – by maximizing loo under various degrees of shrinkage. But statisticians are coming to view this too as exposed to overfitting – the optimization is still responding to the particular sample. The fully Bayesian solution is to use the posterior mean with another prior placed on λ itself. This usually gives a value of loo close to that from direct optimization, but is felt to provide a more reliable result. This is the method used below.

In summary, the advantages of the Bayesian approach are:

- It facilitates calculation of a penalized likelihood measure based on cross validation.
- MLE has Fisher information for parameter uncertainty, but this is not convenient with shrinkage. MCMC automatically generates parameter distribution samples.
- The Bayesian approach can also incorporate a prior for λ, which both estimates λ and samples from a range of λ values instead of just a single λ.
- The frequentist methods end up with the posterior mode, which runs more risk of overfitting. The posterior mean is available from MCMC.
- MCMC packages facilitate using residual distributions inside or outside of the exponential family and can also estimate more complex model formulations – like a combination of additive and multiplicative models.

Moving from GLMM to MCMC

Venter, Gutkovich, and Gao (2017) is a discussion and attempt at using GLMM in standard software packages. It is by a group from the model validation area of AIG who were validating a reserving package by Spencer Gluck that extends his well-known generalized Cape Cod model to include diagonal effects, with the smoothing done by random effects. (Actually it was an anonymous referee from the Committee on Review of Papers who noticed that what Spencer was doing was essentially random effects – actually with correlated effects.)

We started using random effects programs with the default assumption that every random effect has its own variance. We found by an extensive fitting approach similar to classical loo that all these variances act as real parameters that pulled the data strongly towards the sample values. This prevented much actual parameter reduction – but it looked like there was a lot if you just looked at the non-variance parameters. We also found that including the many variances created an estimation problem when parameters got near zero – the likelihood gets very large if the projection and its variance both go to zero. The packages deal with this in ad hoc ways, apparently dropping such parameters by unstated rules. We decided to just use a single variance for all the effects, which we then found out came down to lasso and ridge regression. This led us to Bayesian implementation later on, as we discovered its advantages.

Fitting Plan

The data for the class ratemaking examples comes from Fu and Wu (2007). It is for auto collision, and has total losses and exposures for 8 age classes, 4 use classes, and 4 credit score ranges. It also has a different data set with average severity by these classes but with no credit scores. The modeling is in a few stages.

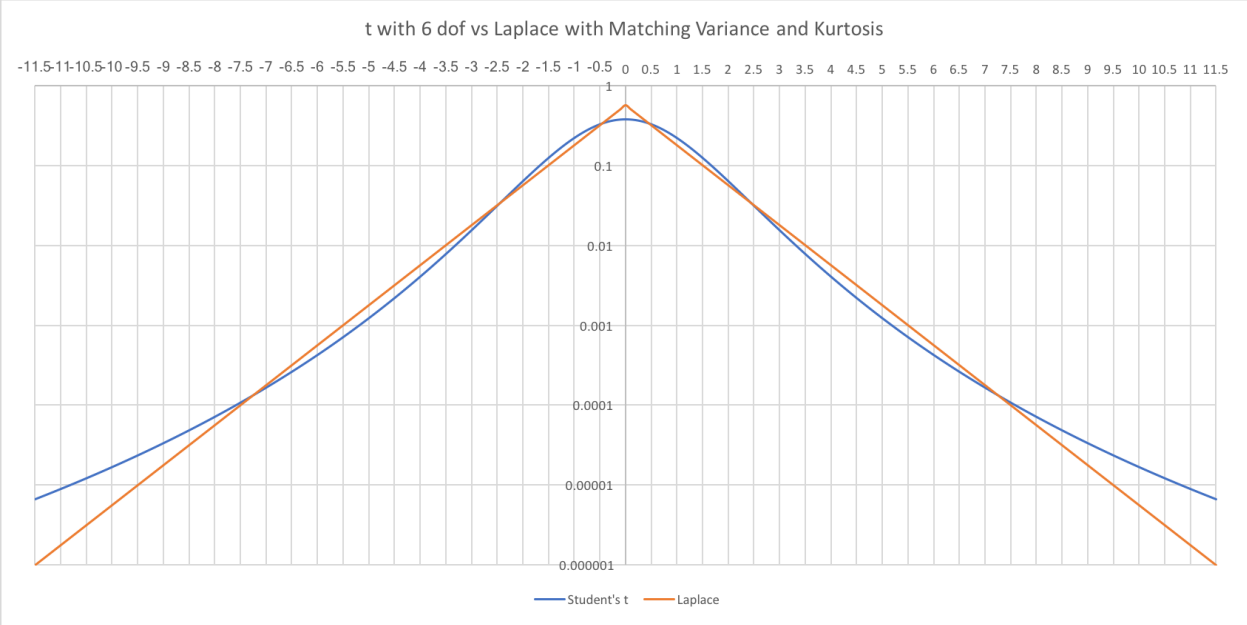


Figure 1: Scaled t with $\nu = 6$ and Double Exponential with Moments Matched

0 Preliminaries

Lasso, MLE regression, and Bayesian lasso are compared using a simplified model.

1 Warmup Model

This uses the less-detailed data to model severity with both an additive and a multiplicative model. The design matrix starts out having a constant plus a parameter for all but one level for each variable. Some variables might be eliminated in the estimation. Cauchy and double-exponential priors are used. The Cauchy is just a t distribution with 1 degree of freedom. The double-exponential has the same kurtosis as the t with 6 degrees of freedom, so is lighter tailed. The variance of this t and the double exponential can be matched using the scale parameters, and the odd moments are 0, so the existing moments of this t are all the same as the double exponential. The densities are graphed on a log scale in Figure 1.

Severity

You can't expect a good estimate for a severity distribution from just the sample mean. All you have on the dispersion of the losses is the degree to which the cell sample means differ from the fitted means. Still you can see the impact of the classification variables on average severity. Also there are some distributions for which the total losses and number of claims will give some information about the distribution. Consider the gamma distribution with mean $= ab$ and variance $= ab^2$. Assume cell$_j$ severity is distributed Gamma$[a_j, b_j]$. Given n_j claims, the sum of the claims is then distributed Gamma$[n_j a_j, b_j]$. For the gamma, b is a

scale parameter, so dividing the variable by a constant gives the distribution with b divided by that constant. Dividing the sum of claims by n_j gives the sample mean severity, and this is thus Gamma$[n_j a_j, b_j/n_j]$. It has mean $a_j b_j$ and variance $a_j b_j^2/n_j$. Thus if you only have the claim counts and sample means for each cell, and fit a Gamma$[a_j, b_j]$ distribution for cell$_j$'s sample mean, the severity distribution for the cell is Gamma$[a_j/n_j, b_j n_j]$, so you have estimated that as well. The normal and inverse Gaussian distributions work similarly.

A simplifying assumption for estimating the severity mean parameters by cell is to assume that either the a or the b parameter is constant across the cells. For severity, a constant a is more likely to give a better fit of the model to the data, as then the variance is proportional to the square of the mean. The a parameter will not be shrunk, so its prior will be uniform in its log. The severity mean will be the fitted shrunk independent variable parameters times the row of the design matrix for that cell.

The normal can also be parameterized to have the variance proportional to the mean-squared, just by using another unshrunk parameter s, and replacing σ everywhere with $s\mu$. This is a typical sort of heteroscedasticity adjustment. This will have similar severity mean and variance to the gamma distribution, but with zero skewness. The observed sample mean from n_j claims is now normal in $[\mu_j, mu_j\sqrt{(n_j s)}]$. Its fit can be compared to the gamma's by loo.

The inverse-Gaussian distribution usually is parameterized with mean μ and variance μ^3/ϕ. I prefer a slightly altered version with $\mu = ab$ and $\phi = a^2 b$. This gives variance ab^2 just like the gamma. Its big advantage is that the sum of claims has parameters $[n_j a_j, b_j]$, just like the gamma. It has skewness 3CV, compared to 2CV for the gamma. Thus it is just a bit more skewed gamma. This is a GGLMM distribution but not a GLMM distribution. It might be too skewed for collision data, but at the end of the severity-distribution example for the smaller dataset it is compared to the other distributions using loo, again assuming variance is proportional to mean-squared, so a is fixed across the cells.

2 Full Model

This data includes credit scores, and has total losses and exposures by cell. Multiplicative and additive models are fit to this, with a constant and a parameter for each level of each variable – leaving out one level of each variable for identifiability. The mean is the base pure premium for one unit of exposure for each cell. The data gives the exposure by cell, so multiplying this by the base pure premium gives the mean for the aggregate losses for the cell.

2a Residual Distributions

The starting point for the distribution of residuals for aggregate losses is the gamma, now

with b fixed across the cells. This has variance proportional to the mean, like the ODP has, and usually works reasonably well for aggregate losses. (Fixing a is usually better for severity, as then variance is proportional to mean squared.) Then a few residual distributions are compared. GLM uses a variance function that expresses the variance as a function of the mean. It can further adjust the cell means by an exposure measure. In his review of the Fu-Wu paper that this data comes from, Mildenhall suggests making the cell variance $V_j = V(\mu_j)/e_j^k$, where V is the variance function, e_j is the cell exposure, and k is a selected adjustment power. Here we can try estimating k as a parameter.

In GGLMM you can parameterize some of this. I try setting $V(\mu_j) = s\mu_j^k$ and estimating s, k as unshrunk parameters with log uniform priors on the reals. This can be used for any distribution. For a gamma or inverse Gaussian with mean $= ab$ and variance $= ab^2$, Stan can be set up to solve for a, b for a cell by taking $b_j = V_j/\mu_j$ and $a_j = \mu_j^2/V_j$. With this done, the variance and mean are specified by the linear model before the making choice of distribution. The distribution can then be selected based on other shape characteristics, such as skewness, using loo to indicate the best fit. This is more flexible than with GLM, where the variance function determines which distribution to use.

To start with I use the Gaussian, gamma, and inverse Gaussian distributions. A combination of the Gaussian and inverse Gaussian distributions – just a weighted sum of the two – can provide a lot of flexibility in the skewness, ranging from 0 to 3CV. Depending on where the skewness seems to be, weighting the gamma with either the normal or the inverse Gaussian is another alternative.

The Weibull is interesting as its skewness can be fairly high or even negative, depending on the mean and variance by cell. Unfortunately you cannot solve for its parameters from the moments in closed form, although it only takes solving a single non-linear equation to match mean and variance. Stan has some built-in functionality for solving specified equations, which I try for the Weibull.

The moments are simplified by using the Weibull parameterization

$$F(x) = 1 - exp[-(x/b)^{1/a}]$$

Then using the notation $n! = \Gamma(1+n)$ even for real n, $EX = a!b$, $1 + CV^2 = (2a)!/a!^2$, and $skew \times CV^3 = (3a)!/a!^3 - 3CV^2 - 1$. The skewness can really blow up for big CV – think of $a = 10$. Then $(3a)!$ is the product of $1 \ldots 30$, while $a!^3$ is the product of $1 \ldots 10$ 3 times. Also the skewness is negative for low CV – but never seems to get as low as -1.2. Figure 2 graphs the skewness as a function of the CV, and compares to the gamma skewness, which is 2CV.

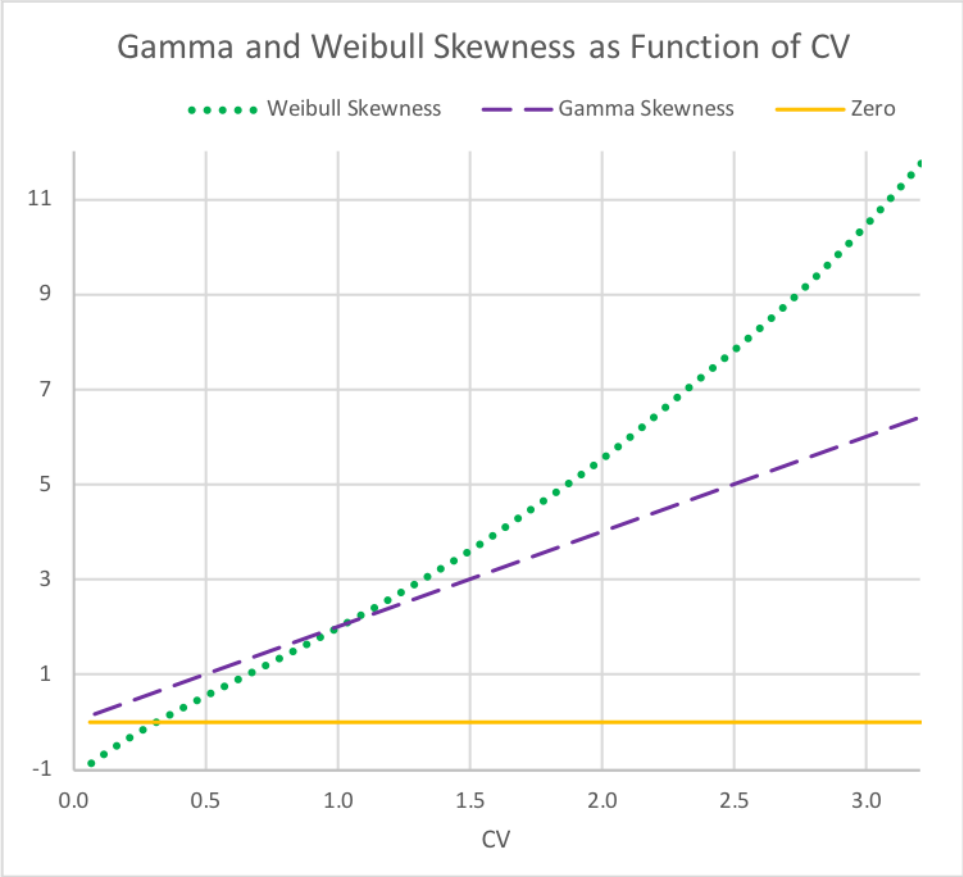


Figure 2: Weibull and Gamma Skewness

The Weibull is fairly different, which is sometimes better and sometimes not.

3 Extensions

This step includes some alternatives and extensions.

3a Interaction Terms

Add interaction terms between age and use. Give each combination its own dummy variable, just leaving out enough to prevent a singular design matrix. Many of these parameters would be expected to shrink to zero. An efficient way to eliminate some of them is classical lasso. This is easiest if run with normal distributions, so as a quick estimate it will be run on the logs of the class losses (leaving out zeros). It does the estimation for about 100 values of λ of its choosing – in a wider range than would be selected in the end. As a starting point, take a λ that is fairly low, so it does not eliminate all of the interaction variables, and use whatever interactions that remain and put them in Stan for Bayesian estimation. Then eliminate any of those that have parameter means near zero with a wide range around that. Compare to the loo of the previous model.

3b Smoothing the Factors

Keep the interaction terms if they improve the fit. Then try fitting a piecewise-linear curve to the parameters (logs of the factors) by type – age, use, or credit. That can be done by making the fundamental parameters the slope changes of the piecewise-linear curves through the original parameters. The slope changes accumulate to the slopes, which accumulate to the original parameters. Dummy variables for the slope changes thus count how many times that slope change gets added up for a particular cell. More detail will be in the report.

If this improves loo, keep that model.

3c Incorporate an Additive Component

Estimate an additive adjustment to the factor model for every rating variable that has a factor. This can be done by duplicating the design matrix and adding columns to the design matrix for each new parameter. The factors and additive terms would be estimated simultaneously. Many of them would likely shrink to zero and thus could be eliminated.

Preliminary Fitting – Simple Severity Fit Methods Comparison

For some perspective on the estimation methods, first we can look at fitting a normal distribution severity model to the simpler data set, which has 8 age groups and 4 use categories – business, long commute, short commute, and pleasure only. Regression, lasso, and Bayesian lasso estimation are compared.

Regression starts by putting all the observed data points to be fit into a column vector, and then making a design matrix with a column for each explanatory variable. Here the variables were every age group except 17-20, which is the base, and the uses except for Pleasure. There is a constant term in the models but the design matrix does not need a column for that in the packages used here. For now each variable is treated as categorical, not numerical, so the columns are just (0,1) dummy variables. For instance, the third age group would have a 1 for every observation in that group, and a 0 everywhere else.

The code for these fits is in Appendix 1. The read_excel function helps read Excel files. The data vector is a column in a file called z_small.xlsx, and the design matrix is in x_small.xlsx. These are put into a vector variable y and a matrix x. But the regression package lm needs x to be a data frame, where the lasso package glmnet and the Bayesian package stan are looking for x to be a matrix. So it has to be read in twice, depending on what you are going to do with it.

Variable selection for straight regression involves taking out insignificant variables, usually with reference to the t-statistic, which is the ratio of the variable's estimate to the standard deviation of the estimate. Usually $t > 2$ is regarded as good, but many practitioners doing

Table 1: Regression and Lasso Output

Variable	regr full	t	lasso min lam=2.2
(Intercept)	328.16	6.949	249.17983
a2	-98.50	-1.729	.
a3	-107.00	-1.879	.
a4	-112.00	-1.966	.
a5	-179.25	-3.147	-64.71981
a6	-141.75	-2.489	-27.22082
a7	-140.75	-2.471	-26.22118
a8	-144.00	-2.528	-29.47190
u2	18.38	0.456	.
u3	52.00	1.291	25.23618
u4	182.00	4.519	155.23635

actual work, as opposed to publishing, will accept a smaller t, maybe down to 1.5. Starting from this perspective, Table 1 gives the regression results for the full regression and some lasso output discussed below. All the age variables had reasonably high t-statistics, but both drive-to-work use classes had pretty low t's. The regression was re-run leaving these out.

Lasso for a normal distribution is the default setting of the glmnet package. This does the estimation for a selection it makes of up to 100 λ values. It produces a plot of the coefficient values for these λs – see Figure 3. As λ decreases, the number of parameters in the model increases (top axis), as does the L1 norm – the sum of the absolute values of the coefficients (bottom axis). Here the base use is Pleasure, so the use parameters are all positive, where the base age is the youngest, and the age parameters are all negative.

The cv.glmnet function does cross validation. It divides the dataset up into subsets – maybe 10 of them – and leaves these out one at a time and looks at how well they are then predicted by the model fit without them. The only output I am using from this is lambda.min, which is the smallest value of λ, giving the most variables in the model, deemed worth using based on their default option for k-fold cross validation. The coefficients for this value ($\lambda = 2.2$ here) are in Table 1. These are a lot smaller than the regression gives, which is due to the parameter shrinkage. All of the age groups with t<2 in the regression were left out, but the drive-long class was retained, even though the t for this was low.

Appendix 2 has the Stan code for this regression example. The data section at the top reads in variables that have been defined in the R session. Much of the code is just defining the variables, which has to be done for Stan to compile the model into C code. The prior for every parameter is assumed to be uniform over its defined range, unless otherwise specified in the model section. The log of the constant term here is uniform on [-10, 10]. The log of

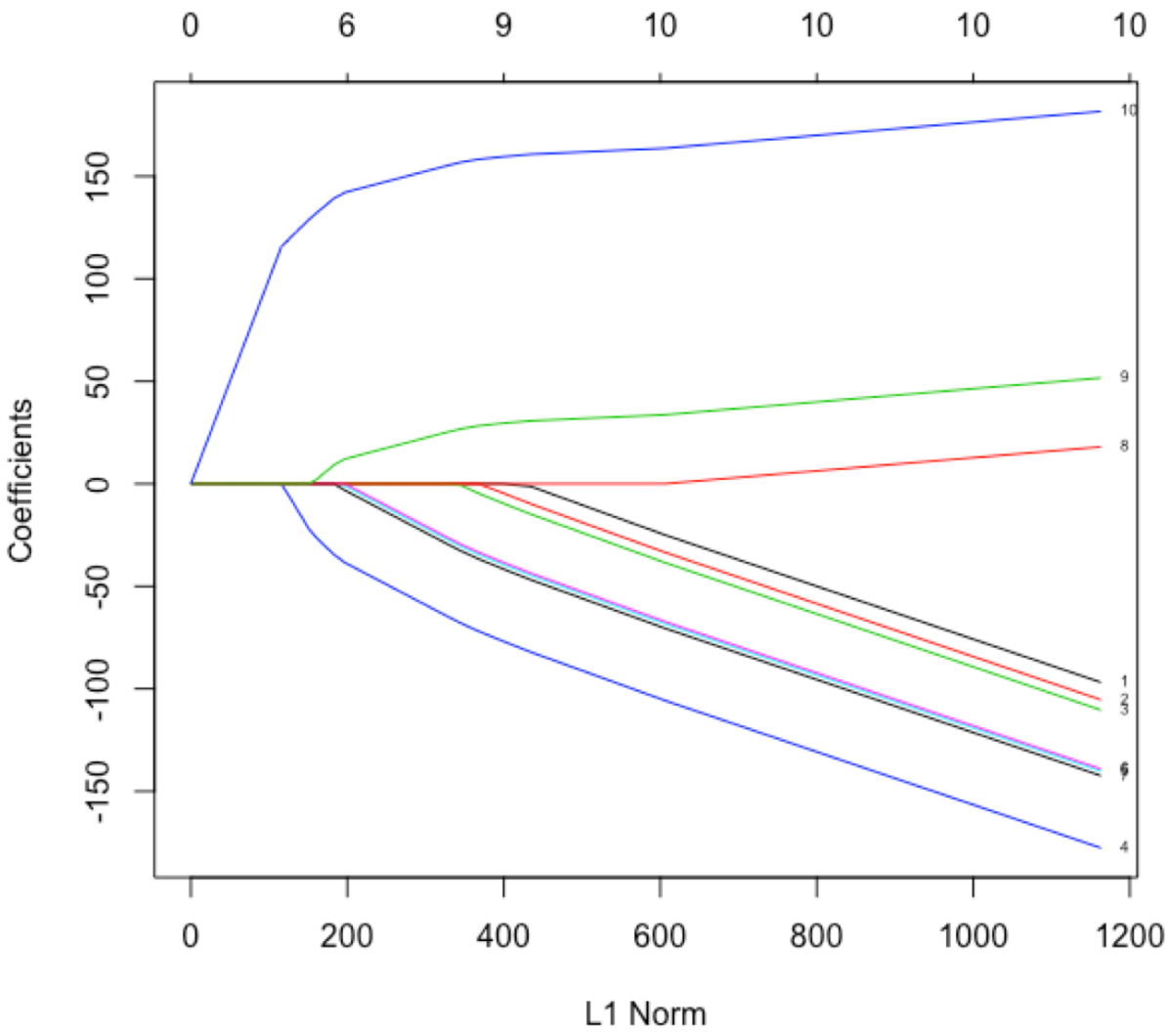


Figure 3: Lasso Shrinkage Graph

the shrinkage parameter s = $1/\lambda$ is made to be positive just because there were convergence problems otherwise. For other applications this parameter has been lower.

For positive parameters, I prefer to start with a distribution proportional to $1/x$. This diverges both as x gets small and as it gets large, so it gives balancing strong pulls up and down, whereas a positive uniform distribution diverges upwardly only. As an example where the integral is known, consider a distribution for the mean β of a Poisson distribution with probability function proportional to $e^{-\beta}\beta^k$. With a uniform distribution for β, which is proportional to 1, the conditional distribution of β given an observation k is also proportional to $e^{-\beta}\beta^k$, which makes it a gammma in $k+1$ and 1, with mean $k+1$. But if the distribution of β is assumed proportional to $1/\beta$, the conditional is proportional to $e^{-\beta}\beta^{k-1}$, which makes it a gamma in $k, 1$, with mean k. Thus the $1/\beta$ unconditional distribution takes the data at face value, whereas the uniform pushes it upwards. Numerical examples with other distributions find similar results. In practice, giving the log of a variable a uniform prior gets the same result but is slightly easier to implement.

A big advantage of Stan, and MCMC in general, over lasso is that it comes with a penalized-likelihood goodness-of-fit measure, loo. This is a cross-validation measure. It calculates the NLL for every point given a fit that used all other data points but not that one. This gives a good estimate of what the NLL would be on an entirely new sample – the population NLL as opposed to the sample NLL. This is a measure of the predictive power of the model.

Although MCMC does not eliminate parameters the way lasso does, it outputs range estimates for every parameter. If a parameter mean is close to zero, with a wide range, it is a candidate for removal from the model. I first try leaving these parameters out then seeing the effect on the loo measure. If it is better, or at least not worse, I leave them out. Eliminating the parameters with means near zero does not usually improve loo very much, but it does simplify and clarify the model. The main point is not to eliminate parameters that are improving the predictive accuracy. This exercise eliminated all but three variables – driver age 35-39, long drive to work, and business use.

This was fewer than lasso had, but not really that different. Lasso also had a small effect for ages above 39. Still neither had much age impacts except for 35-39. The predictive value of the other ages was found to be low by these approaches. This data is for physical damage severity, which might not vary by age as much as frequency does. Also value of the vehicle is not controlled for, and the ages with better drivers could also be those with more expensive cars. Vehicle value could also be a factor in the higher severity for business use.

Figure 4 shows the actual and fitted values for each age-use cell, on a log scale. There is a

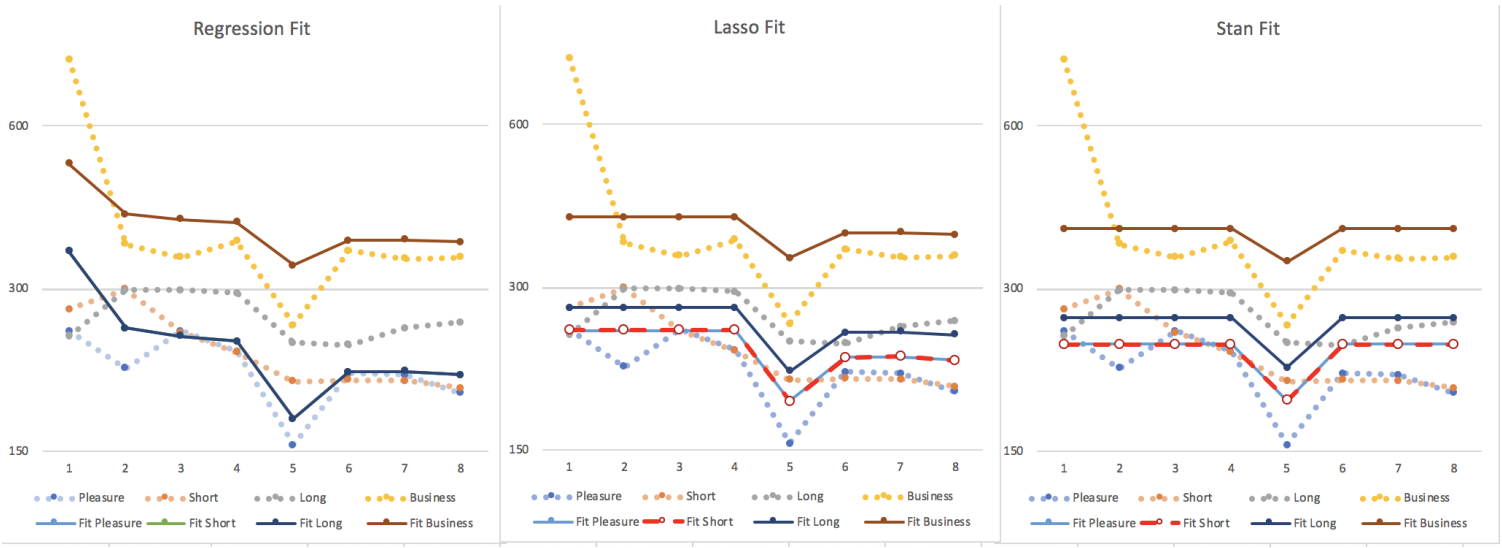


Figure 4: Severity Fits by Age and Use

line for each use class – with dotted lines for the actual data, and the age groups are on the x-axis. Each estimation method gets its own panel. One sort of outlier point is key to watch – business use for ages 17-20. This had average severity of almost 800 on 5 claims. The next highest average severity was 367 for business use for ages 30-34, with 169 claims. In later models, the number of claims will be used as input for the variance of the severity numbers, but that is not part of this model. So it is interesting to see how the different estimation methods are influenced by this cell.

The regression estimate combines the other three use classes. It gives the 17-20 group a higher severity overall, even though this does not show up in the other uses. The lasso and Stan estimates combine pleasure and short drive uses, but keep the long drive class higher. Neither makes the age 17-20 age higher than those near it. Apparently there is no predictive value for omitted observations in doing so. All the estimates make the age 30-34 group lower. You can see that the lasso estimates are a bit lower for the older ages, while Stan is not. The lasso λ was not really optimized – there is no real way of doing this – so it is hard to evaluate what it is doing. The regression estimates get lower in general for older ages.

Once the variables that improve predictive accuracy have been identified by loo, it is usually possible to find a model with just those variables by regression or lasso. For regression, this just requires using those variables. For lasso, the coef function shown in the code gives the implied coefficients for any value of λ input. Lasso and Stan often rank coefficients the same, so that some value of λ in lasso will give the same variables as the method outlined for Stan. The lowest value that chooses the same variables in this case is 4.3. This has the least

Table 2: Stan Regression, and Lasso Coefficients for Three Variable Model

	Stan	Regression	Lasso
Constant	236.2	231.6	243.3
a5	-50.3	-73.1	-33.7
u3	28.0	42.8	8.4
u4	151.0	172.7	138.4

shrinkage among the λs giving those variables. This has more shrinkage than the Stan model, whereas the regression has less shrinkage. Table 2 shows the coefficients for each model. The coefficients here tend to be positive, so the more shrinkage the coefficients get, the higher the constant is.

The mean s parameter in the Stan fit is 121, which does not give a lot of shrinkage. It might be equivalent to 0.8 for glmfit, where lambda.min was 2.2. But the variables were selected by cross validation with loo by taking out any variables that did not make loo worse. The remaining variables thus all improve the predictive value of the model. Then the shrinkage was determined by the Bayes estimate for s in the model with those variables. The shrinkage actually was less once more variables were eliminated. Still it shows some shrinkage compared to the regression. Lasso, on the other hand, needed more shrinkage to get down to those specific variables, and this is reflected in the lasso coefficients.

A couple of take aways from this are first, that the t-statistic in regression does not select variables that stand up under cross validation. Both Stan and lasso, with different cross-validation methods, eliminated a4, which has a t of -2.0, and they both kept u3, with a t of 1.3. Second, while lasso both shrinks coefficients and selects variables, these two tasks are not as compatible as they might seem. Stan with loo uses cross validation to show the useful variables, and then shrinks by the posterior of the shrinkage level s. Lasso determines both variable selection and shrinkage by λ alone, which does not allow this flexibility – and without a clear way to determine λ in the first place.

From here on I will use Stan for the fitting. Recall that this is a kind of credibility approach – estimates are shrunk towards the overall mean. Here that is based on predictive accuracy as opposed to the variance components that credibility uses, but the results are similar.

Distribution Choices

Stan is also quite flexible on distributions. As an example, I fit a gamma instead of normal to this same severity data. The gamma in a, b with mean ab and variance ab^2 can start with a b_j parameter for every cell with a fixed, which makes the variance proportional to the mean squared, or with an a_j for every cell and b fixed, which makes the variance proportional to

the mean. With the same variables as above, the b form had a loo of -185.8, and with a fixed it was -180.5. These are both considerably better than the value of -197.8 for the normal. The a form with variance proportional to mean is the better of the two. The fitted values did not change a lot – the main effect was getting a better distribution around the mean.

It is not necessary to fix either a or b across the cells. For instance, instead of fixing one of these parameters, use two parameters h, k to model variance = h*mean^k. Use the model to fit the mean for the cell and to estimate h, k, and use these to compute the variance for each cell. Moment matching for the gamma gives b = variance / mean, and a = mean^2 / variance, which give the resulting gamma parameters for each cell. Trying this gave $k = 3.6$ and a loo value of 176.6. This is a high value for k and may be arising from trying to get a higher probability for the one outlier cell, which had a high mean but still a large difference from the data. The model is not right in that the variance of the severity mean for a cell is related to the number of claims, and this is not in the model. I will include that in the next section.

Warmup – Severity Distributions on Age-Use Data

Now we will look at adding in claim counts by cell. This model will fit the severity distribution. This gives more than parameter ranges – it could be used for pricing deductibles, for instance. If μ, σ^2 are the severity mean and variance, and there are N claims, the sum of those claims has mean = $N\mu$ and variance = $N\sigma^2$. The sum divided by N has mean $m = \mu$ and variance $s^2 = \sigma^2/N$. These ratios are the data given for each cell. If the claim severity is normally distributed, m, s^2 are the mean and variance of the normal distribution for that data. If the severity is from a gamma distribution in a, b, the sample mean is also gamma distributed, in $\alpha = Na, \beta = b/N$. Either way, the collection of sample means can provide an estimate of the severity distribution parameters. The inverse Gaussian distribution has similar formulas.

The two main things the Stan code requires are the prior distributions of the parameters and the conditional distribution of the observations given the parameters, which is actually the likelihood. The parameters for the age and use classes will have shrinkage priors. We used the double exponential prior in s for this before, with non-shrinkage priors on s, the constant term, and the distribution parameters $a, b, \sigma, h, k, ...$ The Cauchy shrinkage prior is often a bit more efficient than the double exponential, so it will also be used here. It is just the Students-t distribution with one degree of freedom. It has more weight concentrated around zero, but also heavier tails, and this can sometimes more readily distinguish the important contributing variables.

The conditional distributions here are the normal or gamma for the sample mean, although the parameters will be those for the severity, with the claim count as additional information.

The way Stan works is that you have to give it the parameters for every observation. You can do transforms on the severity parameters to get the distributions for the observed sample means. One of the parameters is calculated by cell as the constant plus the design matrix times the parameter vector. In formulas, this is mu = x*v+cn. You can then transform mu by cell to give the parameters for the sample mean. Consider the gamma model with $\sigma^2 = h\mu^k$ for instance. Solve for cell j to get: $a_j = \mu_j^2/\sigma_j^2$, and $b_j = \sigma_j^2/\mu_j$. Then $\alpha_j = N_j a_j$ and $\beta_j = b_j/N_j$ are the gamma parameters for the observed sample mean. The code does these assignments for every cell, then in the distribution statement uses the vector form, like y ~ gamma(alf, bet). (Stan actually defines the gamma distribution with b as what is $1/b$ here, and σ, not σ^2, as the normal parameter, but that will be an adjustment in the code. The text will continue to use the more conventional forms.)

All the variables now are retained in the model, except for the short drive use class. Leaving that variable out combines short drive and pleasure uses. Getting a better model for the cell variances seems to have made the age variables more predictive. The power k came out 2.4, suggesting the variance is proportional to the mean raised to the 2.4 power. This is between 2, which the gamma with fixed a gives, and 3 for the inverse Gaussian. I tried just using a straight gamma for comparison. This eliminates k as a parameter, and loo came out a little better that way, so this model was selected. Still, the selection of the gamma is in part a parameter selection, and that implied parameter is not counted, so keeping $k = 2.4$ might really be just as good. However taking the variance proportional to mean squared is a pretty standard assumption for severity distributions anyway. It makes b a scale parameter, for instance.

The loo measure was -155.7 with the double exponential prior and -156.2 with the Cauchy – both considerably better than in the model without counts. The double-exponential is used below. Figure 5 shows the actual and fitted severity sample means using the gamma fit. The business use class for drivers under 21 did not seem to influence the fit much. With the low claim count for this cell, the variance of the observed sample mean came out high, so the likelihood was better at that point even with a low fit mean. In fact it was 5000 times as high as it was in some of the earlier models.

This was for an additive model. It is easy enough to do a multiplicative model from the same code – just exponentiate the severity mean and make that the severity mean of the gamma, then adjust to make it the sample mean. This starts by fitting an additive model for the log of the losses. Doing this gave virtually the same loo as the additive model. All the parameters, including shrinkage parameter s come out with a lot smaller absolute values. For s – its mean here is 0.17.

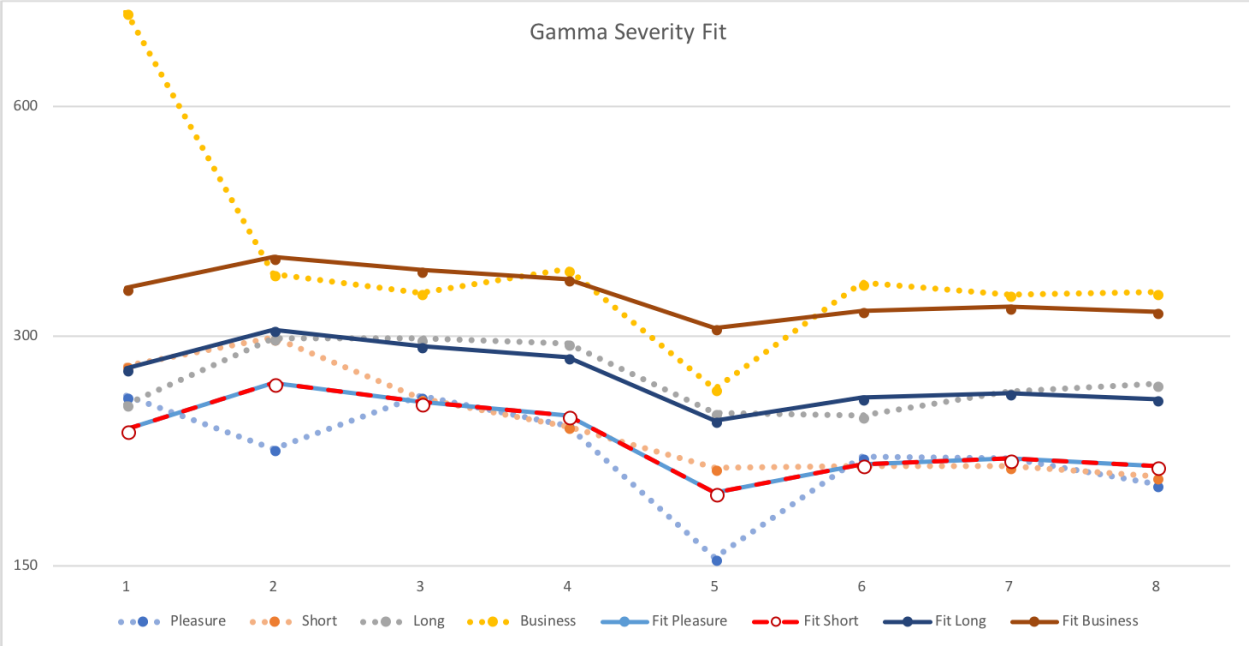


Figure 5: Gamma Severity Using Counts

I also tried the normal distribution but forcing the variance to be proportional to the mean squared. I just made the standard deviation a multiple of the mean to do this. Loo for this model came out -166.6, which is much better than previous normal fits, but not as good as the gamma. The s.e. for comparing this to the gamma is 4.6 by the loo compare function, so the gamma is more than two standard deviations better. A bit of skewness is apparently needed for this data. One way to test for convergence of a model fit is with a measure called Rhat that shows in the print function. It is a ratio of total to within variances among the chains for each parameter. It should be close to 1.0 for a model that converges. It was around 3 for most of the variables in this model in a preliminary run. Forcing the log of the s parameter to be positive restricted the model enough to get convergence.

Inverse Gaussian Distribution

The usual parameterization of the inverse Gaussian is designed for quasi-likelihood estimation, but it is awkward in applications. A more natural parameterization uses parameters a, b with mean ab and variance ab^2, like for the gamma. It then has some other properties of the gamma: the sum of a sample of N claims is IG in Na, b and multiplying by a constant c gives an IG in a, bc. The skewness is 3CV, where CV is the coefficient of variation = standard deviation / mean. The gamma skewness is 2CV, so the IG is more or less a slightly more skewed gamma. Its shape can be a bit different, however. For $a < 1$, the gamma density goes to infinity at $x = 0$, but the IG density is 0 at $x = 0$ – and in fact the density at zero has

slope 0 as well, so grows slowly at first. The density is:

$$f(x|a,b) = \sqrt{\frac{a^2 b}{2\pi x^3}} exp\ \ \frac{-(x-ab)^2}{2bx}\Bigg)$$

Stan does not have this distribution, but it has a provision for adding user-defined functions. This goes in the function block at the start of the program. Here is one I did for the IG:

```
functions{
  real ig_lpdf(real y, real a, real b){
    return log(a/b)-0.5*log(2*pi())-1.5*log(y/b)-b*(y/b-a)^2/2/y; }  }
```

It runs reasonably fast. It gave loo of -157.7, the same as for the gamma. Possibly the best skewness is in between the IG and gamma.

Larger Data Set – with Credit Variables

The larger data set includes four levels of credit scores, with 4 being the best. There are eight age groups and four use classes. The data for each cell is exposure, pure premium, and their product – losses. The model is for pure premium, as a product or sum of a constant and the age, use and credit parameters. The dependent variable is taken as losses, so the modeled pure premium is multiplied by the known exposure by cell to give the expected value of the cell losses.

The initial model assumes losses are gamma distributed with fixed b, so with variance proportional to mean. I tried this for additive and multiplicative versions, and in this case the multiplicative had a clearly larger loo fit measure. The Cauchy and double-exponential priors gave virtually identical fits. I also tried a gamma with fixed a, which makes the variance proportional to the mean squared. This was slightly worse. Then I used a gamma with variance = $s*\text{mean}^k$. The k came out at 1.3, with the loo just about the same as the fixed b version. The better power was not quite worth the extra parameter.

I also tried normal, inverse Gaussian, and Weibull residuals. The Weibull, using variance = $s*\text{mean}^k$, had a loo of -1374.5, which was slightly better than the gamma's -1375.9. The others were all worse fitting. The extensions below use the gamma, as it is faster to estimate.

This form of the Weibull requires a non-linear solver to match the moments. Once the mean and variance have been fit for a cell, the parameters a, b can be fit by solving $EX = a!b$, $1 + CV^2 = (2a)!/a!^2$. The non-linear part is solving for a from the CV. Stan has a facility for solving a system of linear equations with something called the algebraic solver, but the format is picky. For this, you first set up a function. This is what I used:

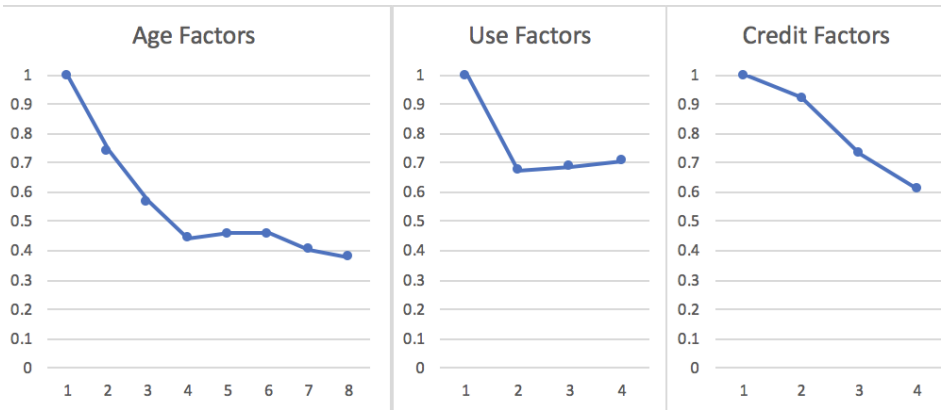


Figure 6: Class Rating Factors

```
functions {  vector system(vector alpha, vector Q, real[] x_r, int[] x_i){
    vector[118] z;
   z = lgamma(1+2*alpha) - 2*lgamma(1+alpha) - Q;
   return z; }  }
```

Q gets $log(1+CV^2)$. The variables x_r and x_i are zero dimensional arrays that don't seem to have any role, but are required. When it solves for $z = 0$, alpha is the a that matches this CV. There are 118 data points. To call it, I used:

```
alpha = algebra_solver(system, start, Q, x_r, x_i );
for (j in 1:N) {   lam[j] = mu[j]/tgamma(1+alpha[j]);
  alpha[j] = 1/alpha[j];      }
```

This solves for the vector of a values for all the observations. Stan's Weibull function uses $1/a$ instead of the a in the above distribution function.

Ten cells with small exposure and zero losses were omitted from the fitting. These distributions are not defined at zero. There were still some cells with small exposure and volatile loss numbers. Mildenhall suggests reducing a cell's assumed variance by dividing it by exposure^k for some $k > 0$. This would allow for smaller exposure cells to be more volatile. For some reason, the estimated k for this data set came out as -0.1, so variance would increase slightly for the larger cells. This gave the same loo as without this adjustment, and it doesn't make much sense, so it was omitted.

Figure 6 shows the factors for each of the rating variables in the gamma model. The use classes are Business, Drive Long, Drive Short, and Pleasure. Figure 7 shows actual vs. fitted averages for use and credit groups by age for this and the combined additive-multiplicative

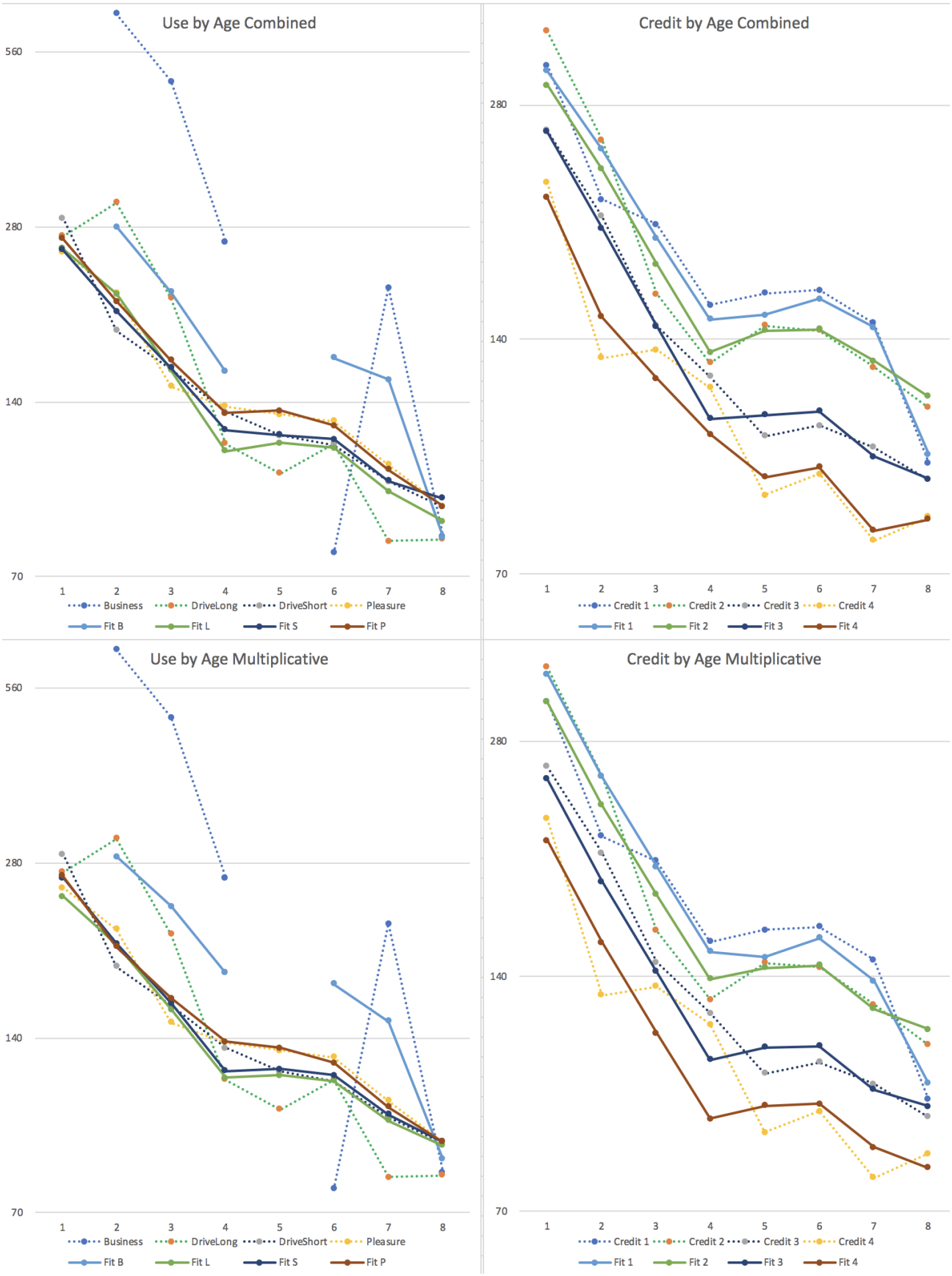


Figure 7, Fitted Class Group Averages

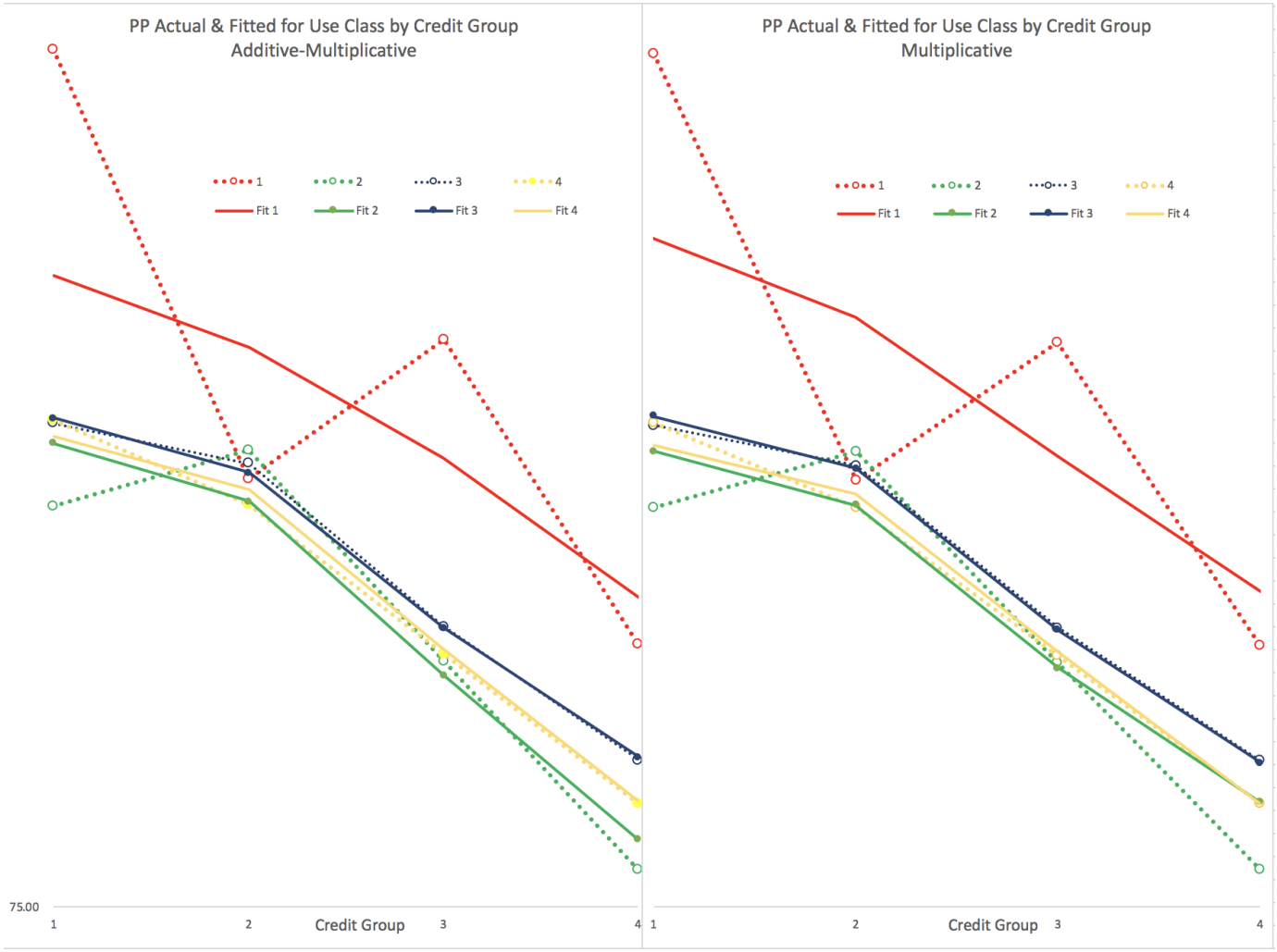


Figure 8: Actual and Fitted Use Class Averages by Credit Group

model discussed below. Figure 8 shows use and credit cell actual and fitted averages over the age groups for these two models. The omitted zero losses show as gaps in some of the graphs.

Use 1, Business, has the smallest volume and is most volatile. This shows up in both the use by age and use by credit graphs. The fitted values do not appear over-responsive to these fluctuations. The Bayesian shrinkage looks like it is doing a lot what credibility would do in giving low weight to those points. The drive long class and the good credit group 1 have poor fits at a few points.

Extensions

Multiplicative-Additive Model

The possible issue with multiplicative models is that cells that have high or low factors in two directions might be over or under estimated by the product of the factors. Here that does not seem to be a problem, in that the estimated values look to be less extreme than the data points. But as an example, I try a model that starts with all the variables used as both factors and additive adjustments. Then variables whose parameters are shrunk close to zero are omitted and the model refit, iteratively, until the best combination is found.

The fitted value for a cell is the constant times the rating factors for the cell plus the additive levels for the same age, use, and credit variables. To estimate it, there are two coefficient vectors, say v and w, and the design matrix is used twice. Call the two instances x and xa. Then the mean for all the cells is the vector mu = exp(xv+cn)+xaw. Different variables probably will be eliminated from x and xa based on parameters being shrunk towards zero. The shrinkage parameter s was set somewhat arbitrarily as 100 times greater for the additive parameters, as it came out in this ballpark in previous fitting. It might be better to have separate priors for these two shrinkage parameters.

Doing this eliminated the factors for age group 2 and credit group 2. Additive levels were fit for all but ages 3 and 8 and use 3, which is drive short. There were thus 21 rating variables in this model, compared to 13 for the multiplicative model. The loo penalized likelihood measure was -1374.2, which is an improvement over -1375.9 for the multiplicative model. The difference of 1.7 is usually considered worthwhile for penalized likelihood. The parameter penalty was 23.1 here, compared to 16.8 for the multiplicative model. This is 7.3 higher for 8 more variables, so less than 1 per variable. This is due in part to shrinkage, but formally due to better predictive accuracy of the larger model. The likelihood was thus higher by 9.0 for the combined model, which is a better fit to the data, but the penalized likelihood is the proper comparison.

Figure 9 shows the factors and additive levels for the variables. They offset each other to

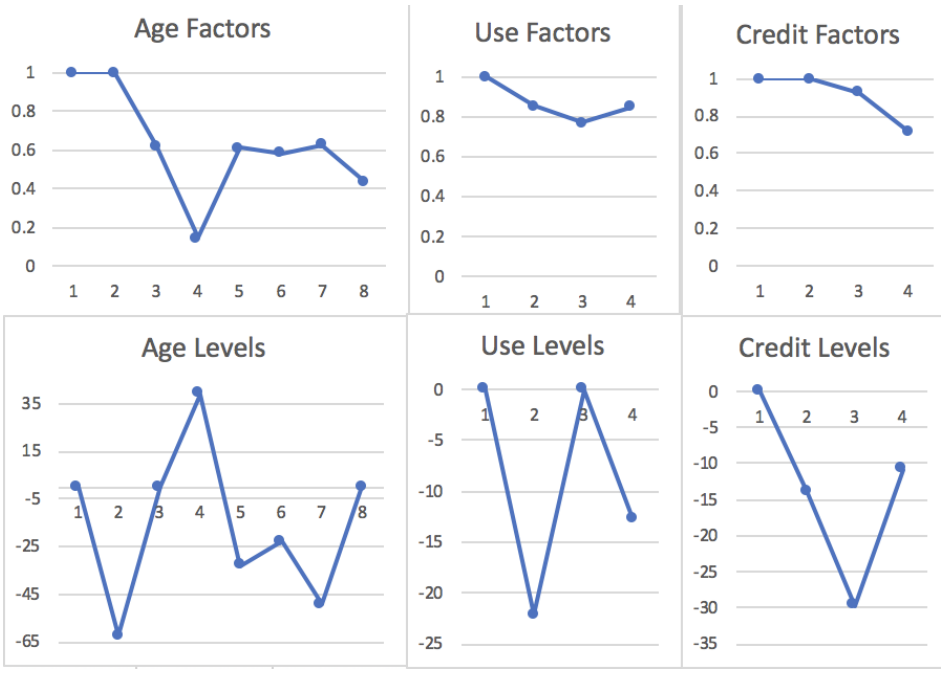


Figure 9: Class Rating Factors and Additive Levels

a degree, as some variables seem to work better additively, and some multiplicatively. The fitted values have small but observable changes. In Figure 7, use groups long and short appear to have slightly better fits across the ages, as do credit groups 2 and 4. In Figure 8, use groups 2 and 4 (long and pleasure), fit a bit better across the credit groups.

This model is pretty intuitive and is easy to fit with MCMC.

Interaction Terms

There may be some combinations of rating elements that interact differently than the overall model. Suppose age 2 and credit 3 is such a combination. Then adding a variable for that combination could pick up the interaction. Since the variables are all (0,1) dummies, the interaction variable would just be the product of the individual variables. There are four observations with that combination – one for each use class. If the variable improved the fit for all four, that would suggest the variable is significant.

Random effects is well set up for estimating this kind of thing, and that is one of its prime uses. You could put in all combinations of two-way interactions, and many of the coefficients would go to zero. If all the variances are the same, this would give lasso or ridge regression, depending on the distribution assumed. Bayesian shrinkage can do that too. Lasso is a good starting point, as it completely eliminates a lot of variables.

I tried interaction focusing on use classes. Each combination of the use variables with age

Table 3: Interaction Factors Age(a), Use(u) and Credit(c)

u2,a2	u2,a7	u3,a2	u3,a3	u4,a3	u4,a4	c4,a2	c4,a4	c4,a5	c4,a6	c4,a7
1.25	0.84	0.84	0.72	0.65	0.77	0.82	1.15	0.89	0.92	0.84

and credit was given a variable, which was a product of the individual variables. The glmnet package is easiest to apply to a normal regression, so I made the dependent variable the log of the cell pure premium, so the regression would give a multiplicative model with lognormal residuals. The least suggested shrinkage, given by cvfit.lambda.min, was for λ about 0.0035. I used that and $\lambda = 0.005$ to review the eliminated variables. Selecting all but those with very small coefficients gave ten combinations. This reduced the interaction variables from 30 to 10.

I put those in the gamma regression in Stan for the multiplicative model and eliminated the ones with small coefficients as long as so doing did not make loo worse. That left six interaction variables, and this resulted in a loo of -1367.0, which is the best result so far, and a fairly big improvement. I also looked at the credit-age interactions, and it looked like credit group 4 (best credit) had the most issues with age interactions. So I added in all seven of those interaction variables, ran Stan, and then again eliminated the non-contributing variables. That left five of those, so eleven interaction terms altogether. This brought the loo measure down to -1366.2, so the credit interaction helped a little, but not much.

Table 3 shows the adjustment factors for these interactions. Figure 10 graphs the resulting use by age and credit by age fits. The biggest improvement seems to come for use 2, drive long, particularly at ages 2, 7, and 8. Credit group 4 looks better than the multiplicative model, but about the same as the additive-multiplicative model. Pleasure use has a strange worsening of the fit at age group 4.

Fitting Curves to Factors

The factors all change fairly gradually across a rating class, so it might be possible to save on parameters by fitting curves to them. A flexible and easy way to do this with parameter shrinkage is to fit piecewise-linear curves to the factors. With no shrinkage, this would replicate the exact parameters. If you do shrinkage on the slope changes between segments, you smooth out the curves, with the degree of smoothing determined by the shrinkage methodology.

I try this for the three curves in the multiplicative model, as in Figure 6, but fitting the curves on the logs of the factors. Starting with zeros for each rating class, the slope changes add up cumulatively to the slopes, which in turn add up to the rating factor logs. This is all

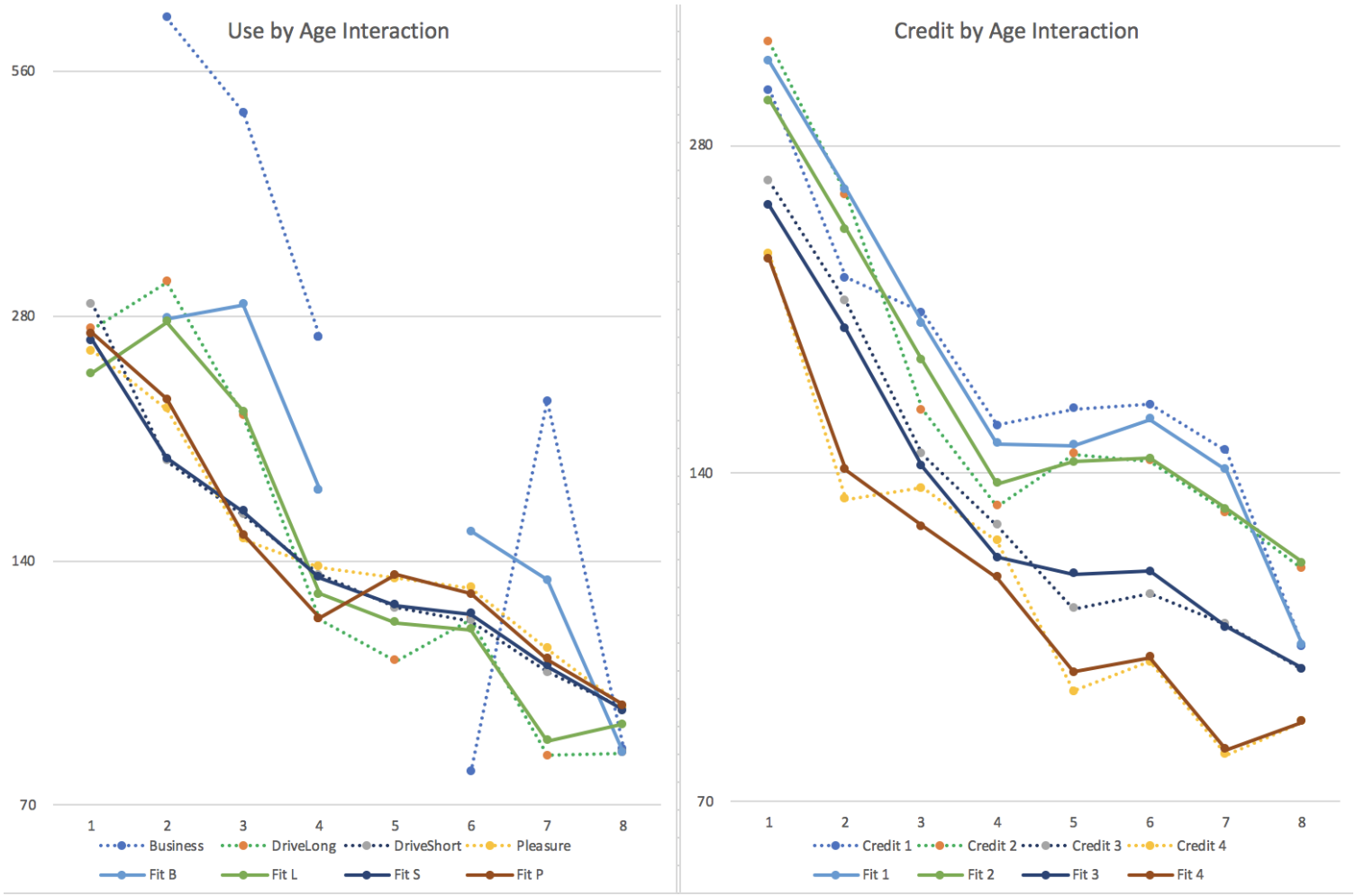


Figure 10: Actual and Fitted Class Group Averages

linear, so can be accomplished with a design matrix. For a cell in class number i in one of the directions, the dummy variable for class u in that direction gets the value $max(0, 1 + i - u)$. The fitted means are still the vector mu = exp(x*v+cn), but now v is the vector of slope change coefficients.

The ages 3, 6, and 8, and credit 4, all got zeros and so were eliminated from the model. A zero slope change just continues the previous line segment. It does not make the log factor zero. The factors all came out very similar to before, but a little smoother. The loo was -1373.3, which is a fair bit better than the -1375.9 for the straight multiplicative model. The main improvement was in the parameter penalty of 12.9, compared to 16.8. The slope-change model is apparently more parsimonious. Figure 11 shows the credit-by-age and use-by-age fits. Some points fit better and some worse than previous models, but the fitted values are on straighter lines, which is related to the model being more parsimonious. Some lines are not parallel, due to differences in mix.

Etc.

The three model enhancements here - additive-multiplicative, interaction terms, fitting curves - can be combined, but methodologically would just repeat what's above.

Summary

Bayesian shrinkage is an improved alternative to maximum likelihood. It has lower estimation and prediction errors, and unlike frequentist shrinkage it comes with a goodness-of-fit measure. It also can use the posterior mean. MCMC software, like Stan, also makes it easy to fit more generalized distributions.

I first used this to fit severity data by class. A few distributions like the normal, gamma, and inverse Gaussian have a known connection between the distributions of the claims and of the sample means, and this allows estimating the severity distributions from the sample means. Then I tried it on a bigger data set with more classes on aggregate losses and pure premium. This produced class factors. Extensions were a combined additive-multiplicative model, interaction terms, and fitting curves to the factors. All of these improved the fits.

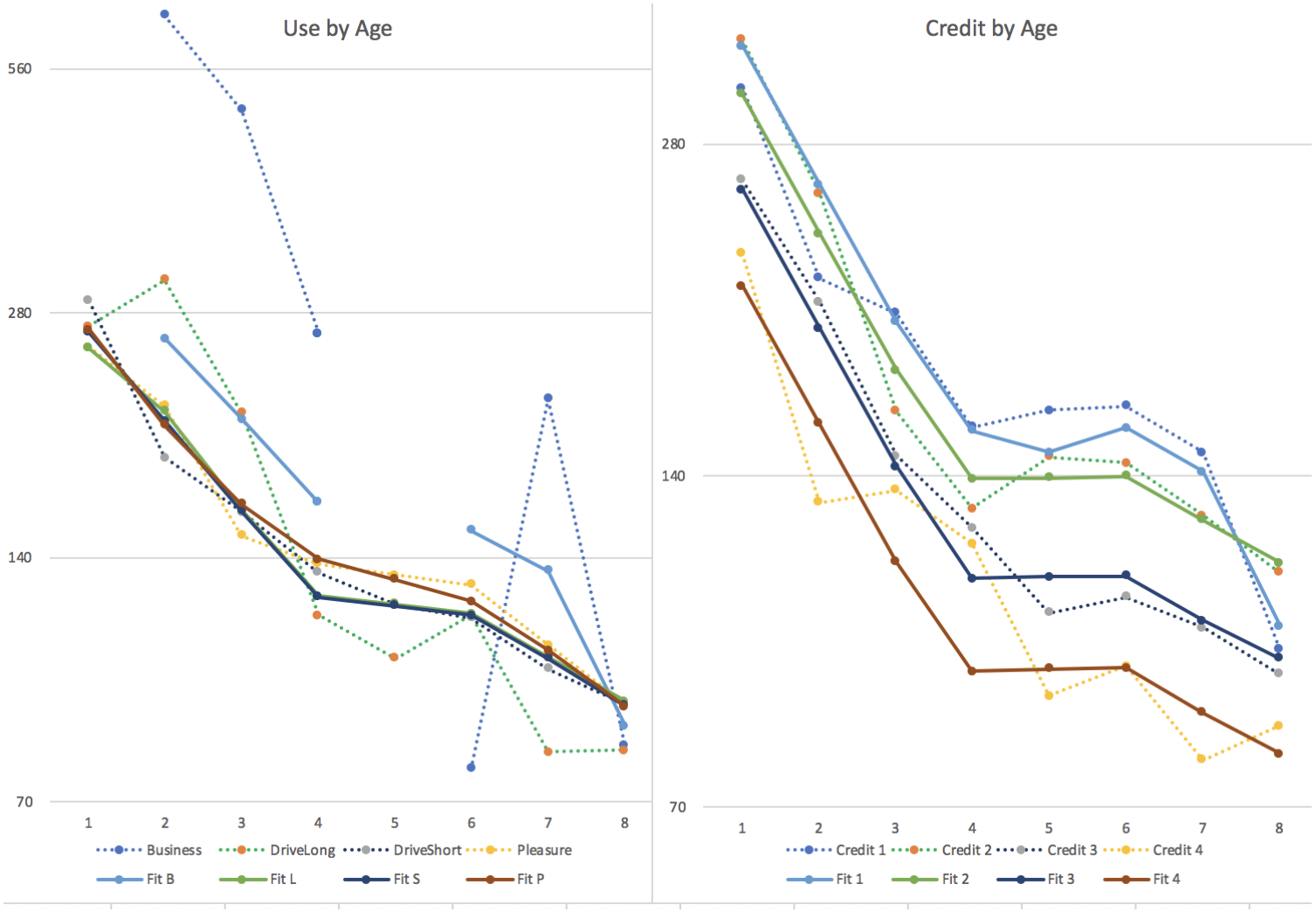


Figure 11: Actual vs. Fitted for Curve-Fit Model

Appendix

1 R code for regression, lasso packages and to feed Stan

```
setwd("~/OneDrive/R/Ratemaking Stan/Severity regression")
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
library("loo")
library(readxl)
library(glmnet)

y = as.vector(as.matrix(read_excel("z_small.xlsx")[,5]))
x = as.matrix(read_excel("x_small.xlsx"))  #do this or the next line
x = read_excel("x_small.xlsx")  #regression function needs x a data frame
U = ncol(x)
N = length(y)
c(N,U)

mod <- lm(y ~ ., data = x)   #full regression
mod <- lm(y ~ ., data = x[c(1:7,10)])  #regression leaving out two uses
summary(mod)  #gives output

fit1 = glmnet(x, y, standardize = FALSE)   #lasso fit
plot(fit1, label=TRUE)
cvfit = cv.glmnet(x, y, standardize = FALSE)  #lasso cross validation
cvfit$lambda.min   #lowest lambda suggested by cross validation
coef(cvfit, s = "lambda.min")  #shows parameters for that lambda

x_full = x
x = as.matrix(x_full[,c(4,9,10)]) #just using selected columns
U = ncol(x)
U

fitsev = stan(file = 'sevregr.stan', verbose = FALSE, chains = 7,
              iter = 7000, warmup = 2000)

log_LL <- extract_log_lik(fitsev)
```

```
loo_LLsev <- loo(log_LL)
loo_LLsev

print(fitsev, pars=c("cn", "v", "s"), probs=c(.05, 0.2, 0.5, 0.8, 0.95),
      digits_summary = 3)
plot(fitsev, pars = c("v", "s"))

out <- get_posterior_mean(fitsev)
write.csv(out, file="out_sregr.csv")
```

2 Stan code for normal regression

```
data {
  int N;           //number of observations
  int U;           //number of variables
  vector[N] y;     //the dollar losses in a column
  matrix[N,U] x;  //design matrix with U columns
}
parameters {  // all except v will get uniform prior, which is default
 real<lower=-10, upper=10> logcn;           //log constant term
 vector[U] v;                            //the parameters
 real<lower=0, upper=10> logs;     //log of s, related to lambda
 real logsig;   //log of sigma parameter
}
transformed parameters {
 real cn;
 real sig;
 real s;                          //shrinkage parameter, like lambda
 vector[N] mu;                 //fitted means
 vector[N] sigma;
 cn = exp(logcn); //for positive parameter, uniform on log is like 1/X
 sig = exp(logsig); //for positive parameter, uniform on log is like 1/X
 s = exp(logs); // Gives more weight to lower values; good if X not big
 mu = x*v+cn; //vector of mu parameters
 for (j in 1:N) sigma[j]=sig; //Stan normal has sigma not squared

}
```

```
model {  // gives priors for those not assumed uniform. This one for lasso.
   for (i in 1:U)  v[i] ~ double_exponential(0, s);
   y ~ normal(mu, sigma);
//for (j in 1:N) y[j] ~ normal(mu[j], sigma[j]);
}
generated quantities { //outputs log likelihood for looic
  vector[N] log_lik;
for (j in 1:N) log_lik[j] = normal_lpdf(y[j] | mu[j], sigma[j]);
}
```

3 Stan code for gamma-alpha regression

```
data {
  int N;           //number of observations
  int U;           //number of variables
  vector[N] y;     //the dollar losses in a column
  matrix[N,U] x;  //design matrix with U columns
}
parameters {  // all except v will get uniform prior, which is default
 real<lower=-10, upper=10> logcn;           //log constant term
 vector[U] v;                           //the parameters
 real<lower=0, upper=10> logs;     //log of s, related to lambda, not too high
 real logalpha;   //log of beta parameter
}
transformed parameters {
 real cn;
 real alpha;
 real s;                          //shrinkage parameter, like lambda
 vector[N] alf;                 //fitted means
 vector[N] beta;
 cn = exp(logcn); //for positive parameter, uniform on log is like 1/X
 alpha = exp(logalpha); //for positive parameter, uniform on log is like 1/X
 s = exp(logs); // Gives more weight to lower values; good if X not big
 for (j in 1:N) alf[j]=alpha; //Stan gamma mean = alpha/beta
 beta = alf ./ (x*v+cn); //vector of beta parameters
}
model {  // gives priors for those not assumed uniform. This one for lasso.
```

```
   for (i in 1:U)  v[i] ~ double_exponential(0, s);
   y ~ gamma(alf, beta);
}
generated quantities { //outputs log likelihood for looic
  vector[N] log_lik;
for (j in 1:N) log_lik[j] = gamma_lpdf(y[j] | alf[j], beta[j]);
}
```

4 Stan code for gamma-k regression

```
data {
  int N;           //number of observations
  int U;           //number of variables
  vector[N] y;     //the dollar losses in a column
  matrix[N,U] x;  //design matrix with U columns
}
parameters {  // all except v will get uniform prior, which is default
 real<lower=-10, upper=10> logcn;           //log constant term
 vector[U] v;                           //the parameters
 real<lower=0, upper=10> logs;     //log of s, related to lambda, not too high
 real<lower=1.0, upper=6> k;
 real  logh;   //log of h parameter
}
transformed parameters {
 real cn;
 real h;
 real s;                          //shrinkage parameter, like lambda
 vector[N] m;                 //fitted means
 vector[N] V;
 vector[N] alf;
 vector[N] bet;
 cn = exp(logcn); //for positive parameter, uniform on log is like 1/X
 h = exp(logh);
 s = exp(logs);       // Gives more weight to lower values, which is good if X not big
 m = x*v+cn; //vector of means
 for (j in 1:N) { V[j] = h*m[j]^k;
  alf[j]= m[j]^2/V[j];
```

```
  bet[j]= m[j]/V[j];}
}

model {
 // gives priors for those not assumed uniform. Choose this one for lasso.
   for (i in 1:U)  v[i] ~ double_exponential(0, s);
   y ~ gamma(alf, bet);
}
generated quantities { //outputs log likelihood for looic
  vector[N] log_lik;
for (j in 1:N) log_lik[j] = gamma_lpdf(y[j] | alf[j], bet[j]);
}
```

References

Demoment, G. 1989. "Image Reconstruction and Restoration: Overview of Common Estimation Structures and Problems." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37:12.

Fu, Luyang, and Cheng-Sheng Peter Wu. 2007. "General Iteration Algorithms for Classification Ratemaking." *Variance* 1:2: 193–213.

Gelfand, A. E. 1996. "Model Determination Using Sampling-Based Methods." *Markov Chain Monte Carlo in Practice, Ed. W. R. Gilks, S. Richardson, D. J. Spiegelhalter* London: Chapman and Hall: 145–62.

Hoerl, A.E., and R. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12: 55–67.

Klinker, Fred. 2011. "Generalized Linear Mixed Models for Ratemaking: A Means of Introducing Credibility into a Generalized Linear Model Setting." *CAS E-Forum* https://www.casact.org/pubs/forum/11wforumpt2/Klinker.pdf.

Miller, Hugh. 2015. "A Discussion on Credibility and Penalised Regression, with Implications for Actuarial Work." *Presented to the Actuaries Institute 2015 ASTIN, AFIR/ERM and IACA Colloquia*.

Morris, Carl N., and Lee Van Slyke. 1978. "Empirical Bayes Methods for Pricing Insurance Classes." *Proceedings of the Section on Business and Economics, American Statistical Association*, 579–82.

Stein, Charles. 1956. "Inadmissibility of the Usual Estimator of the Mean of a Multivariate Normal Distribution." *Proceedings of the Third Berkeley Symposium* 1: 197–206.

Tikhonov, Andrey Nikolayevich. 1943. "On the Stability of Inverse Problems." *Doklady Akademii Nauk SSSR* 39:5: 195–98.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and Waic." *Journal of Statistics and Computing* 27:5: 1413–32.

Venter, Gary, Roman Gutkovich, and Qian Gao. 2017. "Parameter Reduction in Actuarial Triangle Models." *Variance* https://www.variancejournal.org/articlespress/articles/Parameter-Venter-Gutkovich-Gao.pdf.