2002

Journal of Actuarial Practice, Volume 10, 2002

Colin Ramsay, Editor

University of Nebraska - Lincoln, cramsay@unl.edu

Follow this and additional works at: http://digitalcommons.unl.edu/joap

Part of the Accounting Commons, Business Administration, Management, and Operations Commons, Corporate Finance Commons, Finance and Financial Management Commons, Insurance Commons, and the Management Sciences and Quantitative Methods Commons


This Article is brought to you for free and open access by the Finance Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Journal of Actuarial Practice 1993-2006 by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
ARTICLES

Communicating Effectively with Words, Numbers, and Pictures: Drawing on Experience
Karolina Duklan and Michael A. Martin ........................................................ .......................... 1

Unearned Premiums and Deferred Policy Acquisition Expenses in Automobile Extended Warranty Insurance
Joseph Cheng .......................................................................................................................... 6

Can Utility Maximization Models Assist With Retirement Planning?
Zaki Khorasanee .................................................................................................................. 9

Dynamic Funding and Investment Strategy for Defined Benefit Pension Schemes: A Model Incorporating Asset-Liability Matching Criteria
Shih-Chieh Chang, Cheng-Hsien Tsai, Chia-Jung Tien, and Chang-Ye Tu ........................................ 13

Model Risk and Surplus Management Under a Stochastic Interest Rate Process
Jennifer L. Wang and Rachel J. Huang ................................................................................. 15

Some Comments on the Pricing of an Exotic Excess of Loss Treaty
Jean-François Walhin ........................................................................................................... 17

Modeling Size-of-Loss Distributions for Exact Data in WinBUGS
David P.M. Scollnik ............................................................................................................. 19

Improving Mortality: A Rule of Thumb and Regulatory Tool
John H. Pollard ..................................................................................................................... 21

Further Remarks on Risk Sources Measuring:
The Case of a Life Annuity Portfolio
Mariarosaria Coppola, Emilia Di Lorenzo, and Marilena Sibillo ............................................ 22

A Note on the Parallelogram Method for Computing the On-Level Premium
David P.M. Scollnik and Wai Man Sara Lau ......................................................................... 24
EDITORIAL POLICY

The aim of this international journal is to publish articles pertaining to the “art” and/or “science” involved in contemporary actuarial practice.

The Journal welcomes articles providing new ideas, strategies, or techniques (or articles improving existing ones) that can be used by practicing actuaries. One of the goals of the Journal of Actuarial Practice is to improve communication between the practicing and academic actuarial communities. In addition, the Journal provides a forum for the presentation and discussion of ideas, issues (controversial or otherwise), and methods of interest to actuaries.

The Journal publishes articles in a wide variety of formats, including technical papers, commentaries/opinions, discussions, essays, book reviews, and letters. The technical papers published in the Journal are neither abstract nor esoteric; they are practical and readable. Topics suitable for this journal include the following:

- AIDS
- annuity products
- asset-liability matching
- cash-flow testing
- casualty ratemaking
- credibility theory
- credit insurance
- disability insurance
- expense analysis
- experience studies
- FASB issues
- financial reporting
- group insurance
- health insurance
- individual risk taking
- insurance regulations
- international issues
- investments
- liability insurance
- loss reserves
- marketing
- pensions
- pricing issues
- product development
- reinsurance
- reserving issues
- risk-based capital
- risk theory
- social insurance
- solvency issues
- taxation
- valuation issues
- workers' compensation

REVIEW PROCESS

A paper submitted to the Journal first is screened for suitability. If it is deemed suitable, copies are sent to several independent referees. The name of the author(s) of the paper under consideration is usually anonymous to the referees, and the identities of referees are never revealed to the author(s).

The paper is reviewed for content and clarity of exposition. Papers do not have to contain original ideas to be acceptable. On the basis of the referee reports, the editor makes one of the following decisions: (1) accept subject to minor revisions, (2) accept subject to major revisions, or (3) reject.

The editor sends the author(s) of the decision and along with copies of the referees' reports. The referee process is expected to take two to five months (depending on the length of the paper).

See inside back cover for instructions to authors.
Communicating Effectively with Words, Numbers, and Pictures: Drawing on Experience

Karolina Duklan* and Michael A. Martin†

Abstract†

In this paper, we discuss techniques for developing effective communication skills, focusing in particular on technical writing, the use of graphics, and presentation. The key principles of effective communication that we propose to actuaries are as follows:

• Identify your audience and consider their needs and abilities;
• Focus on substantive content;
• Choose appropriate communication tools;
• Use language that is simple, concrete, and familiar;
• Integrate text, numbers, and graphics;

*Karolina Duklan, BActS (Hons), A.I.A.A., is a consulting analyst at Frank Russell Company in Sydney, Australia. She was awarded a University Medal on completion of her bachelor of actuarial studies (Hons) degree at Australian National University, Canberra (1999). She held a position of investment analyst at Towers Perrin in Canberra from 1998 to 2001. In 2001 she commenced her current role at Frank Russell Company in the field of asset consulting.

Ms. Duklan’s address is: Frank Russell Company, GPO Box 5291, Sydney, NSW 2001, AUSTRALIA. Internet address: Kduklan@russell.com

†Michael A. Martin, Ph.D., is reader in statistics at the Australian National University (ANU). He holds a B.Sc. (Hons) degree from the University of Queensland (1986) and a PhD in Statistics from the ANU (1989). He was an assistant professor of statistics at Stanford University from 1989 to 1994 and was Annenberg Distinguished Assistant Professor in statistics at Stanford from 1992 to 1994. In 1994, he accepted a position as lecturer in statistics at the ANU, where he became a senior lecturer in 1995 and reader in 2002.

Dr. Martin’s address is: School of Finance & Applied Statistics, Australian National University, Canberra, ACT 0200, AUSTRALIA. Internet address: Michael.Martin@anu.edu.au

†This work was carried out with financial support from the ANU Faculty of Economics and Commerce Summer Research Grant Scheme. We also wish to thank an anonymous referee for several insightful comments that improved our paper, and Dr. Steven Stern for his assistance with revising our document into LATEX.
We focus in particular on the use of graphics as a communications tool as they are an efficient and potentially highly effective means of conveying information. We also describe several common errors in graphic construction—and how to correct them—using examples from the business world.

Key words and phrases: Communications skills, errors in graphic construction, presentation skills, statistical graphics, technical writing

1 Introduction

Strong technical skills are a hallmark of the actuarial profession. But at least as important as these skills is the ability of the actuary to communicate information accurately, unambiguously, and effectively. The type of information that actuaries routinely communicate is complex. Yet it must usually be made available to people from a wide variety of often non-technical backgrounds. As a result, there is an enormous burden on the actuarial profession to be able to convey information at an appropriate technical level while maintaining enough detail to satisfy professional actuarial standards. This need is more critical now than ever as the actuarial profession assumes more prominent roles in management where the effective flow of information between organizational tiers can mean the difference between success and failure.

The ability to communicate technical information well is a learned rather than an inborn skill. Peculiarly, most current approaches to actuarial education do not specifically try to teach good communications skills, relying instead on students' abilities to pick up the skills as they need them, usually in the post-study workplace environment. This baptism of fire unfortunately can result in the skills being acquired at considerable cost to employers. They cannot necessarily rely on students who arrive straight from their actuarial studies to be able to communicate as effectively as they would like.

At the root of the problem is the fact that traditional actuarial education is focused largely on the development of excellent technical skills—that is, at getting the calculations right—without too much regard to how the results are presented. To a certain extent this approach is reasonable, in that many traditional classroom/study exercises are canned questions, presented without sufficient background or context to make the presentation of the answer a critical concern. In setting exercises, instructors rarely have the luxury of providing strong motivation or background details for the calculations. Hence, students can
struggle to grasp the importance (both practical and theoretical) of the results. Examinations are conducted under strict time limits, so getting details down on the page emerges victorious over learning the ability to explain ideas to others. Unfortunately, any separation of actuarial calculations from their context can be counterproductive for actuarial students.

Finally, and most importantly, one of the critical aspects of technical communication is the ability to judge the level of the intended audience. At the university level, students write solutions for professors and lecturers, so the amount of technical detail presented in answers is appropriately high. This situation is likely to be very different, however, from that encountered by a professional posed with the problem of preparing a report for a client.

In this paper we attempt to increase awareness among actuaries of the importance of communication skills. In so doing, we develop a set of principles that can be used to promote these skills among actuarial students. Learning good technical communication skills is not as simple as picking up a writing manual from the reference section of the book store. While the skills that make for good technical writing overlap with good general writing skills, they need to be developed with special care. Moreover, technical communication can take multiple forms, including written, oral, graphic construction, and presentation skills. Graphic construction skills are particularly useful, as graphics are an extremely efficient way to present, summarize, and describe large sets of numbers. Fortunately, the same set of key principles governs all these disparate forms of communication.

Three major elements distinguish technical communications by actuaries from more general forms of communications:

- The need to describe complicated mathematical ideas and financial concepts;
- The heavy reliance on graphical tools to convey quantitative information; and
- The need to report on complex numerical data analyses in order to describe stochastic (random) behavior.

Each of these elements places a different burden on the communicator, and each can be a significant barrier for non-technical audiences. So learning principles that specifically target these three areas is essential for actuarial students. Along with these particular skills, actuaries also must develop good general communications skills.
The position of actuaries at the intersection of so many disciplines, including banking, finance, economics, and statistics, places a considerable impost on their communications skills. They must speak the languages of multiple disciplines. As both information-consumers and information-providers, actuaries need to be flexible communicators, able to interpret and analyze intricate and complicated numerical information, yet also able to communicate the results of their labors in a way that is both comprehensive and that their clients can understand. The actuary is in this respect an information intermediary, a vital link between raw, unprocessed data and effective decision-making in the presence of risk.

2 Broad Principles for Writing, Talking, Drawing, Presenting

Several key principles apply to all forms of technical communication. The principles are intrinsically linked. Briefly, they are:

- Know your audience. Communications must be framed with a specific audience in mind. The likely ability and background of the audience is an important factor in deciding the level and type of detail communicated. Speaking to a conference of qualified actuaries about a stochastic model fit for calculating insurance premiums is a very different task from that of justifying use of the same calculation to a group of shareholders. The chosen content and style of communication should reflect that difference. The group of actuaries may be interested in learning about the assumptions underlying the model, as well as in obtaining some detail about the process of fitting and assessing the model against historical data. The language chosen to describe these features could be appropriately technical. The shareholder group, on the other hand, might want an overview of how the new technique may affect customer premiums—the language chosen in this case reflects the broader, non-technical nature of the audience. The cost of misjudging the audience is high: the group of actuaries presented with a broad overview might react with distrust ("What are they hiding? Is this even correct?") or even boredom ("Surely there is more to this than what we're being told! Yawn!"); the shareholders confronted with the more technical discussion may react with confusion ("Huh? Why are they telling us this?") or even anger ("What a waste of time!").
• Content is supreme. The most polished presentation, whether it be written, oral, or graphical, cannot make up for lack of substantive content. Florid prose in a report, ornate decorations in a graphic, or gaudy colors on a Microsoft PowerPoint® slide may initially attract an observer's eye to what you have to say, but if there is no substance to it, they will just as readily look away. The old adage that one should open one's mouth only when one has something worthwhile to say is an apt lesson in effective communication. This principle holds for all forms of communication. What constitutes "substantive content" will vary from one audience to the next, so understanding the intended audience is an important part of organizing information for presentation.

• Context is vital. It is critical that information is presented in an appropriate context so that it may be interpreted properly and unambiguously. This goal can be difficult to achieve. A number of factors affect the appropriate context, in particular the intended audience and the nature of the content. Every element of technical communication needs to be carefully considered as to how it might be integrated into an effective, efficient presentation. Technical or mathematical arguments, graphics, and data analyses should be used to complement the main message to be presented—not to overpower it, nor as a substitute for it. Complex technical arguments should be motivated carefully in the context of the information to be presented and should be explained intuitively rather than formally unless the specific circumstance (e.g., a meeting of technical professionals) demands a more formal presentation.

• Language should be simple, concrete, and familiar. The information you present needs to be understood easily, and the best way to ensure such an outcome is to use direct, precise language. Here, the word "language" is intended in a broad sense to include text, pictures, speech, and even gestures. Information should be presented specifically for your target audience, using words and expressions familiar to them. For writing and speech, jargon, acronyms, and obscure technical references should be avoided. Sentences should be short without being terse or choppy. Simplicity is also an important quality for graphics, but it is often an elusive goal as simplicity of construction and simplicity of interpretation can be conflicting aims. Nevertheless, graphics should be constructed to be as simple as possible, avoiding redundant or obstructive graphical elements. Oral presentations accompanied by
Microsoft PowerPoint® slides should integrate text and graphical elements in a clean, seamless way, avoiding garish color schemes, flashy transitions, and distracting background patterns. Lack of content cannot be adequately disguised by these devices, and substantive content is diminished by their use. Simplicity is again an important but elusive quality, as oversimplifying a presentation by reducing it to a set of bullet points can lead to a stilted, formulaic perception.

- Integration is important. Most technical information is multifaceted, so its communication should be diverse. Because different modes of communication are effective for different members of an audience, incorporating information as text, graphics, and numbers is not redundant. The use of data should be embraced rather than avoided, as numbers lie at the heart of the content of much technical information. Creative graphical displays of numerical information are extraordinarily valuable. The primacy of text in technical communications is more a reflection of the history of human communications than an inherent strength of text as a means to communicate information. One of the advantages of modern computer systems is their ability to more thoroughly integrate each of the elements of text, images, and numbers into a single document. Of course, integration needs to be implemented in a way that is both creative and stylish so that the outcome is a cogent, aesthetic whole rather than a piecemeal mess.

- Dealing with complex information. Effective technical communication recognizes the complexity of information and responds to that complexity in a creative manner. Size and dimension are two elements of data complexity that pose immediate problems for analysts and communicators. Nowadays, enormous, complex data sets are common (e.g., minute-by-minute stock prices on a portfolio of stocks), so a key task of technical communicators is that of describing complex data patterns as simply as possible. This idea is akin to data compression, whereby gross features of large data sets can be summarized by using relatively few measures; for instance, representing complex returns series in terms of mean and standard deviation measurements. Sound statistical practice is essential if this data compression is to be a successful strategy in understanding large data sets. In such cases, it is inevitable that some information will be lost in the description. Good statistics is about discovering what is important (signal) at the expense of what is not (noise). Graphics are a particularly efficient means of
representing large amounts of complex information. Technical communicators therefore need to develop excellent graphic construction skills.

Almost all interesting information involves relationships between many variables or factors, so techniques need to be developed that allow such high-dimensional information to be displayed on low-resolution, low-dimensional display surfaces. Computer screens, though seen as modern advances on paper, allow for much lower-resolution images than are possible on paper. As a result, the question of how we reduce the dimension of our data so we can see what is going on is complicated by the low resolution of the device we use to look at the data. Reducing multivariate data to one or two dimensions involves inherent loss of information. So we need to understand the extent and consequences of this information loss.

Exploring high-dimensional data for relationships between variables also raises complex issues such as cause-and-effect. Graphics should help us to assess whether variables are causally related. Unfortunately, this can be a tricky question, which relies not only on good logic, statistics, and experimental design (where that's possible), but also on the luck of asking the right question in the first place.

Comparison is a vital tool in understanding data and communicating what you see. Good communication invites the question "compared to what?" in response to the size and nature of revealed data structures. Such comparisons promote logical thinking about the nature of relationships within data and assist us in deciding what features of a data set are important. Effective communication is interactive in the sense that it engages the audience to understand and actively participate in the thought processes underlying what is being presented. Their response can feed back into the analysis to allow an even better understanding of the data. Good communication elicits the right questions from the audience. Good graphic construction facilitates useful comparisons through careful thinking about the locations of graphical elements within a single graph or on a single page and through the use of techniques such as small multiples (many graphics located close to one another so that they may all be compared in a single sweep of the eye).

The information actuaries must deal with can be incredibly complex. Effective communicators do not seek to deny the complexity
of the information they describe. Rather, they try to exploit what simpler structures explain the mechanisms underlying the complex data structures.

- Good communication is hard work! Even the most brilliant technical work can be spoiled by sloppy presentation. Poor grammar, crowded or careless graphics, poor organization of presentation layout, and even poor choice of presentation font, are all inimical to good communication. To communicate effectively, you must win your audience's attention—and fight to keep it. If your report is riddled with typographic errors, readers may assume that your carelessness extends beyond your typing. If your graphics are misleading, your audience may distrust other things you say. If your Microsoft PowerPoint® presentation is full of flashy transitions of bullet points, your audience may overlook your substantial content and simply enjoy the sideshow. The solution is simple: practice, practice, practice. Read your own writing and critique it. Think about the presentations or reports that have engaged you, and remember what about them made you pay attention. Then employ these strategies in your own communications.

All the modes of communication described—writing, speech, graphics, and presentations—benefit from the application of these broad principles, but the principles apply in different ways to each of them. In what follows, we explore the various modes of communication and offer particular advice on how to convey ideas effectively.

3 Writing Technical Documents

Beginning, middle, end: the importance of structure: Since mankind first developed language, storytellers have held a revered place in society. Ancient storytellers such as Aesop remain famous today. One of the central tenets of good storytelling is the idea that a good story must have a well-defined beginning, middle, and end. So it is with good technical writing. Without such basic structure, even brilliantly conceived ideas cannot be conveyed effectively. In story-telling, the beginning is used to set the scene, giving readers the chance to understand the basic setting of the story; the important elements of the story unfold in the middle, hopefully engaging the reader's full attention; the ending is the climax of the tale, tying together loose storylines and presenting the moral of the tale. Each of these components is fundamental to the story as
a whole: a tale that begins in the middle is confusing; one without
an ending is unsatisfying and frustrating; without a middle, there
is no story!

**Humble beginnings—“what's my motivation?”**: The storyteller's craft
remains useful for technical writers today. The beginning is par-
ticularly important for technical writers; in some cases, it is all that
will be read! The introduction to a technical work should motivate
the work carefully, summarize the main results, and highlight the
important conclusions. In documents typically produced by actu-
aries, the executive summary plays this role. It is also useful for
the introduction to signpost briefly what is in the remainder of
the document—some readers may be interested only in a partic-
ular section and may ignore the document if they cannot find it
easily.

One of the worst sins of technical writing is that of failing to moti-
vate the work adequately. Writers mistakenly believe that they are
wasting the reader's time by giving background to their work. Not
at all! By making the purpose of the work apparent immediately,
authors make it more likely that readers will read past the first
few paragraphs. Authors who fail to adequately motivate their
work, fail to engage reader interest and may even prevent some
readers from understanding their work.

How should technical work be motivated? The answer lies in the
key principles described earlier. First, the intended audience for
the work needs to be considered. Ask yourself what the audience
can be expected to know about the topic. Anticipate the question
"Why are you telling us this?" from the audience. Clearly state at
the start what the main issue is, why it is important, and what your
solution is—outline the content. For some of the audience, this
may be all they want; for others, it will allow them to frame your
work in a context that is familiar and important to them. State up
front what solutions you are offering. Also try to state what the
work does not do.

**The power of summary—drawing conclusions**: The introduction and
the conclusion are equally important. The introduction foreshad-
ows the importance and relevance of the results, and the con-
clusion must summarize the findings of the work, both technical
and practical, and, if possible, make a recommendation based on
those findings. The conclusion must be both concise and precise.
It must also be written in an authoritative style and should be self-contained.

**Why is technical writing different?:** Technical writing, whether it be in business or science or engineering, differs from general prose in three major ways: the use of mathematical or other technical detail; potentially heavy reliance on graphical tools; and the important role numerical work plays in the underlying content. Each element poses a different burden on the technical communicator, but the first element represents the most formidable challenge to the effective communication of ideas. Mathematics and its associated disciplines are akin to foreign languages to many people. To them, comprehending a technical document full of mathematical ideas is difficult. Yet, much of what actuaries do could not exist without mathematics and statistics, and the nature of actuarial business requires that actuaries must frequently communicate with professionals less technically-inclined than themselves. For example, actuaries designing new life insurance products must coordinate their activities with legal professionals who establish the contracts under which the products will operate. How can such disparate groups communicate effectively? The answer lies in the principle that information must be presented in an appropriate context so that the audience can tackle the information in a language and manner that is reasonable given their backgrounds.

So, how are technical or mathematical arguments to be presented? The key principles of audience identification and appropriate context guide us. For example, an actuarial consulting report should contain the minimum amount of mathematical detail necessary for addressing the key consulting questions. Numerical advice should be included as necessary. The client does not wish to (and may not be able to) read detailed mathematical arguments; nor do they need to see every interim calculation. They do wish to know, however, that the recommendations made by the consulting actuary are based on appropriate assumptions and correct, logical thinking. Answers to questions should be framed in the same language as the original question is framed, and abstractions should be avoided. A common problem experienced by novice technical writers is that they forget that the excitement or enjoyment they derived from developing the intricate mathematical arguments that supported their work is rarely shared by the readers of the report. If mathematical arguments disrupt the main message of the paper, then they do not belong there.
Many actuarial reports are intended for both actuaries and non-actuaries. One way in which writers can make technical arguments available is by way of a technical appendix. This idea is well-known in academia. Many academic journals discourage authors from including too much mathematical detail within a paper itself, but offer authors the opportunity to include technical proofs in an appendix. This approach keeps authors focused on the main issues in the paper, but also allows them to express mathematical ideas freely. We must recognize the huge difference between a published academic paper and a consulting report or a technical report for a client. There remains room within certain types of reports for what is essentially tangential material. How else can interested actuaries recreate what the author has done, either for their own interest or to learn the technique and apply it to their own work? In other types of reports—for instance, client-only documents—there is often no room for such luxuries. Sometimes an unusual amount of detail is necessary, either because the problem demands a mathematical solution and nothing else will do, or because the solution is particularly unusual or novel and becomes, of itself, the subject of interest. Again, context drives the types of decisions to be made here.

When mathematical or technical ideas need to be presented in detail, their presentation makes an enormous difference to the way in which the material is perceived. Most importantly, readers are more able to comprehend mathematical or technical arguments if they are able to grasp the main ideas intuitively. Necessary mathematical complexities should be prefaced by remarks that attempt to explain the goals to be achieved and the means of achieving them in an intuitive way. For example, if a new pricing methodology is best explained by a mathematical statement, then a report introducing that idea should motivate the need for the new technique. It should explain how the development avoids problems with the existing practice. Alternatively, the writer should explain how the new technical developments facilitate a solution where none was previously possible. Such an approach transforms a continuous stream of mathematics into a more fleshed-out argument based on logical principles that most people can understand even if they fail to grasp the mathematical detail itself.

There is a broad literature dedicated to mathematical writing. Its focus is more appropriate to mathematicians or engineers than to actuaries. If your requirements include writing that is heavy
in mathematical notation, however, it is worth reading the style
guides produced for the Mathematical Association of America by
Gillman (1987) and Knuth (1989) and those produced for the Amer­
ican Mathematical Society by Krantz (1997) and Steenrod et a1.,
(1981). Other notable works in this area include the books by Al­
ley (1986), Barras (1978), and Higham (1993) and the article by
Ehrenberg (1982). In addition, there are countless corporate style
guides and manuals for writers in business and finance, though
these tend to focus on more routine business correspondence and
reporting than actuaries usually have to produce.

Writing is not the same as problem solving: One of the difficulties fac­
ing novice technical writers is that they seldom have any writing
training or experience. Most students write solutions rather than
reports, and the main drive is to write enough detail in obtaining
the correct answer so that a high grade will be awarded. Such a
strategy can be disastrous when the same person must write a con­
sultancy report because the parameters governing good solution­
writing are very different from those governing good technical
writing. No longer is it necessary to recreate the sequence of steps
that lead to the answer—the reader probably does not care about
those details. Rather, the reader is more likely to care about the
interpretation and consequences of the answer.

A useful way of discovering whether your own technical writing
style is cumbersome is to read it aloud, as if you were trying to
verbally explain it to someone. If the writer finds that they have
to stop repeatedly and explain one thing or another, then they
have not written the document well. Technical writing places the
writer as a filter between raw inputs (usually extremely technical)
and comprehensible outputs.

Heuristics: One resort to which technical writers can turn when the
detail underlying their work is too technical is to use heuristic ar­

Heuristics: One resort to which technical writers can turn when the
detail underlying their work is too technical is to use heuristic ar­
proof? Nevertheless, heuristic arguments are usually seductive to
readers. As long as the writer makes clear that the arguments are
informal educated guesses, most readers will appreciate them for
what they offer. Writers need to be careful, however, not to put
forward heuristic arguments as if they were formally correct, as
this will only annoy or confuse the reader who is able to follow
the arguments closely.

A similar comment applies to the use of analogies. Most analogies
are imperfect. The writer must be prepared to acknowledge where
the analogy fails. Otherwise, readers may unwittingly extend it to
an area where it does not apply, thereby drawing incorrect con­
clusions.

The role of jargon: Like many technical fields, actuarial science and its
associated disciplines of finance and statistics are awash with jarg­
on. Accrual rates, annuities due, net present values, discounted
cash flows, preserved benefits, life tables, mortality, exposure,
loss adjustment expenses, loss reserving, rating factors, reinsur­
ance, written premiums versus earned premiums, ...: the list goes
on and on. What is worse, each of the main actuarial practice
areas (such as life insurance, general insurance, pensions, or su­
perannuation) has its own jargon! Fortunately, actuarial training
is broad, and qualified actuaries must be familiar with all of the
major fields. Nevertheless, given the increasing complexity of the
profession, an actuary in one practice area may find it difficult to
converse with an actuary in another. For clients and other busi­
ness professionals, the situation can be even more frustrating.
Further (admittedly anecdotal) evidence that jargon is now a se­
rious impediment to sound actuarial practice is that a cursory
search of actuarial consultancy web pages reveals that many con­
sultancies now explicitly promise explanations “in simple English,
free from actuarial jargon” as part of their terms of business.
Yet, it is easy to see how beginning actuaries fall into the trap of
routinely using jargon. Almost all actuarial education ingrains the
jargon indelibly into the lexicon of the training actuary. A report
on examination performance for examinations of the Institute of
Actuaries in 1997 reads as follows¹ (our emphasis indicated in
italics):

Candidates were asked to prepare a letter explaining dif­
f erent interest rates quoted for the same loan. This re­

¹<http://www.actuaries.org.uk/pastpapers/1997apr/q2a97rep.pdf>
port summarizes the main points which the examiners were looking for in the solution. Scripts were expected to read like letters to friends, without jargon and with any technical terms clearly explained. Many candidates did not achieve this.

It is easy to spot jargon in a technical report you have written: spell-check it in your word processor (without using any custom dictionaries you have created). Assuming your usual spelling skills are fairly good, most of the words highlighted as incorrectly spelled fall under the rubric of jargon. The best advice is to avoid jargon wherever possible. Of course, the difficulty is that most jargon is usually shorthand that connotes some complicated idea.

The general principle of “know your audience” is critical to deciding whether jargon is appropriate. If the paper or report is to be read by a group of the author’s peers, then it is usually acceptable to use relevant jargon. On the other hand, if the report is to be read by a client, jargon should be kept to a minimum. Writers who follow this advice will often have to create lengthier documents. The increase in length, however, is worth the effort if it means that their reports can be read by the intended audience.

A writer should particularly avoid creating jargon. Unless an idea is novel, the temptation to coin a new phrase to describe it should be avoided. Most ideas, even very good ones, do not warrant the introduction of a new word into the already-crowded actuarial lexicon.

**TMA (Too Many Acronyms):** Along with the excessive use of jargon, perhaps the most annoying tendency of technical writers is that of creating acronyms. Some use of acronyms is acceptable, provided it is absolutely clear to the writer that all members of the audience are familiar with them. For example, in a report to a government minister, it is acceptable to refer to government departments using well-known acronyms such as IRS or FBI. Other cases are less clear. While it may be perfectly clear to the writer that MoS stands for “margin on services”, it may be far from clear to a non-actuarial audience. Finally, only a frustrated mathematics graduate would end a technical argument with QED, an acronym so old that the language from which it was drawn is now dead!
4 The Role of Graphics—Modern Cave Painting?

Graphics are a powerful way to communicate technical information. They can summarize and describe vast amounts of information in a compact, efficient, and eye-catching way. Well-constructed graphics can transcend the barriers of language and numeracy because they rely on the almost automatic response of the human brain in interpreting shapes and patterns. Visual information is processed in a different part of the brain than language or numerical information—in much the same way as a modern computer hands off complex video or audio processing to dedicated hardware away from the main processor. Even people without specialized training in pattern recognition or statistics are able to interpret graphs reasonably well. Unfortunately, the reliance of graphics on human visual perception also leads to their greatest weaknesses—the human eye is easily tricked. Thus graphics must be constructed with care lest they lead to misinterpretations and confusion.

Graphics are, by their nature, demonstrative, and the purpose for which they are constructed needs to be clear and unambiguous. Like effective writing techniques, effective graphic construction is a skill that needs to be learned. Howard Wainer, who has published several highly-readable papers on statistical graphics, says that "like motor car driving and making love, drawing graphs is an activity that most statisticians feel they can do well without instruction. The results, of course, are usually disastrous." While humorous, this sentiment is, regrettably, all too true. Moreover, with the ready availability of graphics capability within computer packages such as Microsoft Excel®, truly abominable yet visually attractive graphics are at the fingertips of anyone who can switch on a computer.

Edward Tufte's beautifully-crafted and insightful books on visual displays of information (1983, 1990, 1997) belong in the library of anyone working with data. Tufte's works examine the rationale behind graphics as a tool for communicating information, and they set out definitively and elegantly the key principles of information design and display. Other notable work includes the books by William Cleveland of Bell Laboratories (1985, 1993), who has been at the forefront of developments in statistical graphics, the book and articles by Howard Wainer from the Educational Testing Service (1997, 1984, 1990) whose discussions of the good and bad of statistical graphics are both insightful and entertaining, and articles by Anscombe (1973) and Tukey (1990).

The role of graphics in business and financial communications has attracted some attention in the actuarial, accounting, and finance lit-

The first, and most important, rule of graphic construction is to identify the likely audience for the graphic. It is no coincidence that this is the same “golden rule” as for writing and presenting! Tailoring information for the specific audience is critical for all forms of technical communication. For example, while a survival curve is immediately meaningful to a life actuary, it is unlikely to be an effective graphical tool for a more general, non-statistical audience. Similarly, standard statistical tools such as quantile-quantile plots, while they are models of graphic construction and invaluable statistical analysis tools, are of limited to no use in presentation to general audiences.

4.1 Two Types of Graphics

We will distinguish here two types of graphics: presentation graphics, which are explicitly designed for use in a published report for the consumption of others, and analysis graphics, which are routinely produced as part of a larger analysis and would generally not be part of the ultimate report. Examples of presentation graphics include bar charts, histograms, time series plots, and pie charts. Graphics such as survival curves, quantile-quantile plots, residual plots, and so on are more often classed as analysis graphics. In our present context, most of our comments are applied to presentation graphics, but many also hold true for analysis graphics.

Bar charts and time series plots can be very useful graphical tools, but careful attention should be paid to the principles of good graphic construction even when creating such simple graphics. Pie charts should probably be avoided altogether, as they suffer from several deficiencies that limit their effectiveness. They rely on a reader being able to spot slight differences between areas of sectors of a circle, a feat many people find difficult and unnatural. Moreover, pie charts usually encode only a handful of numbers, and a table is usually a much more efficient way to present such information. Edward Tufte opines that “given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used” (1983). While we agree with Tufte’s sentiment, we concede one point in favor of pie charts: they are very familiar to audiences in business and finance, and this familiarity can make them easier to interpret. Nevertheless, variants of pie charts
such as three-dimensional or exploded pie charts and the aptly-named doughnut chart, are anathema to effective communication.

4.2 Rules for Effective Graphic Construction

4.2.1 Substantive Content Should Drive the Need for Graphics

A graphic should represent a significant piece of information. In simple terms, graphics are designed to attract readers' attention, but they must also be meaningful. It makes no sense to encode only a few numbers into an overblown graphic—in these cases, a small table makes more sense, and gives readers direct access to the numbers involved—see Figure 1. Tufte states that "visually attractive graphics also gather their power from content and interpretations beyond the immediate display of some numbers. The best graphics are about the useful and important, about life and death, about the universe. Beautiful graphics do not traffic with the trivial" (1983). Of course, decisions as to what is important are highly subjective!

Lack of purpose in graphic construction is betrayed by several telltale signs. The first is low data density, a measure of how much data is represented in the space allotted to the graphic. We have seen numerous annual reports that force the reader through a forest of bar charts or pie charts, each of which represent only a handful of numbers. A better option is to use a single, moderate-sized table.

Graphics adorned with excessive decoration also can conceal a lack of content. Figure 2 depicts such a case, where "chartjunk" (i.e., extraneous decoration with no informational content) dominates the graphic. Good graphics answer the question "Why?" as well as the questions "What?" and "How?" about a set of data. Where possible, they should reveal cause-and-effect relationships. Two excellent examples of how graphics might achieve this goal are given in Chapter 2 of Tufte's book Visual Explanations. He shows how graphics were instrumental in the discovery of the means of cholera transmission and how graphics, had they been more thoughtfully constructed, may have prevented the launch of the ill-fated 1986 space shuttle Challenger.
This graphic shows how a whole page of a report can be taken up describing just two numbers! This bar chart encodes only two numbers—about 112,000,000 for 1998 and about 110,000,000 for 1999. The bevelling and greyscale gradient on the bars reduces their perceived height, while the small amount of data encoded makes it a shame to waste an entire page in a report. A small, two-number table, or even a short sentence, would suffice. The report from which it was drawn contained eight similar graphics, each describing just two numbers. (Source: ©Australian Venture Capital Association Limited—Year 2000 Yearbook.)

How well a graphic achieves its purpose can be difficult to judge, as graphics can show both what is present in the data and what is not. As a result, graphics can surprise and delight us by making apparent features of the data that were not originally anticipated. Detailed data analyses involving graphs are best thought of as iterative processes. Preliminary, mainly graphical exploration of the data is followed by a deeper investigation based on model formulation, fitting, and assessment. Graphics are an integral part of each stage of the analysis. Individual graphics also benefit from an iterative approach to their design, whereby each element is carefully considered in the context of its interaction with other graphical elements.
This graphic encodes just four numbers: 5553 in 1961, 6645 in 1965, 6166 in 1971, and 5808 in 1985. The decoration dominates the graphic to such an extent that it misrepresents the data hideously. Note that the horizontal distance that represents the 14 years between 1971 and 1985 is shorter than the preceding intervals of four years, presumably so the "mouth"—a decoration—remains in proportion with the face. Also, the final amount (5808) appears smaller (positioned lower than) than the initial amount (5553). Curiously, the authors have applied some good statistical practices—the amounts reported are medians rather than means, and the currency is adjusted to 1972 levels. (Source: ©MBC (Makati Business Club) Economic Papers, September 1988.)

4.2.2 Good Graphics Promote Comparisons

Good graphics must be based on sound logical principles and good statistical practice. Graphics must not lie! Almost all interesting and important arguments involving numbers are relative—how big is one number compared with another number, and what does the difference in their size mean in the context of the problem at hand? Difference and change are the drivers of almost all decision-making. Comparison is the
most important tool of scientific inquiry that we have. Good graphics reflect the same principles as sound, logical reasoning, invoking wise comparisons in a way that is both natural and aesthetic. There are several ways in which graphics can be constructed to facilitate meaningful comparisons:

**Figure 3**
The Dominant Face Reworked

A re-working of Figure 2 shows a shape that is not mouth-like at all! In fact, it appears as if the rate of decline is slowing. Unfortunately, the long time period between measurements makes it difficult to sustain this argument.

- If two curves are to be compared, consider plotting their difference or their ratio rather than simply putting both curves onto a single axis. This technique forces comparison along a horizontal baseline and takes advantage of the fact that humans can perceive even slight deviations from straight lines, especially horizontal and vertical ones;

- Plots should be augmented by the addition of visual elements such as fitted lines wherever possible so that patterns in the data are easily recognized;

- Graphical elements that are close to one another are more easily compared than those that are far apart. As a result, placing multiple lines on the same set of axes or multiple graphs on a single
sheet of paper is an effective way to promote comparison. The latter idea, referred to as the *use of small multiples*, is a particularly effective way to describe large amounts of multivariate data efficiently. The concept behind small multiples is that a large number of similar graphics can be explored within a single eye span, so even small differences become readily apparent. The worst case is where several graphics to be compared are spread over several pages.

Mere proximity of graphical elements does not guarantee wise comparisons, as the optical illusion in Figure 4 shows. Figure 5 shows an example of the use of small multiples to compare the marketing budgets of several different kinds of firms.

An excellent graphic design which uses small multiples is the scatterplot matrix, which depicts all two-dimensional relationships among pairs of variables in a multivariate data set. This graphic manages to render high-dimensional information into two-dimensions and does so in a way that allows the reader to quickly explore each panel for evidence of correlation—see Figure 6 for an example.

**Figure 4**  
An Optical Illusion

Which of the two "middle circles" is larger? Most people answer that the one on the right is larger, when in fact they are the same size. Size is judged in relation to the outer array of circles in each case. The left middle circle is small compared to the circles surrounding it, while the right middle circle is larger than its surrounding circles. The result is an incorrect perception in comparing the two middle circles.
Here, 16 small graphical elements are placed close to one another to facilitate comparison between them. Unfortunately, the nested cylinders' size bears scant relation to the numbers they represent. Overall, a good idea (small multiples) is ruined by poor choice of symbols (nested cylinders). Note also the heavy use of jargon ("midicorps", "microcorps"). A caption describes in three lines what the graphic was unable to impart. *(Source: © Business Today magazine, Feb 22—Mar 6, 1998, page 67.)*

The graphic design of Figure 6 is good because it allows the viewer to examine many two-dimensional slices of the high-dimensional data space quickly. Care must be taken, though, in interpreting the graph, as interesting directions in the data may not include those involving only two variables at a time. Also, while the graphic is capable of showing association between variables, it cannot address the question of whether such relationships are causal. The establishment of causality must be more than a visual process—it also requires careful logic and, typically, good experimental design.
A scatterplot matrix exploring the relationships between various economic indicators and a disability index (number of disability insurance claims scaled by a measure of exposure). Covariates included were a measure of consumer confidence, employment participation rate, long-term unemployment, real GDP per capita, and a total bankruptcy rate. The data were reported by Service and Ferris (2001). There appears to be an association between the disability index and each of the other variables—see the top row of the array. There also appears to be several relationships among the covariates (e.g., Bankruptcy and GDP, Participation Rate and GDP, Participation Rate and Bankruptcy), which makes separating the individual effects of each variable difficult.

4.2.3 Graphics Should Be Designed to Be Aesthetically Pleasing

Proportion, perspective, and scale are important elements of graphic construction. Graphics need to be eye-catching without being garish. Every design element of a graphic should be considered in terms of how it may affect the viewers’ ability to perceive the content of the graphic. The aspect ratio of a plot, the ratio of the height of a graphic to its width, can affect how the content of a plot is perceived—Figure 7 shows such a case. Also, careful attention needs to be paid to the layout of graphics...
within a page. For example, two histograms sharing the same set of horizontal axes and bin-widths should be aligned vertically rather than side-by-side to facilitate easier comparison of their shapes.

**Figure 7**
The Importance of Aspect Ratios

![Can you see the pattern?](image1)

![What about now?](image2)

Each of these two graphics plot the same data, yet the pattern (a simple sine wave) only emerges clearly in the lower graphic. What has changed? Aspect ratio and choice of vertical axis.

Seemingly benign design aspects, such as choice of axes, are critical to drawing graphs that are easy to interpret. For example, the practice of including the zero point on all axes reflects a poor design choice, as zero may be nowhere near the bulk of the data. Unfortunately, such choices are often not left to users, as popular computer packages such as Microsoft Excel® offer the feature of axes including zero as a default for some choices of line graph (and, strangely, not for others). While the default can, of course, be changed, many users will never exercise this choice. Figure 8 shows two versions of a graphic depicting household income data from Figure 2. Which is the more truthful?
The forced inclusion of 0 on the vertical axis of the left plot de-emphasizes the extent of the change in income over time.

Other design factors that affect graphical perception include choice of plotting symbol and the use of colors or shadings. A good general principle is that when a graphical element is used to encode numbers, that element and the information it encodes should be of the same physical dimension. For example, bar charts violate this principle because they encode single numbers as two-dimensional objects (bars), rather than as one-dimensional objects (lines). Three-dimensional bar charts are even worse, as they encode a single number using a threedimensional object. The introduction of redundant dimensions promotes ambiguity in how one interprets the graphic (does the depth of the 3-D bar have any meaning?), and ambiguity is the enemy of effective graphic construction. Three-dimensional elements also risk introducing unusual perspective effects into graphics, the overall impact of which can be unexpected—see, for example, Figure 9.

Color is a potentially effective tool in graphic construction, though its use has been historically low. Color needs to be used carefully, as they are not strongly visually ordered, whereas grey scales are. As a result, grey scales are preferable in many instances. Moreover, up to 5% of the male population suffers some form of color-blindness, so designs should not rely exclusively on color. Shading patterns such as cross-hatching can also lead to unusual and distracting optical effects such as moiré vibration.
Unfortunately, the introduction of a spurious third dimension into the plot causes bars that have negative borrowing components to appear as if they are in front of the other bars—they have leapt into the third dimension! (Source: © Report on Public Sector Borrowing, Australian Public Service, 1994.)

Consistent choices improve the impact and comprehension of graphical forms. We have seen annual reports in which several flavors of bar charts (stacked, three-dimensional, bevelled) have appeared on consecutive pages of the report. This practice causes readers to constantly switch frames of reference and makes comparison across graphics difficult.

4.2.4 Graphics Should Be Simple In Interpretation and Perception

Of the four principles discussed, this one is the most elusive. It is not always possible to attain graphics that are both simple to visually perceive and simple to interpret. A case in point is the use of transformations—data are often transformed so that when they are summarized by a graphic, the main features are readily apparent. Yet, when a viewer comes to interpret the graphic, they must do so remembering that a back-transformation is necessary before any conclusions.
can be drawn. To attain perceptual simplicity, one must keep in mind the way human visual perception works. For example, people can see very easily when a pattern of data points deviates from a straight line, but may be unable to perceive similar scale deviations from a curved line. Equally, it is perceptually easier to observe deviations from horizontal or vertical lines than it is from lines at arbitrary angles—this is the principle that makes residual plots such an effective tool for assessing the quality of a statistical model fit.

Quantile-quantile plots, designed to assist in detecting when data do not plausibly arise from a bell-shaped distribution, are an example of excellent graphical construction. They achieve simplicity in perception, but they are not simple to interpret without training. The graphical premise underlying quantile-quantile plots is that data are transformed onto a particular scale where departures from normality are associated with non-linear patterns in the plot. Discovering such patterns is much easier, perceptually, than the process of deciding whether a histogram of the data looks bell-shaped. Most viewers cannot adequately envisage what “bell-shaped” means, whereas deciding whether a pattern is linear or not is easy. The difficulty arises once the visual pattern has to be interpreted in the original context. Inexperienced viewers make the mistake of interpreting the pattern as meaning that the original data have a linear relationship with some other variable—that is, they attempt a literal interpretation of the shape of the plot. Only when the link between distribution shape and the associated Q-Q plot is made do viewers realize the correct interpretation—see, for example, Figure 10.

Other design elements also impact the simplicity of graphics. Impediments to simplicity include the abundant use of abbreviations on a plot, overuse of legends and different line types (e.g., dotted, dashed, dot-dash lines), and excessive decoration.
Baxter, Coutts, and Ross (1980) report data on total cost of claims for 128 combinations of claimant age, vehicle age, and vehicle type categories. Histograms of total claims for the 128 categories and the log of total claims are shown at the right, with associated normal Q-Q plots on the left. The distribution of total claims is highly skewed to the right (indicated in the Q-Q plot by a non-linear, concave-up curve), while the distribution of log total claims is closer to bell-shaped, but slightly skewed to the left (notice the slight concave-down curvature in the Q-Q plot).

4.3 Major Errors in Graphic Construction

4.3.1 Misrepresentation of Data

The most common error in graphic construction is the use of graphical elements that either deliberately or accidentally fail to accurately represent the data they encode. The simple paradigm to which all graphics should adhere is that graphical elements that represent numbers should be drawn in proportion with those numbers. This straight-
forward rule is breached surprisingly often. Two simple cases are where bars are not started at zero but at some other, arbitrary value, or where long bars are broken; see Figure 11 for two examples. In each instance, the relative heights of the bars are not in the correct proportion—the relevant visual metaphor is broken.

Figure 11
Broken Bars

Vodafone has fewer call drop-outs.

It's confirmed - Vodafone has fewer dropouts than Telstra or Optus in NSW, Qld, Vic, SA and the ACT.

So you can enjoy clearer calls and less noise against throughout Australia.

Bar lengths should be proportional to the numbers they represent! In the left graph, the largest bar should be only 1.005 times as large as the smallest bar, but visually the ratio of their sizes is about 5. In the right graph, the larger bar should be about 1.6 times the length of the smaller bar, but visually the ratio of their sizes is about 6. In each case, the error favors the company producing the graphic. Amusingly, the fine print on the left graphic admits that the graphic is not drawn to scale—why bother to print it then? (Left: © Vodafone, Source: Australian Communications Authority, March 2000. Right: Wesfarmers Retail Pty Ltd share offer for Howard Smith, June 2001.)

This case can be contrasted with that discussed in Figure 8 for line plots, where the most relevant visual element was the slope of the line rather than its height above the baseline. Hence, the lack of a zero baseline is not as critical a problem for line plots as it is for bar charts, which use relative heights of the bars to visually encode the numerical information to be transmitted.
Figure 12 shows another obvious misrepresentation where the time scale is seriously distorted. Ironically, the headline for this graph, when translated, reads "A picture is worth a thousand words"—unfortunately, almost all of the words we can use to describe this graphic are critical.

A key error common with financial data is the failure to adjust monetary amounts for factors such as inflation. Invariably, if inflation is not accounted for, strong positive trends in variables such as spending are generally overstated.

Another, more subtle form of misrepresentation is data aggregation prior to graphing the data. Aggregation is a form of data smoothing that allows for long-range trends to be observed in volatile data. If the data are aggregated too coarsely, important short-run information can be lost. For example, reducing quarterly or monthly data to annual data by aggregating quarters/months into years can cause significant seasonal variations to be obscured. In extreme cases, this approach can lose the most important or interesting information.

As a simple illustration, in some classes of general insurance, claims are likely to rise in certain seasons (e.g., storm and fire insurance claims will tend to rise in summer and decline in winter), and these critical trends will be missed if data on such claims are annualized. Of course, the amount of aggregation appropriate for a particular set of data depends on the question being asked. In the preceding example, annualized data would be appropriate if the goal of the graphic were to display the overall growth in claims over the last ten years. If, on the other hand, finer detail were required, the amount of aggregation would need to be reduced.

Smoothing and aggregation inherently involve loss of information. The key to a satisfactory graphical outcome is to identify what extent of information loss can be tolerated for the question at hand. A straightforward way to avoid misrepresentation is to experiment with differing amounts of smoothing before deciding which graphic gives the most useful and truthful account of the data. Remember that good graphic construction is a process of iterative refinement—the search for truth in graphics is neither short nor easy.

Other subtle misrepresentations in bar charts include failure to begin bars on a common baseline (making it harder to judge their relative sizes) and varying the width of bars (the perceived size of a bar is related both to its height and its width, so equal width bars should always be used)—see Figure 13.
In this graphic, the time scale is so distorted (the two years from 1985 to 1987 are represented on the time scale using the same distance as for the seven years from 1978 to 1985), and the viewing angle and perspective so distracting that the graphic is almost useless as a visual tool for understanding the U.S. dollar/Swiss Franc exchange rate. It is extremely difficult to judge the extent to which the Swiss Franc had recovered its value in 1985 after the initial drop in the late seventies. Also, the width of the exchange rate curve increases as the curve moves down the page. The strong visual impression is that the relative size of the dollar to the franc is growing over time (see the increasing width of the curve, and the decoration of growing dollars rolling off the end of the curve). Of course, during the period under study, the currencies generally moved in the opposite direction. A simple time chart would communicate the correct information much more efficiently and unambiguously—see the plot on the right, which clearly shows the 1985 recovery of the franc to over half its 1970 value. (Source: Computerworld Schweiz, 1989, © Cash magazine.)
Perhaps the most common form of data misrepresentation occurs when bar charts are constructed using decorative elements other than fixed-width bars to represent data. One only has to pick up a copy of USA Today or browse their website\(^2\) to find a vast array of exotic shapes (e.g., hot dogs, bears, arms, legs, hats) presented as bars in a bar chart. The problem with these decorative elements posing as bars is that in order to make them look real, their widths must remain in proportion to their height, so that as their heights grow so do their widths. As a result, although the relative heights of these bars are correct, their relative perceived sizes—usually their areas—are not. This design variation distorts viewers’ perception of the data, and hence misrepresents the true situation.

\(^2\) See, for example, <http://www.usatoday.com/snapshot/news/snapindex.htm>
4.3.2 Redundant Dimension

Graphical elements should have the same dimension as the information they encode. So, a single number $A$ is better represented by a line segment with length proportional to $A$ than by a square whose side-length is proportional to $A$. This preference is based on what we know about human visual perception—when people are presented with a two-dimensional figure such as a square, they usually perceive its size as its area ($A^2$) rather than its side-length or diagonal length.

The introduction of spurious dimensions into a graphic also causes ambiguity. Some viewers will interpret characteristics in that extra dimension as carrying meaningful information, while others will not. The use of three-dimensional bars in bar charts also can create unusual depth and perspective effects. If the useful information in a bar chart is only represented by the height of the bars, then the bars should be rendered as lines, not bars, and certainly not as three-dimensional blocks, or, worse, cylinders or cones. In the case of three-dimensional bars, the perceived size of objects is their volume (proportional to $A^3$), rather than their heights ($A$). The use of fixed-width, two-dimensional bars is acceptable only because their areas are in the same proportion as their heights and because such bars are aesthetically nicer than simple lines. Nevertheless, varying the widths of two-dimensional bars introduces spurious information into the redundant second dimension and hence the perceptual properties of the graphical element. Some examples of problems arising from redundant dimensions are shown in Figures 14 and 15.

Not only does the problem of redundant dimension create distortion of information for viewers of a graphic, but it also slows their comprehension of the information in the graphic. An interesting article by Fischer (2000) explores the issue of whether redundant dimension in bar charts materially affects the speed of comprehension among viewers. He finds that there is, indeed, a significant slowing of cognition for graphs containing such irrelevant depth cues. The last chapter of Cleveland (1985) describes a number of other visual perception experiments with analogous results.
Here bars are presented as pieces of a cake (a three-dimensional object). To maintain the proportions of a piece of cake, bars are of varying width. Other notable errors in this graphic include a non-zero baseline for bars and a distorted time scale at the left of the graph. As a result of these errors, the ratio of perceived size for the largest to the smallest piece of cake is about 50 (using volumes) and about 15-20 (using areas), while the actual ratio of their sizes should be $3.5/2.75 = 1.27$. (Source: Australian Education Union)

4.3.3 Excessive Decoration

Graphics should be eye-catching, but not to the extent that the real information is drowned out by the extraneous decoration. Tufte refers to such elements as chart junk. Decorations cannot rescue a graphic based on low or no substantive content. When decorations dominate a graphic, the graphic becomes itself a decoration, and it ceases to be a useful tool for communicating information. Worse still is when substantive content is hidden or distorted by decoration, because viewers may misinterpret or even distrust the information they receive.
Some fundamental problems arising from redundant dimensions include hidden bars and oblique baselines. Unfortunately, many of the measurements for life insurers cannot be recovered at all from this graphic, as they are completely obscured. The third dimension on this graph could be collapsed so that grouped side-by-side bars for each entity could be presented on a two-dimensional bar chart with investments on the horizontal axis and percentage on the vertical axis. (Source: ©Australian Taxation Office (1999), Tax Reform: not a new tax, a new tax system.)

Put simply, if the information is important, you do not need to highlight it with ornate decorations—the substantive content you provide will hold the viewers' interest. Examples of excessive decoration are given in Figures 16 and 17.
The graph encodes four numbers in almost the worst way possible. The decoration is the graphic! The information to be conveyed is that the average bull market lasts about four years and has a real return of about 100%, while the average bear market lasts about a year and has a real return of -25%. It is not clear how this comparison is served by depicting a bull that is, perceptually, about 10-15 times the size of a bear. This two-dimensional rendering of the data is largely meaningless. (Source: © Professional Investor magazine, October 1997.)

4.3.4 Multiple Vertical Axes

Authors commonly construct graphics in which several data series are plotted on a single plot with vertical axes on each side of the plot corresponding to the different series. While this device saves some space, it almost always introduces visual effects that encourage inappropriate comparisons between the two series. If two series are thought to be related, scatterplots are a far better tool for assessing any relationship. Intersections between lines on a plot with multiple vertical axes are particularly easy to misinterpret. Visually, intersections between the series suggest a sudden change in the ordering of the two series—one suddenly appears larger than the other. Of course, the effect is usually
spurious, as the series are on entirely different scales, and the intersection is an artifact.

Similarly, varying slopes on a plot with multiple vertical axes lead to a misinterpretation as to the relative rate at which the two series are changing. The rate of change of each series is completely dependent on its vertical scale, so relative rates of change in a plot with two vertical scales are meaningless—by changing the scale on one of the axes, one can change the viewers' perception of the plot completely.

Figure 17
A Pie Spiral of Pension Fund Capital

Each step of the spiral represents the amount of pension fund capital at five year intervals. The amount of capital is encoded as the heights of the sections, which appear to be in roughly the right proportions. Yet, the angle subtended by each section also systematically grows as the eye moves up the spiral, so the perceived size of segments—measured as volumes—grow much faster than they should. The amounts given are cumulative, though this is noted nowhere on the graphic. Conventionally, time is depicted as increasing from left to right. Here time grows in a spiral, purely as a decorative effect. The graphic encodes only six numbers. (Source: © Computer Graphix AG, Info, January 1990.)
The upper graphic attempts to show the relationship between height and relative mortality risk by plotting two series, height and mortality, on the same graph. Although the heights are presented in ascending order on the plot, they are shown as equidistant from one another when, in fact, they are not. Graphics must respect the fact that numbers have not only order but also magnitude. Apart from the initial error of plotting multiple, different-scaled series on the same graphic, the graphic suffers a number of other weaknesses. These include the lack of explicit vertical axes, choice of stylized human figures as bars (redundant dimension), and ambiguity about where the shadow bars begin (do they begin at the feet of the human bars, or at the line separating red background from blue background?). The heights in the data are almost equidistant from one another, so the shape of the mortality curve depicted in the original graph is almost, but not quite, right. The lower graphic is a much better graphic for examining the relationship between height and mortality. It is a simple scatterplot relating the two. A horizontal reference line was added to the new plot at height 1 to reflect a baseline mortality risk. ( compra/shape © TIMS magazine, November 11, 1996.)
The graphic on the left shows the yen declining in value at about the same rate as the Australian dollar between January 1997 and June 1998. The differing scales for the two exchange rates mean that the actual rates of change were different. Note the right vertical axis is in reverse numerical order. If, as the title of the graphic on the left suggests, the goal is to show that Australian and Japanese currencies were moving together, a simpler and more direct method would be to plot the Yen/SA exchange rate against time as in the right graphic. (Source: (Left) © Business Review Weekly, June 1998.)

Figure 20 shows an example of where two series plotted on the same graphic interact particularly poorly, even though the series are just two ways of considering the same information.

4.3.5 Breaking with Established Conventions

We all view graphics through a set of inherent filters that allow us to perceive information quickly and easily. Some of these rules are obvious, such as lines going upward on a page representing increase while downward sloping lines connote decrease; words should read left to right; when comparing graphical elements, larger objects represent larger numbers than smaller objects; and so on. Other rules are less obvious: time on a plot evolves from left to right on a horizontal axis and from bottom to top on a vertical axis; white represents absence while black represents presence; an object in the background of another, same-size object will look smaller. When these conventions
are broken, we become confused as our fundamental assumptions are challenged. Comprehension is also slowed as visual information usually processed automatically must be analyzed anew. Conventions exert enormous impact on what we can understand easily; we should not break them frivolously or carelessly. When you construct a graphic, think about what you see and relate it to what you mean others to understand from the graphic. Linking the visual to the cognitive is an essential part of graphic construction. Figures 21 and 22 show how breaking conventions can radically alter—or even reverse—viewers' perceptions of a graphic.

**Figure 20**

**Avoid Mixing Raw Amounts and Percentages**

The graphic on the left depicts a series of raw amounts (bars) and a series of percentage growth rates from previous years (line). Although both series are presented on the same plot, explicit labeled vertical axes are not provided (numbers are instead presented on the graphic itself). The plotted line is like a second derivative of the original series, a quantity that is not easily visualized. The two plotted series are visually at odds—although the series of raw amounts always grows, the superimposed line plot suggests regular declines. In fact, the declines depicted are in the relative rate of growth. The left graphic also suffers other construction errors. The mixing of graphical elements (bars and lines) is visually jarring. The lack of explicit vertical axes is also a problem as it forces the data to be presented directly on the graphic (which begs the question of why a graphic is needed at all). A more effective, and simpler, way to view the data depicted would be to present a simple line chart of the original annual valuations—see the graphic on the right. Variations in growth rate are easily seen in such a graph as the line either moves up or down from its previous angle. *(Source: Carrett and Stitt (2001), *Australian Actuarial Journal*.)
A quick look at this graphic suggests that sales and profits are falling for the Freshwater Fish Marketing Corporation. Time to get out of business? Hardly! A closer examination of the time axis at the bottom of the plot reveals that time is plotted in reverse from 1989 to 1982 - in fact, profits and sales have risen since 1982. (Source: © Winnipeg Free Press, obtained from <http://www.stat.sfu.ca/~cschwarz/Stat-301/Handouts/Descriptive/BadGraphs/seafood.gif>)
The speed of microchips has increased exponentially since 1977. Yet the curve traced by the spheres in this graphic is turning in the opposite direction to that expected of an exponential increase. The fact of an increase is apparent, but the nature of the increase is not. Also, instead of increasing in equal-sized steps from left to right, the time axis in this graphic is traced by a set of concentric curves. The spheres have volumes that are not in the same proportions as the numbers they represent. The graphic is more decorative than informative. (Source: © Scientific European, October 1990.)

4.3.6 A Modern Problem: Too Much Power, Too Much Choice

Ten years ago, the task of producing graphics was the domain of the graphic artist. Unfortunately, graphics artists were more often trained in art not in statistics. As a result, many information graphics were beautiful to look at, but did not convey information accurately. Today, almost anyone with a personal computer can produce professional-looking displays. Popular spreadsheet programs can produce a dizzying array of graphics. For example, Microsoft Excel can produce 14 standard types of graphics—column, bar, line, pie, scatter, area, doughnut, radar, surface, bubble, stock, cylinder, cone and pyramid—each with multiple variants for a total of 73 basic designs plus numerous custom charts. Yet only three of them—the basic bar/column chart, line chart, and scatterplot—are worthy of common use.

Bar and column charts differ only in whether the bars run horizontally or vertically. We prefer vertical bars as up is a more natural direction to connote increase or aggregation than left-to-right. Three-
Microsoft Excel’s Chart Wizard offers an astonishingly long list of graphical possibilities. Pictured are the six variants of pie charts: basic, 3-D, basic with sub-pie, exploded, 3-D exploded, and basic with sub-bar.

dimensional variants of these charts always introduce redundant dimensions and should be avoided. Stacking bars makes comparisons of their components difficult, as each component has a different baseline as we move from bar to bar. Variants that replace fixed-width bars by other geometric figures such as cones or pyramids suffer redundant dimensions, but also fail because the shapes chosen are narrower at the top than the bottom, so the bar’s height is de-emphasized.

Line charts are particularly useful for representing data developing through time. Our preference is for line plots rather than bar charts for time series data, as the joining of adjacent points in a line plot emphasizes the movement of the series through time. Three-dimensional variants of line charts are particularly hard to interpret. Area charts, created by filling the area under line charts, are generally ineffective. It is usually the height of the line above a baseline, not the area under the line, that encodes the appropriate information. Ambiguity about which of these features is the relevant graphical element in a line chart creates different perceptions for different viewers.

Stock charts are a variant of bar or line charts that track high, low, and closing stock prices for a particular stock over a number of days. These charts share the properties of bar or line charts, with the further advantage that professionals in finance are familiar with their interpre-
tation. Nevertheless, there is little to distinguish them from simple bar charts or line charts. We do, however, recommend line charts be used for stock prices rather than bar charts to emphasize the flow of stock prices over time.

Pie charts fail largely because although humans perceive straight lines effectively, our ability to perceive subtle differences between sectors of a circle is unreliable and variable. Three-dimensional, exploded, and doughnut varieties of pie charts only complicate our perceptual difficulties. Tables of numbers prove more effective.

Scatterplots are useful for exploring two-dimensional relationships, and higher-dimensional information can be encoded through the appropriate choice of plotting symbols. For instance, a relationship in four dimensions can be represented in two dimensions by plotting the first two variables as coordinates of a scatterplot, but instead of points, plotting rectangles whose height and width are proportional to the third and fourth variables, respectively. In a similar vein, Microsoft Excel*'s bubble charts use circles to encode a third variable. However, bubble charts are not effective because the values of the third variable are encoded as the diameters of the circles, whereas the perceived size (area) of a circle is proportional to the square of its diameter.

Radar charts (also known as star charts) can be useful for visualizing continuous multivariate data, but the resultant shapes need to be interpreted carefully. They are best used in small multiples, where a large number of radar charts can be compared quickly to assess variability in multivariate data. Surface charts, including perspective plots, contour plots, and wireframe variants, may be useful for three-dimensional data, but have limited use in other contexts.

It is important for creators of graphics not to confuse artistic sophistication with graphic sophistication. In graphic construction, simpler is better. Decorative or realistic effects such as three-dimensional bars or complex shadings can dramatically affect viewers' perceptions. Remember, your content will drive interest in your graphic, and lack of content cannot be concealed by flashy visual effects. Modern software gives us unprecedented choice when constructing graphics—exercise that choice wisely, always keeping in mind the needs of readers.
5 Presenting Numbers in Technical Communications

Actuaries convey complex numerical information to other actuaries, professionals from other disciplines, and clients. From simple reporting of numbers to complex sensitivity analyses and financial simulations, the types and range of numerical information to be communicated are broad. These requirements impose a burden on actuaries because many less-technical audiences regard numerical analyses and simulations as a "black box" from which the desired outcome emerges. Actuarial techniques are iceberg-like in that only a small fraction of the work performed is displayed in detail. How the results of many hours of complex calculations can be pruned into a manageable, easily explained form is a major problem faced by the profession. Actuaries cannot assume that policy- and decision-makers are familiar with common actuarial terms. All communications must be cast in as simple and familiar a language as possible. Wise use of graphics is part of the answer to this problem, but non-graphical techniques are also important.

The most important rule for communicating numerical work is that the author must present the results in a way that addresses the original question or issue in the same language as that in which the question was raised. The divide between formal, numerical answers and plain-language answers must always be crossed by the communicator, and the audience must never be forced to take this responsibility. Plain-language answers will be appreciated by the audience no matter what their technical level. Even for those at the highest technical levels, communications that summarize numerical results in a direct, easy to understand way will be regarded as insightful and useful.

The second rule for communicating numerical work is that, although a plain-language approach should be used in summarizing the work, it is nonetheless important to recognize the power of numerical arguments and, therefore, not to avoid using numbers. Numbers carry meaning beyond what can be transmitted using only words and images, and their inclusion remains an important part of technical communications.

The key is to strike a balance between detail and summary. Identifying the needs of the audience is critical to knowing what balance will work well, but some broad advice can assist communicators to find the right mix.
• Simulations and sensitivity analyses need to be broad enough to draw proper conclusions. Numerical results also need to be supported by appropriate rigor. Analyses that do not consider enough cases or which are based on inappropriate assumptions usually lead to incorrect conclusions and warped logic. If a new methodology is to be recommended for use with real data, it should be tested on real data. We have seen new techniques only tested on data drawn from three theoretical distributions that foundered when they were applied to real data. If real data cannot be obtained, then simulation studies must be broadened to reflect real-world experience. Similarly, sensitivity analyses must reflect the types of departures from set assumptions that are both credible and probable in the real world. Whenever new methodology is to be presented, the onus always rests with the presenter to demonstrate the merits of the new technique.

• Actuary, heal thyself! Actuaries are well-trained in statistics, but all too often we see in papers and reports large tables of numbers presented, undigested, accompanied only by a dismissive statement that the author’s conclusion is “clear from the information presented in the table”. No matter how careful or correct the calculations underlying the creation of such tables of numbers, the content of the tables must be analyzed with care, using proper statistical techniques. All conclusions should be delineated and specified. Nothing should be declared as simply “clear from looking at the table” as it puts the burden of analysis on the reader. Analyses of results often show the stated conclusions to be far from “clear given information in the table”!

• Actuarial modeling is an iterative process that generally cannot be described in a technical report or paper. More likely, the report will contain the final results of a modeling procedure, with little or no discussion of what other models were considered, what diagnostics were performed through the analysis, and so on. It may be acceptable to detail the model-fitting process, but in that case it is critical to also mention what assumptions underlie the analysis and to defend their tenability. Diagnostic procedures should only rarely be reported in the final document; authors should resist the temptation to report their results as a step-by-step description of the analysis. Any data uncovered as unusual should be noted, especially if they have been removed in the course of the analysis.
Consider a general insurance setting in which the actuary is required to implement and describe a stochastic model for claims experience based on a number of rating factors. First, the actuary needs to consider carefully what assumptions will underlie the analysis. These assumptions need to be stated in describing the model, along with a reasoned discussion of choice of rating factors, availability and source of data, and so on. In this case, a generalized linear model for claims experience might be appropriate, and a formal statistical analysis of building a model would proceed. The actuary might present the results of the model fitting, perhaps in a table, together with $p$-values for the various rating factors. While the formal modeling process is complete, the actuary's task in communicating the information has just begun. Readers cannot be required to draw their own conclusions from a set of $p$-values alone—many non-technical readers will be unable to make such judgments unassisted. The actuary must describe the final model in plain language, note what rating factors were found to be important in the analysis, and note which factors were not. Any anomalies that arise from the modeling process, such as factors that were considered a priori important but which were not included in the final model, need to be explained, in plain language. Finally, the actuary must draw a proper conclusion, which may include a formal recommendation that the new model be considered for evaluating future claims experience.

6 General Issues in Effective Writing

As technical writing is a subset of writing in general, the principles that govern good writing undoubtedly apply to technical writing. Many of these principles are covered in general style guides, such as the University of Chicago's *The Chicago Manual of Style* (1993) and Strunk and White's *The Elements of Style* (1979).

**Grammar is important:** Nothing annoys readers more than reports that are poorly written from a grammatical perspective. Writers should ask a colleague to read their work. Writers should use the automatic spell-checking and grammar-checking facilities built into modern word processors. All reports should be proof-read carefully, as spell-checkers will not find all mistakes.

We are reluctant in this forum to enter the ever-raging debate on the use of “I/We” in technical papers, and we feel that this is more a matter of style than of correctness. Masculine pronouns, and the ubiquitous “he/she”, should be avoided in favor of plural pro-
nouns (they, their) to minimize the risk of alienating a large proportion of the audience. Other more formally grammatical issues, such as active versus passive voice and issues such as maintaining the appropriate tense within a paragraph are also important, but are more than adequately covered in popular style guides.

**Precise, concise, wise—keep it simple:** Writers should adopt a precise, concise style. Strategies that promote such a style are those that avoid complicated grammatical structures. Writers should avoid tangential comments or asides, particularly when the writing style required is formal. Writers need to think critically about what they are saying, even down to the level of small phrases.

Each writer has their own personal style, and it is futile to try to produce automatons who each write in a uniform, simple manner. It is possible to adopt a straightforward writing style without sacrificing your individuality. An easy step is to re-read each document you write and proof it, searching not only for grammatical and technical errors, but also for stylistic gaffes. These can be just as damaging to your report's ultimate fate as even the most grievous technical error.

**Words and phrases to avoid:** Certain style guides essentially prohibit the use of long or difficult words. We argue that these are matters more of style than substance. Simple, short sentences that use primarily simple, straightforward prose are unlikely to be ambiguous. Unfortunately, they also tend to be fairly dry and uninteresting to read. When writing, one should always have access to a dictionary and a thesaurus. (We simply point our web browser to <http://www.dictionary.com>.) Examples of words that can and should be avoided (with alternatives shown in parentheses) include: utilize (use); facilitate (help); endeavor (try); terminate (stop); transmit (send); demonstrate (show); initiate (begin or start); necessitate (need); elucidate (explain); and so on.

**Don't be a draft dodger!** These days, composing documents at the keyboard is widespread. Of course, such electronic innovations have done wonders for productivity, subjectively measured in terms of pages of output. But they have also been partly responsible for a sharp decline in the ability of people to carefully craft their documents.

The benefits of on-line composition are obvious: changes to your document can be made easily; sections can be added or deleted
at a whim; grammar and spell-checkers can eradicate typographical and other errors; and "intelligent agents" built into modern software can automatically structure documents into a variety of familiar and impressive formats. It is easier than ever for the written word to look professional. The same cannot always be said, however, for the quality of the content! No software yet available can alter the fact that writing is a craft that benefits from reflective thought and practice. No matter how skilled the writer, the first draft of a piece of technical work is rarely ever acceptable as a finished work. Indeed, the second, third and fourth drafts often need polishing. Good writing requires not only skill, but also patience. A writer must be prepared to read and re-read a document several times and make changes as appropriate before the document can be considered for submission.

**Titles, abstracts, and references**: Many reports are judged solely on their title, summary, and reference list. For many written works, these are the only parts of the document that are widely read.

- **The Title**: Titles should be brief and descriptive. Obviously, these two goals are at odds. The temptation to incorporate all the ideas from the report into the title should be avoided. The main question authors should ask themselves before selecting a title is "what would make me want to read this paper".

- **The (Executive) Summary**: Material for the summary needs to be chosen even more carefully than the title. It needs to cover the main ideas from the report without overwhelming the reader with unnecessary detail. A good basic structure is to have a separate short sentence describing each of the main ideas in the report. The summary generally should not exceed a single page in length. Mathematical or technical symbols are usually not useful in the summary.

- **References**: Where a technical document is meant for wide distribution, the reference list is an important part of the work. Almost all technical work is derivative in some sense, and it is critical that the relevant research of others be cited fairly and appropriately. Only directly relevant items should be cited, unless the paper is a review article. Writers should be careful not to unduly reference their own previous works.
7 Effective Presentation Skills

Just as critical as the ability to write well is the ability to present technical information to a live audience. Although the principles of effective communication also apply to live presentations, a number of factors distinguish this form of communication. Good skills in this area are increasingly important as the actuarial profession evolves and actuaries assume more prominent management roles within companies.

Timing is a dominant issue. Speakers should always adhere to time limits. Arrive early for your talk as it allows you to check that any equipment works, allows you to relax a little, and often gives you the opportunity to make yourself familiar with some of the audience.

The live component of your presentation is ephemeral—unlike writing, there is usually no chance to edit mistakes or to recast part of the presentation. On the plus side, there is a spontaneity associated with speaking to a live audience that cannot be captured in writing or other forms of communication.

Presentations afford the communicator a unique opportunity to directly interact with their audience, either explicitly through a genuine dialogue or implicitly by adjusting the presentation on the fly in response to audience reaction. A good communicator can sense the mood of the audience and respond by varying the tempo of the talk—by focusing on particular issues that seem to pique the audience's interest or de-emphasizing topics of less importance. The ability to skip ahead in a presentation or to slow down gives the communicator an enhanced opportunity to engage the audience's attention.

The forum of a live presentation often allows the speaker to use language less formal than in a written report. While speakers still need to structure what they say according to good grammatical rules, natural speech is considerably more free-form than writing. Other forms of non-verbal communication such as eye-contact, gestures, and facial expressions evincing emotions (e.g., smiling!) are possible with live presentations. If they are used wisely they can make the presentation come alive more than any written report can.

Audiences for live presentations are precious. If you lose an audience during a talk, you may never get them back. Live audiences are notoriously variable, with factors such as time of day having a significant impact on their behavior—a successful pre-lunch presentation could be a post-lunch bomb! Finding the right level for a presentation is a critical but tricky proposition. Never assume an audience knows nothing and speak down to them—this will alienate a proportion of the listeners. Use plain language, and avoid jargon and colloquial expres-
visions. Regard the audience as intelligent, but potentially uninformed, and fashion your presentation accordingly. Look at your audience's faces and eyes, and judge from their reactions whether you are pitching your material at the right level—and adjust your presentation if necessary.

Try to learn beforehand the culture of the group to whom you will be presenting. Will the audience interrupt to ask a question or will they wait until the end of the presentation? Will there be any particular people to watch—e.g., a senior manager who always asks a question out of left field—and think beforehand how you will handle that situation should it arise. Unlike writing, live communications can be fluid, and you will be judged on both your presentation and on how you react to the moment. The best approach is to be enthusiastic and confident. If there is a lectern available, do not use it! Move freely and engage the audience whenever possible. Speak at a comfortable volume, modulating your voice as you would when talking to a friend. Above all, be natural and as relaxed as possible.

7.1 Structure Your Presentation

Just because a presentation is live does not mean that it should not be carefully scripted. Like all forms of communication, it should be structured in a way that is logical and clear. Begin by stating the overall goal of your presentation. Make it clear why what you are discussing is important, to whom it is important, and what your solution is. Audiences conditioned to sound-bites need to know the main message of your talk up front.

Whatever your preferred mode of presentation—overhead, Microsoft PowerPoint® slides, physical charts, or just plain speech—always prepare a handout for the audience. A handout is a tangible reminder of your presentation, it gives it a life beyond the hour in which you speak, and it makes it possible for people who cannot attend the presentation to receive your message. A handout also signifies that you stand behind what you say—you are willing to commit it to paper and, hence, to close scrutiny. Because the audience will be taking the handout with them, it needs to be prepared carefully.

First, the handout must contain your name and contact details—if a question occurs to a person the day after your talk, they will want to contact you. It also must contain the date of the presentation to place what you say in some historical context. The handout should be a document prepared specifically for that purpose, not just a copy of your slides or a copy of the full report. The handout you prepare
should summarize your talk, cover each point you raise in plain lan-
guage, and include any key graphics or tables on which you want the
audience to focus. The handout is also a safety valve in case you forget
to mention an important point. It is a special-purpose document, not
an afterthought to the presentation. It will represent you far longer
than the presentation will.

Live presentations are inevitably less-structured than written doc­
ments. This aspect of presentations is a double-edged sword. Some
presenters fail to recognize that some structure is critical, and their
presentations typically meander and never make any real impact. Good
presenters impose structure, but can take advantage of the freedom of­
fered by the live environment to adjust their presentation in a variety of
ways that promote audience engagement. For instance, judicious repe­
tition of key ideas can be extremely effective in delivering the required
message.

7.2 The Mechanics of Presenting

Presentation has an element of presence and a physicality not found
in written communications. It usually relies heavily on technology for
delivery, and logistical preparation is an important component of giving
an effective presentation. How material will be shown (overheads or
Microsoft PowerPoint®), how your output should look, what the room
is like (lighting, layout), how large the audience is likely to be, even the
time and day of the presentation, are all critical questions in planning
your talk.

Most business presentations are made using Microsoft PowerPoint®,
although some holdouts still use overhead projectors. If you are using
Microsoft PowerPoint®, it is best to bring your own laptop computer
and your own data projector. This practice avoids difficulties related to
operating system differences (Windows/Macintosh/Linux), logins (you
may not have an account on the system at the delivery site), and versions
of available software. You should bring two copies of your presentation
on separate disks (or on CD) as well as a copy appropriate for use on
an overhead projector. Be prepared for the worst!

• Overhead slides:

Overhead slides should be typed rather than handwritten and
should not be too crowded as people at the back of the room need
to be able to read them. Dark ink should be used to promote vis­
ibility. Avoid at all costs what Tufte refers to as the "trapezoidal
strip tease," the practice of concealing the overhead and revealing
the contents one line at a time. This technique discourages the audience from engaging as it forces them to follow the presentation at an artificially imposed pace and it encourages linear thinking.

• **Microsoft PowerPoint** slides:

Presentations delivered in Microsoft PowerPoint suffer most of the same problems as presentations delivered on overheads, plus some new problems related to the features of the software. Microsoft PowerPoint gives users an incredible number of choices for the display of information, but our recommendation is to use many of the features conservatively. The interface for your presentation needs to be selected carefully, paying particular attention to the following issues:

- **Colors:** Color schemes need to be chosen thoughtfully to present the appropriate image. Dark writing on light backgrounds is recommended, as it provides the best readability at a distance. Light writing on dark backgrounds is also a reasonable, high-contrast choice, but is less legible at a distance. Other color choices are usually disastrous, particularly red writing on blue backgrounds which results in uncomfortable vibration effects. Also, combinations of red and green cause particular difficulties for members of the audience who are red-green colorblind.

- **Transitions, Advances, and Fades:** Microsoft PowerPoint provides multiple sophisticated visual and sound effects that can be used in transition between slides or even from one line within a slide to the next. These effects are flashy, distracting nuisances. The audience is not there to see a movie, complete with special effects. Transitions used to advance from one line to the next within a slide are the Microsoft PowerPoint equivalent of the trapezoidal strip tease, and they should be avoided.

- **Fonts:** As far as possible, use standard, sans serif fonts for your presentation. These fonts have maximum readability at a distance and are guaranteed to be available on any standard computer on which you can run your presentation. Odd or decorative fonts should be avoided at all costs—they play the same role in presentations as chartjunk plays in graphic construction—that is, they de-emphasize your content.

- **Backgrounds:** Also avoid the use of distracting logos or backgrounds to your slides. Company logos should be discreet
and tastefully placed. People only need to know where you are from once, so large distracting reminders on every slide are overkill. Similarly, Microsoft PowerPoint's default collection of clip-art is familiar to most people.

- Layout: Slide layout is important, and Microsoft PowerPoint's default layouts are reasonable, though they favor the use of dot points more than we recommend. Dot points promote simplistic, linear thinking. We suggest a more flexible strategy that uses ideas such as hyper-linking creatively to allow the presentation to respond to audience reactions. An approach that invites audiences to ask questions and make comparisons assists in turning the presentation from a monologue to a dialogue. Slides that mix text, graphics, and numbers tend to be more interesting and promote such dialogues. Care must be taken not to clutter the slides too much.

7.3 Speaking Strategies

Speaking before a live audience can be a traumatic experience. Good speakers use this nervousness to their advantage by channeling the resultant energy into an enthusiastic delivery style. Effective communication is fostered if the audience feels comfortable with the speaker. Good eye contact, natural gestures, and a relaxed attitude will help create such an atmosphere. If possible, have the lights on in the room when you are delivering your presentation. This choice will make it easier to establish and sustain eye contact with your audience.

The best presentations are those delivered in a steady voice using a natural tone, much as if you were involved in a conversation with another person. Being natural will help you avoid nervous habits such as uttering "umm's and aah's" or fidgeting. The best way to avoid these nervous habits is to realize the power of silence in a presentation. Pauses between sentences or ideas can be extremely useful, as they give the audience time to absorb what has just been said. Pauses also can be used deliberately to give more effect to the preceding statement. Silence can be golden. Fidgeting can be controlled by holding a laser pointer or a pen, but such props should be used sparingly.

Finally, always be ready for questions from the audience. Never react with surprise, dismay, or disdain. Your overall performance may be judged by how you handle direct interactions with the audience. Many questions are directed at drawing attention to the questioner, not embarrassing the speaker. Treat all questions with respect. Be prepared to take time answering them and to admit you do not know the answer if
necessary. Questions can enliven your presentation. Encouraging them sends the signal that you are confident and competent. A useful strategy is to plant a colleague in the audience who will ask a pre-arranged question. This approach can induce others from the audience.

8 Conclusion

While the content of most actuarial communication is technical, the craft of communicating such information effectively is as much an art as it is a science. Just because the information is structured and detailed does not mean that you cannot exercise creativity and style. Nevertheless, technical communications must conform to certain guidelines to be effective. Understanding your audience's abilities and needs is critical to the successful communication of your work. Good technical communications result from meaningful content described in a straightforward way. Further, the best technical communications recognize the power of combining text, images and numbers in a compelling presentation.

Finally, spend some time in the shoes of your audience. Learn from your own experience in listening to and reading the communications of others. Remember what attracted you to presentations you enjoyed and what repelled you from those you disliked. Attempt to emulate the techniques used in the good presentations. Try to be as objective as possible in assessing how effective your communication style is. If you cannot be objective, ask a colleague to assist by critiquing your style.

Learning effective technical communication skills is a difficult and frustrating task, but the rewards of possessing such skills are well worth the price of obtaining them.

References


Unearned Premiums and Deferred Policy Acquisition Expenses in Automobile Extended Warranty Insurance

Joseph Cheng*

Abstract

A prorata formula is commonly used to calculate unearned premium reserves in property-casualty insurance. I believe, however, that an exposure-adjusted formula is more appropriate in automobile extended warranties. This paper describes the exposure-adjusted approach to calculate the unearned premium reserves of an automobile extended warranty insurance program, to test the adequacy of the calculated reserves, and to determine the allowable deferred policy acquisition expenses from an insurance company’s perspective.

Key words and phrases: exposure, unearned premium reserves

1 Introduction

In the 1970s auto manufacturers introduced a one-year/12,000 mile bumper-to-bumper warranty on new vehicles as a response to consumer demand for quality. Reception from consumers was so good that third party companies (i.e., companies that do not produce automobiles) began to market an extension to the manufacturer’s warranty; hence the

*Joseph Cheng, F.C.A.S., F.C.I.A., M.A.A.A., is the President of J.S. Cheng & Partners Inc., a Canadian actuarial consulting firm that specializes in all areas of property-casualty actuarial services. He obtained his bachelor’s degree from California State University at Chico in 1972 and his master’s degree from University of Illinois at Urbana in 1974. Prior to founding J.S. Cheng & Partners Inc. in 1991, Mr. Cheng was a partner with Eckler Partners Inc. He has 26 years of P&C industry experience including careers in insurance companies, rating bureaus, and consulting firms. Currently, he is the appointed/valuation actuary to more than 25 Canadian insurance/reinsurance companies.

Mr. Cheng’s address is: J.S. Cheng & Partners Inc, 1500 Don Mills Road, Suite 706, Toronto, ON M3B 3K4, CANADA. Internet address: jscp@jscp.com
name extended warranty.\textsuperscript{1} Third party automobile extended warranties are now subject to insurance regulation because many third party uninsured extended warranty programs became insolvent in the 1980s and numerous consumers were left uncompensated. As the market for automobile extended warranties expands, the manufacturers also market their own brand of extended warranties. Unlike third party companies, each manufacturer provides extended warranties only for its own vehicles. Most, but not all, jurisdictions exempt the manufacturer's brand of extended warranty from insurance regulation.

A new automobile extended warranty (hereinafter called an extended warranty) usually is defined by two limits: time and mileage. An extended warranty expires when either of the two limits is reached. For example, a five-year/60,000 mile extended warranty means the warranty will expire either in five years or when the odometer reading reaches 60,000 miles, whichever comes first. The extended warranty for new vehicles usually does not come into effect until coverage under the manufacturer warranty has expired. Most manufacturers now offer a three-year/36,000 mile full (bumper-to-bumper) coverage; a few offer even longer ones. As a result, only 10 percent to 30 percent of new vehicle owners have purchased extended warranties.

As the exposure of an extended warranty is measured from the initial registration date of the new vehicle, the age of any extended warranty is the time elapsed between the initial registration date and the valuation date. In this paper, an extended warranty is assumed to be effective on the first day of the effective month.

Most extended warranty programs give the original owner up to 36 months to purchase an extended warranty as long as the three-year/36,000 mile portion of the manufacturer's warranty has not expired. A delayed purchase does not increase the exposure of the insurer at the time of purchase because both the expiration date and mileage limit remain the same as if the extended warranty were bought on the registration date of the vehicle. Also, a change in vehicle ownership does not automatically eliminate the exposure of the extended warranty insurer. In almost all extended warranty programs, the coverage can be transferred to the next vehicle owner by paying a modest transfer fee. In some cases, the transfer fee is waived.

\textsuperscript{1}The manufacturer's warranty that comes with a new automobile is not considered insurance in North America.
1.1 Objectives

The objectives of this paper are twofold: (i) to introduce a methodology to calculate the unearned premium reserves for an extended warranty; and (ii) to describe a general procedure to test the adequacy of the calculated unearned premium reserves and estimate the allowable deferred policy acquisition expenses.

Section 2 discusses the differences between unearned premium reserves calculated on an exposure-adjusted basis or on a prorata basis. The extended warranty data are introduced in Section 3. The data are organized by the effective month of the manufacturer warranty. In Section 4 the method used to calculate the unearned premium reserves for an automobile extended warranty contract is discussed. Section 5 deals with testing the adequacy of unearned premium reserves. Unearned premium reserves plus future investment income derived thereof are compared against future claims and expenses to determine if premium deficiency exists. Investment income is estimated from interest-bearing assets, taking into account credit risk, interest rate risk, and payment pattern risk. Section 6 discusses the treatment of deferred policy acquisition expenses. In U.S. and Canadian GAAP financial statements, insurance companies are allowed to defer policy acquisition expenses to the extent they meet the test of recoverability. The impact of reinsurance on a mono line warranty insurance company's balance sheet is discussed. Section 7 gives a summary and conclusions. The appendix contains details of FASB Statement No. 60 (FASB60), CICA Accounting Guideline No. 3 (AcG3), and the nature of credit risk, interest rate risk, and payment pattern risk.

2 Unearned Premium Reserves

An insurance company can calculate the unearned premium reserves of an extended warranty insurance program on an exposure-adjusted basis or on a prorata basis. As the manufacturer provides the first three-year/36,000 mile coverage, the insurer's exposure in the first three years is minimal because few drivers will exceed 36,000 miles in this period. A prorata unearned premium reserve formula will overstate the earned premiums in the first three years. From an income and outgo matching perspective, the exposure-adjusted basis is a better approach. Under the exposure-adjusted approach, premiums are earned in proportion to the emergence of the expected losses; when 5 percent of the ultimate losses are expected to be the cumulative incurred at
the end of year two, the formula should have 95 percent of the written premiums as unearned premiums. As an illustration, a hypothetical six-year/72,000 mile (6/72) extended warranty with an underlying three-year manufacturer warranty might have the following cumulative expected loss, expected earned, and unearned pattern shown in Table 1. The earned pattern, with a proper amortization of acquisition expenses, theoretically would match the income and outgo of the 6/72 contract throughout the life of the extended warranty.

Table 1

<table>
<thead>
<tr>
<th>Cumulative</th>
<th>Time (in Months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Losses</td>
<td>0%   2%   5%   15%   45%   75%   100%</td>
</tr>
<tr>
<td>Earned Premium</td>
<td>0%   2%   5%   15%   45%   75%   100%</td>
</tr>
<tr>
<td>Unearned Premium</td>
<td>100%  98%  95%  85%  55%  25%   0%</td>
</tr>
</tbody>
</table>

Suppose an insurer has $M$ different types of extended warranty contracts each running for $n$ months. For a contract type $m$, let $P_{im}$ denote the expected monthly pure premium paid at the start of month $i$, $i = 1, 2, \ldots, n$. The expected undiscounted pure premium for contract type $m$ is $P_m$ where

$$P_m = \sum_{i=1}^{n} P_{i,m}.$$  

The unearned premium ratio for a contract type $m$ at $i$ months is:

$$R_{i,m} = 1 - \frac{\sum_{k=1}^{i} P_{k,m}}{P_m} = \frac{\sum_{k=i+1}^{n} P_{k,m}}{P_m}$$  \hspace{1cm} (1)

with $R_{nm} = 0$. Let $G_{i,m}$ represent the written premiums of a group of type $m$ extended warranties that are $i$ months old; then, the total unearned premiums is:

$$U_m = \sum_{i=1}^{n} R_{i,m} G_{i,m}.$$  \hspace{1cm} (2)

The total unearned premiums of the entire program is:
Equations (1), (2), and (3) hold true for either the prorata method or the exposure-adjusted method. In the case of the prorata method $P_{1,m} = P_{2,m} = P_{3,m} = \ldots = P_{n,m}$ for contract type $m$.

Under the prorata method, premiums are earned in proportion to the time expired on the contract. Notwithstanding its simplicity, the prorata method produces a severe overstatement of premiums earned in the early part of the contract and a corresponding understatement of earned premiums near the end of the contract.

At this time, there is no consensus as to which method is proper. The accounting profession has limited guidance on extended warranty unearned premium reserves. Under National Association of Insurance Commissioners (NAIC) rules, extended warranties with contract terms greater than 13 months are considered as long duration contracts. There is a three-way test to determine the unearned premium reserve. (For the P&C Statutory Statement of Actuarial Opinion, see the American Academy of Actuaries Property and Casualty Practice Note of December 2000 for guidance.)

A straight interpretation of FASB60 would suggest the following two approaches: prorated by months elapsed or by mileage driven. The first approach (Table 2) assumes there is no exposure in the first three years, i.e., no policyholder drives more than 12,000 miles per year. This we know is an implausible assumption. The second approach (Table 2) assumes that one can determine the odometer readings of all policyholders on a valuation date. The second approach is more accurate than the first, but it is impractical. The exposure-adjusted method is a blend of both approaches. When supported by loss experience, the exposure-adjusted method is the only one that follows the intent of FASB60.

Under Canadian GAAP, unearned premium reserves of extended warranties can be computed on an exposure-adjusted basis or prorata basis. Most companies, however, use the exposure-adjusted method.

---

2Developed by the Committee on Property and Liability Financial Reporting of the American Academy of Actuaries, 1100 17th Street NW, 7th Floor, Washington DC 20036.

3Summary of FASB Statement No. 60, paragraph 3 is given in Appendix A.
Table 2
Prorata Earned Exposure Over Extended Warranty Period

<table>
<thead>
<tr>
<th>Cumulative</th>
<th>Time (in Months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mileage (in 1000s)</td>
<td>0</td>
</tr>
<tr>
<td>Earned</td>
<td>0</td>
</tr>
</tbody>
</table>

3 Data Organization

As most extended warranty programs allow the vehicle owner to purchase an extended warranty while the three-year/36,000 mile manufacturer's warranty is in effect, it is convenient to track the exposure and claim payments of an extended warranty by the date of sale of the vehicle rather than by the date of sale of the extended warranty.

Table 3
Historical Data for Contract Type $m$

<table>
<thead>
<tr>
<th>Effective Month ($j$)</th>
<th>Age ($i$)</th>
<th>1/91</th>
<th>2/91</th>
<th>...</th>
<th>...</th>
<th>10/98</th>
<th>11/98</th>
<th>12/98</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$A_{1,1,m}$</td>
<td>$A_{1,2,m}$</td>
<td>...</td>
<td>...</td>
<td>$A_{1,94,m}$</td>
<td>$A_{1,95,m}$</td>
<td>$A_{1,96,m}$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$A_{2,1,m}$</td>
<td>$A_{2,2,m}$</td>
<td>...</td>
<td>...</td>
<td>$A_{2,94,m}$</td>
<td>$A_{2,95,m}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$A_{3,1,m}$</td>
<td>$A_{3,2,m}$</td>
<td>...</td>
<td>...</td>
<td>$A_{3,94,m}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>73</td>
<td>$A_{73,1,m}$</td>
<td>$A_{73,2,m}$</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Notes: $A_{i,j,m} =$ Total claim amount paid during the month $i$ from type $m$ contracts with effective month $j$ and age $i = 1 +$ Valuation Month/Year $- j$.

Claim payments are used here in lieu of incurred claim amounts for two reasons: (i) outstanding claim estimates are not exact and actual payments after the valuation date (e.g., December 31, 1998) may differ from the original estimates; and (ii) a small number of rejected claims is submitted for a second adjudication, and some of them are approved for payments later. Table 3 shows sample historical data for contract type $m$, and Tables 4 and 5 show the actual data for a two-year/24,000 mile plan with a one-year/12,000 mile manufacturer warranty. Tables 4 and 5 use data points from 12/96 to 12/98.
<table>
<thead>
<tr>
<th>Age</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/96</td>
<td>100</td>
<td>105</td>
<td>94</td>
<td>95</td>
<td>89</td>
<td>92</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/97</td>
<td>100</td>
<td>105</td>
<td>94</td>
<td>95</td>
<td>89</td>
<td>92</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/97</td>
<td>100</td>
<td>105</td>
<td>94</td>
<td>95</td>
<td>89</td>
<td>92</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/97</td>
<td>100</td>
<td>105</td>
<td>94</td>
<td>95</td>
<td>89</td>
<td>92</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5/97</td>
<td>100</td>
<td>105</td>
<td>94</td>
<td>95</td>
<td>89</td>
<td>92</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6/97</td>
<td>100</td>
<td>105</td>
<td>94</td>
<td>95</td>
<td>89</td>
<td>92</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/97</td>
<td>100</td>
<td>105</td>
<td>94</td>
<td>95</td>
<td>89</td>
<td>92</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/97</td>
<td>100</td>
<td>105</td>
<td>94</td>
<td>95</td>
<td>89</td>
<td>92</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/97</td>
<td>100</td>
<td>105</td>
<td>94</td>
<td>95</td>
<td>89</td>
<td>92</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/97</td>
<td>500</td>
<td>525</td>
<td>470</td>
<td>475</td>
<td>445</td>
<td>460</td>
<td>655</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/97</td>
<td>500</td>
<td>525</td>
<td>470</td>
<td>475</td>
<td>445</td>
<td>460</td>
<td>655</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/97</td>
<td>500</td>
<td>525</td>
<td>470</td>
<td>475</td>
<td>445</td>
<td>460</td>
<td>655</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13/97</td>
<td>12000</td>
<td>12600</td>
<td>11280</td>
<td>11400</td>
<td>10680</td>
<td>11040</td>
<td>15720</td>
<td>16800</td>
<td>8160</td>
<td>13680</td>
<td>12360</td>
<td>12240</td>
</tr>
<tr>
<td>14/97</td>
<td>12000</td>
<td>12600</td>
<td>11280</td>
<td>11400</td>
<td>10680</td>
<td>11040</td>
<td>15720</td>
<td>16800</td>
<td>8160</td>
<td>13680</td>
<td>12360</td>
<td>12240</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>15</td>
<td>12000</td>
<td>12600</td>
<td>11280</td>
<td>11400</td>
<td>10680</td>
<td>11040</td>
<td>15720</td>
<td>16800</td>
<td>8160</td>
<td>13680</td>
<td>12360</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>18000</td>
<td>18900</td>
<td>16920</td>
<td>17100</td>
<td>16020</td>
<td>16560</td>
<td>23580</td>
<td>25200</td>
<td>12240</td>
<td>20520</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>18000</td>
<td>18900</td>
<td>16920</td>
<td>17100</td>
<td>16020</td>
<td>16560</td>
<td>23580</td>
<td>25200</td>
<td>12240</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>23000</td>
<td>24150</td>
<td>21620</td>
<td>21850</td>
<td>20470</td>
<td>21160</td>
<td>30130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>23000</td>
<td>24150</td>
<td>21620</td>
<td>21850</td>
<td>20470</td>
<td>21160</td>
<td>30130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>23000</td>
<td>24150</td>
<td>21620</td>
<td>21850</td>
<td>20470</td>
<td>21160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>23000</td>
<td>24150</td>
<td>21620</td>
<td>21850</td>
<td>20470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>32500</td>
<td>34125</td>
<td>30550</td>
<td>30875</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>32500</td>
<td>34125</td>
<td>30550</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>32500</td>
<td>34125</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>95000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#Ins  10000  10500  9400  9500  8900  9200  13100  14000  6800  11400  10300  10200

Notes: Age = Age of Contracts; and #Ins = Number of Insured Contracts by Manufacturer Effective Month.
Table 5

*Historical Claim Amount by Manufacturer Effective Month and Age of Warranty (12/97-12/98)*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>135</td>
<td>115</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td>64</td>
<td>120</td>
<td>115</td>
<td>147</td>
<td>65</td>
<td>110</td>
<td>96</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>135</td>
<td>115</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td>64</td>
<td>120</td>
<td>115</td>
<td>147</td>
<td>65</td>
<td>110</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>135</td>
<td>115</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td>64</td>
<td>120</td>
<td>115</td>
<td>147</td>
<td>65</td>
<td>110</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>135</td>
<td>115</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td>64</td>
<td>120</td>
<td>115</td>
<td>147</td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>135</td>
<td>115</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td>64</td>
<td>120</td>
<td>115</td>
<td>147</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>135</td>
<td>115</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td>64</td>
<td>120</td>
<td>115</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>135</td>
<td>115</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td>64</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>135</td>
<td>115</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>135</td>
<td>115</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>675</td>
<td>575</td>
<td>425</td>
<td>435</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>675</td>
<td>575</td>
<td>425</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>675</td>
<td>575</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>16200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5 (Continued)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#Ins</td>
<td>13500</td>
<td>11500</td>
<td>8500</td>
<td>8700</td>
<td>9500</td>
<td>6400</td>
<td>12000</td>
<td>11500</td>
<td>14700</td>
<td>6500</td>
<td>11000</td>
<td>9600</td>
<td>8300</td>
</tr>
</tbody>
</table>

Notes: Age = Age of Contracts; and #Ins = Number of Insured Contracts by Manufacturer Effective Month. No data available for ages 15 to 25.
4 Methodology and Assumptions

First, the exposures (in contract months) must be determined. Let $E_{i,j,m}$ be the number of exposures for a specific contract type $m$, age $i$ months, and effective month $j$. For a given effective month (based on manufacturer's warranty effective date) and contract type, the number of exposures $E_{i,j,m}$ can be projected for each month subsequent to its effective month. Lapses are ignored in this projection.

As an example, there are 1,000 type $m$ contracts in a six-year/72,000 mile program with effective month in July 1991. The exposures are projected in Table 6. These projections assume that all contracts are effective on the first day of each month. The extra month (73rd month) is used to capture all late payments or repairs done in the last month of the contract. These projections also assume that after a cooling off period (usually 60 days for consumers to reverse their impulsive decisions to purchase extended warranties), the extended warranty count remains constant until expiration. In practice a small percentage of warranties are cancelled mid-term because their underlying vehicles have been written off in accidents. Ignoring these cancellations will not have a material effect on the future claim projections because

$\text{Expected Future Claims} = \text{Pure Premium} \times \text{Exposure.}$

Though the exposure term is overstated by the inclusion of canceled extended warranties, the pure premium term is understated by roughly the same percentage. (The no-lapse assumption can be removed if one
keeps track of exposures, not only by effective month and contract type, but also by age of each contract.) For the rest of this paper the no lapse assumption is used.

From the data, the monthly pure premiums can be estimated by age for each contract as follows:

\[ N_{i,j,m} = \text{Number of claims in month } i \text{ of the contract term for type } m \text{ contracts with effective dates in month } j; \]

\[ E_{i,j,m} = \text{Number of type } m \text{ contracts in force in month } i \text{ of the contract term with effective dates in month } j; \]

\[ A_{i,j,m} = \text{Total actual claim payments made in month } i \text{ of the contract term for type } m \text{ contract with effective dates in month } j; \]

\[ P_{i,m} = \text{Average pure premium in month } i \text{ for the contract type } m. \]

Thus

\[
P_{i,m} = \text{Claim Frequency } \times \text{Average Claim Size} = \frac{\sum_j N_{i,j,m}}{\sum_j E_{i,j,m}} \times \frac{\sum_j A_{i,j,m}}{\sum_j N_{i,j,m}} = \frac{\sum_j A_{i,j,m}}{\sum_j E_{i,j,m}}. \tag{4}
\]

The average pure premium is usually calculated using the last 12 calendar months of data available for each age \( i \). If it is necessary to use more than 12 months of data, some inflation adjustment to equation (4) is needed. For newer contracts, the data have not reached the part of the contract term when claims are more likely to be made. Therefore, the pure premiums have to be estimated from the more mature contracts with similar features. In all cases, \( P_{i,m} \) should be smoothed and adjusted to the valuation date cost level. The resulting \( P_{i,m} \) becomes the expected monthly pure premiums for contract type \( m \).

Using a six-year/72,000 mile contract as an illustration, with only one type of contract, i.e., \( m = 1 \), the monthly expected pure premiums are \( P_{1,1} \) through \( P_{73,1} \). The expected undiscounted pure premium of a six-year/72,000 mile contract that is \( i \) months old is:

\[
\text{Expected Pure Premium} = \sum_{k=i+1}^{73} P_{k,1}. \tag{5}
\]
Let $n$ be the valuation month. It is assumed that $E_{i,j,1} = E_{i,.,1}$ only when $j = n - i + 1$. Assuming there are $E_{i+1,.,1}$ contracts that are $i$ months old, the expected payments of these contracts is:

$$\text{Expected Payments} = \sum_{k=i+1}^{73} E_{i+1,.,1} P_{k,1}. \tag{6}$$

Let us assume that the valuation date is December 31, 1998 and that there are $E_{73,.,1}$ contracts effective (in-force) in January 1993, ..., $E_{25,.,1}$ contracts effective in January 1997 ... $E_{2,.,1}$ contracts effective in December 1998. As there usually is a cost inflation in warranty repairs due to the fact that the cost of parts and labor tend to increase over time, and $P_{i,1}$ from equation (6) is at the December 1998 cost level, it follows that $P_{i,1}$ must be adjusted for inflation after the valuation date. Let $r$ denote the monthly inflation rate. Equation (7), the expected undiscounted inflation-adjusted payments for contracts that are $i$ months old, $(\text{EUIAP}_i)$, is given by:

$$\text{EUIAP}_i = E_{i+1,.,1} \sum_{k=i+1}^{73} P_{k,1} (1 + r)^{k-i}. \tag{7}$$

The expected undiscounted inflation-adjusted payments for contracts with four years to expiry, $(\text{EUIAP}_{24})$, is expanded as in Table 7. The total expected losses (i.e., the total expected undiscounted inflation-adjusted payments) from all 6/72 contracts (after the valuation date) is given by

$$\text{Total Expected Losses} = \sum_{i=1}^{72} \text{EUIAP}_i. \tag{8}$$

Figure 1 shows the $(\text{EUIAP})$ calculation for all 6/72 contracts. The rows in the upper triangle represent the age of the contracts, and the columns represent the effective month of the contracts. The sum of column $i$ is $(\text{EUIAP}_i)$. Each diagonal, however, represents a calendar month of payments starting from January 1999. The upper triangle can be re-oriented so that each diagonal becomes a row corresponding to the calendar month in which payments are expected. (See the lower triangle.)
Figure 1
Expected Undiscounted Inflation-Adjusted Payment Calculation for all 6/72 Contracts

<table>
<thead>
<tr>
<th>Age</th>
<th>Effective Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jan. 93</td>
</tr>
<tr>
<td>2</td>
<td>Dec. 98</td>
</tr>
<tr>
<td>73</td>
<td>Dec. 98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Payment in</th>
<th>Effective Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 99</td>
<td>Dec. 98</td>
</tr>
<tr>
<td>Feb. 99</td>
<td>Dec. 04</td>
</tr>
</tbody>
</table>
### Table 7

**Expected Undiscounted Inflation-Adjusted Payments**

*For Contracts with Four Years to Expiry, EUIAP*$_{24}$

<table>
<thead>
<tr>
<th>Claim Age</th>
<th>Payment Month</th>
<th>Pure Prem</th>
<th>Inflation Factor</th>
<th>Exp</th>
<th>Expected Payments</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Jan 1999</td>
<td>$P_{25,1}$</td>
<td>$(1 + r)$</td>
<td>$E_{25,\cdot,1}$</td>
<td>$(1 + r)P_{25,1}E_{25,\cdot,1}$</td>
</tr>
<tr>
<td>26</td>
<td>Feb 1999</td>
<td>$P_{26,1}$</td>
<td>$(1 + r)^2$</td>
<td>$E_{25,\cdot,1}$</td>
<td>$(1 + r)^2P_{26,1}E_{25,\cdot,1}$</td>
</tr>
<tr>
<td>27</td>
<td>Mar 1999</td>
<td>$P_{27,1}$</td>
<td>$(1 + r)^3$</td>
<td>$E_{25,\cdot,1}$</td>
<td>$(1 + r)^3P_{27,1}E_{25,\cdot,1}$</td>
</tr>
<tr>
<td>28</td>
<td>Apr 1999</td>
<td>$P_{28,1}$</td>
<td>$(1 + r)^4$</td>
<td>$E_{25,\cdot,1}$</td>
<td>$(1 + r)^4P_{28,1}E_{25,\cdot,1}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>Dec 2002</td>
<td>$P_{72,1}$</td>
<td>$(1 + r)^{48}$</td>
<td>$E_{25,\cdot,1}$</td>
<td>$(1 + r)^{48}P_{72,1}E_{25,\cdot,1}$</td>
</tr>
<tr>
<td>73</td>
<td>Jan 2003</td>
<td>$P_{73,1}$</td>
<td>$(1 + r)^{49}$</td>
<td>$E_{25,\cdot,1}$</td>
<td>$(1 + r)^{49}P_{73,1}E_{25,\cdot,1}$</td>
</tr>
</tbody>
</table>

**Notes:** Prem = Premium; and Exp = Exposure.

Equation (8) essentially sums each column of Figure 1 from right to left. As inflation is applied on a calendar month basis, it is more convenient to use the lower triangle in Figure 1 and sum each row. Thus, equation (8) can be re-written as:

\[
\text{Total Expected Losses} = \sum_{i=1}^{72} \text{EUIAP}_i \\
= \sum_{i=1}^{72} \sum_{k=i+1}^{73} P_{k,1} (1 + r)^{k-i} E_{i+1,\cdot,1} \\
= \sum_{k=2}^{73} \sum_{i=k}^{73} P_{i,1} E_{i+2-k,\cdot,1} (1 + r)^{k-1}. \quad (9)
\]

5 Adequacy of Unearned Premiums

The basic approach to testing the adequacy of unearned premiums [see, for example, Cantin and Trahan (1999)] is to compare two quantities: (i) the sum of unearned premiums and investment income from the funds backing the liabilities, and (ii) the sum of expected losses, claim adjustment expenses, and policy maintenance expenses. If the sum in (i) exceeds the sum in (ii), then the test indicates that the unearned premium reserves are adequate and some acquisition expenses may be de-
ferrable for GAAP financial reporting. If the sum in (ii) exceeds the sum in (i), then the test indicates that there is a premium deficiency. When there is premium deficiency, both U.S. and Canadian GAAP require a premium deficiency provision. A premium deficiency first should be recognized by writing off any unamortized deferred policy acquisition expenses to the extent required. If the premium deficiency is greater than the unamortized deferred policy acquisition expenses, a separate liability should be provided for the excess deficiency. This has the same effect as increasing the unearned premium reserves to meet the future claims and expense obligations.

As extended warranties usually have terms shorter than seven years, it is reasonable to use a portion of the company's bond portfolio to support the unearned premium reserves. The expected investment yield of this portfolio must be estimated in order to determine the future investment income attributable to the assets supporting the unearned premium reserves. Besides the portfolio market yield, there are several risks to consider: credit risk, interest rate risk, claim payment pattern risk, liquidity risk, and foreign exchange risk. Because most insurance companies invest in high-grade bonds and because extended warranty claims tend to be in one currency, liquidity and foreign exchange risks are ignored in this paper. The expected investment yield of the bond portfolio is first obtained and then the margins for credit risk, interest rate risk, and payment pattern risk are subtracted to obtain an estimate of the expected yield of this portfolio. The following example illustrates how the expected yield is estimated.

(1) Market yield (annual) of portfolio 5.75%
(2) Credit risk (annual) of portfolio 0.10%
(3) Interest rate (annual) risk 0.30%
(4) Payment pattern (annual) risk 0.35%
(5) Expected yield (annual): (1) − (2) − (3) − (4) 5.00%
(6) Expected yield (monthly) 0.4074%

Once the expected investment yield and the expected losses are known, the run-off experience of the warranty program can be forecasted. Starting with the market value of the bonds backing the unearned premium reserves, monthly claim payments, claim adjustment

---

4 Relevant sections of FASB60 and ACG3 are reproduced in Appendix A.
5 The credit risk, claim payment pattern risk, and interest rate risk are often referred to by actuaries as the C-1 risk, C-2 risk, and the C-3 risk, respectively.
6 The margin calculation for each risk category is discussed in Appendix B.
expenses, policy maintenance expenses are deducted and monthly investment income is added to the account. These calculations assume that the payments are made in the middle of the month and investment income is the product of average monthly assets and the selected yield. Let

\[ A_1 = \text{Total asset value on 1/1/99, i.e., at the start of the first month}; \]

\[ A_i = \text{Total asset value at the start of month } i; \]

\[ C_i = \text{Total claim payments made during month } i \text{ (from equation (9))}; \]

\[ \text{CAE}_i = \text{Total claim adjustment expenses during month } i, \text{ usually a percentage of } C_i; \]

\[ \text{PME}_i = \text{Total policy maintenance expenses to keep policies in force, usually a flat amount or a percentage of unearned premiums; and} \]

\[ I_i = \text{Total investment income earned during month } i; \text{ i.e.,} \]

\[ I_i = 0.4074\% \times \frac{1}{2} [2A_i - (C_i + \text{CAE}_i + \text{PME}_i)]. \]

It follows

\[ A_{i+1} = A_i - C_i - \text{CAE}_i - \text{PME}_i + I_i. \] (10)

If the final surplus run-off, \( A_{73} \), is negative, there is a premium deficiency. Otherwise, the unearned premiums are adequate.

6 Deferred Policy Acquisition Expenses (DPAE)

6.1 DPAE Before Reinsurance

Extended warranty is a single premium policy. Acquisition costs are paid upfront. If they are expensed in the year when the policy is written, there will be a large operating loss in that year. U.S. and Canadian GAAP allow deferral of acquisition expenses, provided they meet the test of recoverability. There are two parts to the test. The first part tests whether there is a reasonable expectation that the insurer will recover some of the acquisition expenses (e.g., brokerage/commission/premium tax,
etc.) if a policy is canceled. The second part tests whether the insurer can expect a reasonable profit when all the extended warranties expire. If both questions are answered affirmatively, then some policy acquisition expenses are deferrable. The amount that is deferrable, however, is still unknown. A reasonable inference from the guidance on premium deficiency (FASB60, paragraph 32, CICA-AcG3 paragraphs 5, 8, and 10) is that unearned premiums less DPAE ought to be sufficient to discharge future claims and expenses related to the in force business. That is, DPAE should not exceed the surplus in the run-off. Also, expenses that have not been incurred cannot be deferred. Therefore the allowable deferred policy acquisition expenses should be limited to the lesser of:

(a) The surplus ($A_{73}$) in the run-off; or

(b) Acquisition expense ratio times unearned premium reserves.

The following illustrates this approach for a 6/72 contract at the end of year two. In practice, the DPAE calculation is only done for the entire extended warranty program, not a portion of it.

(1) Written Premium $105.26
(2) Acquisition Expenses Paid $42.11
(3) Acquisition Expense Ratio, (2)/(1) 40.00%
(4) Unearned Premiums [95% x (1)] $100.00
(5) Expected Losses\(^7\) $50.00
(6) Claim Adjustment Expenses $5.00
(7) Policy Maintenance Expenses $2.00
(8) Investment Income (see Appendix D) $16.00
(9) Expected Surplus in Run-off, (4) - (5) - (6) - (7) + (8) $59.00
(10) Allowable DPAE, Minimum [(3) x (4), (9)] $40.00

The details of the run-off are shown in Table 8.

\(^7\)For most property-casualty insurance policies, expected losses are derived as expected loss ratio times unearned premiums.
## Table 8
Run-Off of an Extended Warranty Program with $\text{PME}_i = 0.04167$

<table>
<thead>
<tr>
<th>Month</th>
<th>$A_i$</th>
<th>$C_i$</th>
<th>$\text{CAE}_i$</th>
<th>$I_i$</th>
<th>$A_{i+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100.00000</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40633</td>
<td>99.88221</td>
</tr>
<tr>
<td>2</td>
<td>99.88221</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40585</td>
<td>99.76394</td>
</tr>
<tr>
<td>3</td>
<td>99.76394</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40537</td>
<td>99.64519</td>
</tr>
<tr>
<td>4</td>
<td>99.64519</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40489</td>
<td>99.52595</td>
</tr>
<tr>
<td>5</td>
<td>99.52595</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40440</td>
<td>99.40623</td>
</tr>
<tr>
<td>6</td>
<td>99.40623</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40391</td>
<td>99.28602</td>
</tr>
<tr>
<td>7</td>
<td>99.28602</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40342</td>
<td>99.16532</td>
</tr>
<tr>
<td>8</td>
<td>99.16532</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40293</td>
<td>99.04413</td>
</tr>
<tr>
<td>9</td>
<td>99.04413</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40244</td>
<td>98.92245</td>
</tr>
<tr>
<td>10</td>
<td>98.92245</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40194</td>
<td>98.80026</td>
</tr>
<tr>
<td>11</td>
<td>98.80026</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40144</td>
<td>98.67759</td>
</tr>
<tr>
<td>12</td>
<td>98.67759</td>
<td>0.43860</td>
<td>0.04386</td>
<td>0.40094</td>
<td>98.55441</td>
</tr>
<tr>
<td>13</td>
<td>98.55441</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.39848</td>
<td>97.46385</td>
</tr>
<tr>
<td>14</td>
<td>97.46385</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.39403</td>
<td>96.36885</td>
</tr>
<tr>
<td>15</td>
<td>96.36885</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.38957</td>
<td>95.26939</td>
</tr>
<tr>
<td>16</td>
<td>95.26939</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.38509</td>
<td>94.16545</td>
</tr>
<tr>
<td>17</td>
<td>94.16545</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.38060</td>
<td>93.05701</td>
</tr>
<tr>
<td>18</td>
<td>93.05701</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.37608</td>
<td>91.94406</td>
</tr>
<tr>
<td>19</td>
<td>91.94406</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.37155</td>
<td>90.82657</td>
</tr>
<tr>
<td>20</td>
<td>90.82657</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.36699</td>
<td>89.70453</td>
</tr>
<tr>
<td>21</td>
<td>89.70453</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.36242</td>
<td>88.57791</td>
</tr>
<tr>
<td>22</td>
<td>88.57791</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.35783</td>
<td>87.44671</td>
</tr>
<tr>
<td>23</td>
<td>87.44671</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.35322</td>
<td>86.31090</td>
</tr>
<tr>
<td>24</td>
<td>86.31090</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.34860</td>
<td>85.17047</td>
</tr>
</tbody>
</table>

The deferral of acquisition expenses does not affect the insurer's liabilities. It creates an asset\(^8\) on the insurer's balance sheet. As a result, the expenses charged to the income statement for an extended warranty in year one are reduced substantially.

---

\(^8\)Deferred policy acquisition expenses are classified as an asset (FASB60 paragraph 29, CICA, AcG3 paragraph 10).
### Table 8 (Continued)

Run-Off of an Extended Warranty Program with $\text{PME}_i = 0.04167$

<table>
<thead>
<tr>
<th>Month $i$</th>
<th>$A_i$</th>
<th>$C_i$</th>
<th>CAE$_i$</th>
<th>$I_i$</th>
<th>$A_{i+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>85.17047</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.34395</td>
<td>84.02538</td>
</tr>
<tr>
<td>26</td>
<td>84.02538</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.33929</td>
<td>82.87563</td>
</tr>
<tr>
<td>27</td>
<td>82.87563</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.33460</td>
<td>81.72120</td>
</tr>
<tr>
<td>28</td>
<td>81.72120</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.32990</td>
<td>80.56206</td>
</tr>
<tr>
<td>29</td>
<td>80.56206</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.32518</td>
<td>79.39821</td>
</tr>
<tr>
<td>30</td>
<td>79.39821</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.32044</td>
<td>78.22961</td>
</tr>
<tr>
<td>31</td>
<td>78.22961</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.31567</td>
<td>77.05624</td>
</tr>
<tr>
<td>32</td>
<td>77.05624</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.31089</td>
<td>75.87810</td>
</tr>
<tr>
<td>33</td>
<td>75.87810</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.30609</td>
<td>74.69516</td>
</tr>
<tr>
<td>34</td>
<td>74.69516</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.30127</td>
<td>73.50740</td>
</tr>
<tr>
<td>35</td>
<td>73.50740</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.29644</td>
<td>72.31480</td>
</tr>
<tr>
<td>36</td>
<td>72.31480</td>
<td>1.31579</td>
<td>0.13158</td>
<td>0.29158</td>
<td>71.11735</td>
</tr>
<tr>
<td>37</td>
<td>71.11735</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.28719</td>
<td>70.15673</td>
</tr>
<tr>
<td>38</td>
<td>70.15673</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.28328</td>
<td>69.19220</td>
</tr>
<tr>
<td>39</td>
<td>69.19220</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.27935</td>
<td>68.22374</td>
</tr>
<tr>
<td>40</td>
<td>68.22374</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.27540</td>
<td>67.25133</td>
</tr>
<tr>
<td>41</td>
<td>67.25133</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.27144</td>
<td>66.27497</td>
</tr>
<tr>
<td>42</td>
<td>66.27497</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.26746</td>
<td>65.29462</td>
</tr>
<tr>
<td>43</td>
<td>65.29462</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.26347</td>
<td>64.31028</td>
</tr>
<tr>
<td>44</td>
<td>64.31028</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.25946</td>
<td>63.32193</td>
</tr>
<tr>
<td>45</td>
<td>63.32193</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.25543</td>
<td>62.32956</td>
</tr>
<tr>
<td>46</td>
<td>62.32956</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.25139</td>
<td>61.33314</td>
</tr>
<tr>
<td>47</td>
<td>61.33314</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.24733</td>
<td>60.33266</td>
</tr>
<tr>
<td>48</td>
<td>60.33266</td>
<td>1.09649</td>
<td>0.10965</td>
<td>0.24325</td>
<td>59.32811</td>
</tr>
<tr>
<td></td>
<td>50.00000</td>
<td>5.00000</td>
<td>16.00000</td>
<td>59.00000</td>
<td></td>
</tr>
</tbody>
</table>

Notes: PME$_i$ is policy maintenance expenses in month $i$. 

**Rounded** 50.00000 5.00000 16.00000 59.00000
Cheng: Automobile Warranty

As deferrable expenses are expressed as a percentage of unearned premiums, the choice of a prorata or an exposure-adjusted method affects the amount of deferred policy acquisition expenses. The use of the prorata method, however, can lead to a premium deficiency situation in the latter part of the extended warranty program because the insurer has declared too much profit in the early part of the program.

6.2 DPAE After Reinsurance

Thus far no reinsurance has been assumed in the above calculations. As warranty is a high frequency and low severity class, reinsurance, if applicable, will tend to be quota share or aggregate stop loss in nature. The effect of reinsurance on DPAE is best illustrated with the example provided below:

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>Ceded</th>
<th>Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Unearned Premiums</td>
<td>$100.00</td>
<td>$75.00</td>
<td>$25.00</td>
</tr>
<tr>
<td>(2) Expected Losses</td>
<td>$50.00</td>
<td>$37.50</td>
<td>$12.50</td>
</tr>
<tr>
<td>(3) Claim Adjustment Expenses</td>
<td>$5.00</td>
<td>$3.75</td>
<td>$1.25</td>
</tr>
<tr>
<td>(4) Policy Maintenance</td>
<td>$2.00</td>
<td>$0.00</td>
<td>$2.00</td>
</tr>
<tr>
<td>(5) Deferrable Expenses</td>
<td>$40.00</td>
<td>$26.25</td>
<td>$13.75</td>
</tr>
<tr>
<td>(6) Investment Income</td>
<td>$16.00</td>
<td>N/A</td>
<td>$4.00</td>
</tr>
<tr>
<td>(7) Expected Surplus in Run-off</td>
<td>$59.00</td>
<td>N/A</td>
<td>$13.25</td>
</tr>
<tr>
<td>(8) Allowable DPAE, Min. [(5), (7)]</td>
<td>$40.00</td>
<td></td>
<td>$13.25</td>
</tr>
<tr>
<td>(9) Unearned Commissions</td>
<td></td>
<td>$26.25</td>
<td></td>
</tr>
</tbody>
</table>

Notice that the program, before 75 percent quota share reinsurance, generates enough surplus ($59) to allow the insurer to defer 40 percent of the unearned premiums. In this quota share reinsurance transaction, the insurer receives 35 percent ceding commissions and an agreement to share claim adjustment expenses on a prorata basis. The net acquisition expense ratio after reinsurance is 55 percent (13.75/25), higher than the 40 percent obtained on a direct basis. Furthermore, the cash flow (as a percentage of the unearned premiums) is reduced due to the 100 percent retained policy maintenance expenses; investment income is reduced to $4. Consequently, the surplus in the run-off is reduced to $13.25 as opposed to 25 percent of $59 on a direct basis (i.e., $14.75).

The net allowable DPAE is $13.25, being the lesser of deferrable expenses ($13.75) and the surplus in the run-off ($13.25).
6.3 Impact of Reinsurance on DPAE

It is worthwhile to look at the insurer's balance sheet before and after reinsurance. Before reinsurance, there are unearned premiums of $100 and DPAE of $40. The balance sheet is shown in Table 9. After reinsurance, the balance sheet (on GAAP gross up basis) is shown in Table 10.

### Table 9

<table>
<thead>
<tr>
<th>The Balance Sheet Before Reinsurance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assets</strong></td>
</tr>
<tr>
<td>Bonds</td>
</tr>
<tr>
<td>Ceded Unearned Premiums</td>
</tr>
<tr>
<td>Ceded Unpaid Claims</td>
</tr>
<tr>
<td>DPAE</td>
</tr>
<tr>
<td><strong>Total</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Liabilities &amp; Shareholders' Equity</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unearned Premiums</td>
</tr>
<tr>
<td>Unpaid Claims</td>
</tr>
<tr>
<td>Unearned Commissions</td>
</tr>
<tr>
<td>Shareholders' Equity</td>
</tr>
<tr>
<td><strong>Total</strong></td>
</tr>
</tbody>
</table>

The net unearned premiums are $25, being $100 on the liability ledger less $75 on the asset ledger. On the valuation date, the reinsurer's ceding commissions are classified as unearned commissions to the insurer (i.e., a liability) because they have to be returned if the ceded premiums are returned. The gross DPAE on the asset ledger is no longer $40 because in a run-off the insurer will earn $26.25 commissions from the reinsurer and realize an expected surplus of $13.25 from the net retained premiums. Therefore, the gross up DPAE should not exceed ($26.25 + $13.25) or $39.50. Furthermore, the gross up DPAE should not exceed $40 (the deferrable expenses before reinsurance) because the insurer cannot defer more than its actual deferrable policy acquisition expenses. In this example the first limitation is lower. Therefore, the balance sheet should show $39.50 as the gross allowable DPAE. By entering into a quota share reinsurance, some DPAE ($40 – $39.50) is lost in the form of frictional cost or profit to the reinsurer. Shareholders' equity is also reduced by $0.50.
### Table 10
The Balance Sheet After Reinsurance

<table>
<thead>
<tr>
<th>Assets</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonds</td>
<td>91.25</td>
</tr>
<tr>
<td>Ceded Unearned Premiums</td>
<td>75.00</td>
</tr>
<tr>
<td>Ceded Unpaid Claims</td>
<td>Small</td>
</tr>
<tr>
<td>DPAE</td>
<td>39.50</td>
</tr>
<tr>
<td>Total</td>
<td>205.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Liabilities &amp; Shareholders' Equity</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unearned Premiums</td>
<td>100.00</td>
</tr>
<tr>
<td>Unpaid Claims</td>
<td>Small</td>
</tr>
<tr>
<td>Unearned Commissions</td>
<td>26.25</td>
</tr>
<tr>
<td>Shareholders' Equity</td>
<td>79.50</td>
</tr>
<tr>
<td>Total</td>
<td>205.75</td>
</tr>
</tbody>
</table>

7 Closing Comments

This paper describes an exposure-adjusted methodology to calculate the unearned premium reserves of an automobile extended warranty insurance program and test the adequacy of the calculated reserves. It also presents a general formula to estimate the allowable deferred policy acquisition expenses on a before and after reinsurance basis for all property-casualty insurance companies.

Appendix A—Relevant Aspects of FASB Statement No. 60 and AcG3

FASB Statement No. 60 (Issued June/82)

FASB Statement No. 60\(^9\) extracts the specialized principles and practices from the AICPA insurance industry related guides and statements of position and establishes financial accounting and reporting stan-

---

\(^9\)FASB Statement No. 60, *Accounting and Reporting by Insurance Enterprises*, is copyrighted by the Financial Accounting Standards Board (FASB), 401 Merritt 7, P.O. Box 5116, Norwalk, Connecticut 06856-5116, U.S.A. Portions are reprinted with permission. Complete copies of this document are available from the FASB.
Standards for insurance enterprises other than mutual life insurance enterprises, assessment enterprises, and fraternal benefit societies.

Insurance contracts, for purposes of this statement, need to be classified as short-duration or long-duration contracts. Long-duration contracts include contracts, such as whole-life, guaranteed renewable term life, endowment, annuity, and title insurance contracts that are expected to remain in force for an extended period. All other insurance contracts are considered short-duration contracts and include most property and liability insurance contracts.

Premiums from short-duration contracts ordinarily are recognized as revenue over the period of the contract in proportion to the amount of insurance protection provided. Claim costs, including estimates of costs for claims relating to insured events that have occurred but have not been reported to the insurer, are recognized when insured events occur.

Premiums from long-duration contracts are recognized as revenue when due from policyholders. The present value of estimated future policy benefits to be paid to or on behalf of policyholders less the present value of estimated future net premiums to be collected from policyholders are accrued when premium revenue is recognized. Those estimates are based on assumptions, such as estimates of expected investment yields, mortality, morbidity, terminations, and expenses, applicable at the time the insurance contracts are made. Claim costs are recognized when insured events occur.

Costs that vary with and are primarily related to the acquisition of insurance contracts (acquisition costs) are capitalized and charged to expense in proportion to premium revenue recognized.

Investments are reported as follows: common and nonredeemable preferred stocks at market, bonds and redeemable preferred stocks at amortized cost, mortgage loans at outstanding principal or amortized cost, and real estate at depreciated cost. Realized investment gains and losses are reported in the income statement below operating income and net of applicable income taxes. Unrealized investment gains and losses, net of applicable income taxes, are included in stockholders' (policyholders') equity.

For our purposes, the relevant paragraphs pertaining to acquisition costs are FASB60, paragraphs 28–31:

28. Acquisition costs are those costs that vary with and are primarily related to the acquisition of new and renewal insurance contracts. Commissions and other costs (for example, salaries of certain employees involved in the underwriting and policy issue functions,
and medical and inspection fees) that are primarily related to insurance contracts issued or renewed during the period in which the costs are incurred shall be considered acquisition costs.

29. Acquisition costs shall be capitalized and charged to expense in proportion to premium revenue recognized. To associate acquisition costs with related premium revenue, acquisition costs shall be allocated by groupings of insurance contracts consistent with the enterprise's manner of acquiring, servicing, and measuring the profitability of its insurance contracts. Unamortized acquisition costs shall be classified as an asset.

30. If acquisition costs for short-duration contracts are determined based on a percentage relationship of costs incurred to premiums from contracts issued or renewed for a specified period, the percentage relationship and the period used, once determined, shall be applied to applicable unearned premiums throughout the period of the contracts.

31. Actual acquisition costs for long-duration contracts shall be used in determining acquisition costs to be capitalized as long as gross premiums are sufficient to cover actual costs. However, estimated acquisition costs may be used if the difference is not significant. Capitalized acquisition costs shall be charged to expense using methods that include the same assumptions used in estimating the liability for future policy benefits.

For our purposes, the relevant paragraph pertaining to premium deficiency is FASB60, paragraph 32:

32. A probable loss on insurance contracts exists if there is a premium deficiency relating to short-duration or long-duration contracts. Insurance contracts shall be grouped consistent with the enterprise's manner of acquiring, servicing, and measuring the profitability of its insurance contracts to determine if a premium deficiency exists.
Accounting Guideline AcG-3 Financial Reporting by Property and Casualty Insurance Companies (Issued April 1986)¹⁰

1 The Insurance Industry in Canada is governed by Federal and Provincial statutes. These statutes, supplemented by the regulations and annual statement forms issued thereunder, are primarily designed to monitor the solvency of insurance companies so as to ensure the protection of policyholders. Financial statements prepared for issuance to shareholders, policyholders and other interested parties have, in the past, been greatly influenced by these regulatory requirements.

2 The Canadian Council of Superintendents of Insurance has now affirmed that the basic financial statements in the standardized annual forms filed with them shall be drawn up to enable property and casualty insurance companies to report in accordance with generally accepted accounting principles. This action permits property and casualty insurance companies to follow the Accounting Recommendations in the CICA Handbook in the preparation of financial statements for issuance to shareholders, policyholders and other interested parties.

3 This Guideline covers those areas specific to property and casualty insurance companies that are not directly covered by the Accounting Recommendations in the CICA Handbook and identifies certain other relevant matters that are covered by those Recommendations. It outlines the practices the Accounting Standards Steering Committee considers should be followed by such companies in preparing financial statements in accordance with generally accepted accounting principles.

Deferred policy acquisition expenses

4 Certain costs such as commissions and premium taxes, vary directly with, and are directly related to, the acquisition of business (i.e., new and renewal premiums written during the accounting period) and can be associated directly with specific revenues. Other costs, such as salaries of certain employees involved in underwriting and policy issuance functions, inspection report fees, and fees paid to boards and bureaux, may vary indirectly with the acquisition of business but are directly related to the premiums written.

¹⁰This Guideline is to be read in conjunction with the Introduction to Accounting Guidelines contained in the CICA Handbook and is reproduced with the permission of the Canadian Institute of Chartered Accountants
during the period in which the costs are incurred. These costs meet the criteria for deferral and association with the related premiums as they are earned, provided such costs are expected to be recovered. Certain other costs incurred during the period, such as collection expenses and uncollectible accounts, professional fees and general administrative expenses, do not vary directly with, and are not directly related to, the acquisition of business and therefore are charged to expense as incurred.

5 Deferred policy acquisition expenses may be expected to be recovered to the extent that a premium deficiency is not apparent at the balance sheet date (see paragraphs 8-10).

6 Deferred policy acquisition expenses should be determined by reasonable groupings of business, consistent with an insurer's manner of acquiring, servicing, and measuring the profitability of its business. Each individual company must consider its own situation in determining such groupings. There are situations, for example, where a number of lines of insurance may be written under a composite policy; in this case, it would appear that these lines should be aggregated in determining recoverability. Indeed, total aggregation might be considered unless particular lines are sold and administered in an entirely different fashion from others. Pricing of itself is not considered an appropriate criterion for segmentation; indeed, certain products that are priced differently may support each other in a total business context. Segmentation might be considered, for example, when there is a distinct department for reinsurance assumed, as opposed to direct writing. Within reinsurance assumed, segmentation might be considered as between treaty business and facultative business or between quota share and excess reinsurance. Each case must be assessed on its merits with regard being given to the actual operations of the company concerned.

7 "The Insurance Companies Acquisition Expenses Regulations," applicable to federally registered insurers and certain of the provincial statutes or requirements, limit the deferral of acquisition costs to 30% of the unearned premiums. This is not in accordance with generally accepted accounting principles. The Departments of Insurance, however, permit financial statements to be prepared in accordance with generally accepted accounting principles in this respect provided an appropriation of retained earnings is made in the annual statement forms for amounts in excess of 30%.
Premium deficiencies

8 In those instances where anticipated future claims and expenses exceed unearned premiums, a premium deficiency exists and provision should be made therefor. Anticipated future claims and expenses include expected claims (including adjustment expenses), maintenance expenses (nondeferrable costs which can be attributed to maintaining the policies in force), policyholder dividends (if applicable) and unamortized deferred policy acquisition expenses. Premium deficiencies should be determined by reasonable groupings of business.

9 Consensus has not been reached as to the necessity of including consideration of anticipated investment income on policyholders' funds in the determination of a premium deficiency. Until this matter is resolved, the practice followed should be disclosed in the note on accounting policies.

10 A premium deficiency should first be recognized by writing off any unamortized deferred policy acquisition expenses to the extent required. If the premium deficiency is greater than the unamortized deferred policy acquisition expenses, a separate liability should be provided for the excess deficiency. This procedure acknowledges that where an asset has been impaired, such impairment should be recognized before any additional liabilities are recorded.

Appendix B—Risk Margins in Discount Rates

The three major risks associated with the discount rate in an actuarial valuation are credit risk, payment pattern risk, and the interest rate risk. As was pointed out earlier, the credit risk, claim payment pattern risk, and interest rate risk are often referred to by actuaries as the C-1 risk, C-2 risk, and the C-3 risk, respectively. For an actuarial approach to measuring the C-1 and C-3 risks, see, for example, Sega (1986) and Vanderhoof et al., (1989) for the C-1 risk and Mereu (1989) for the C-3 risk.

Credit Risk

Not all investments are of the same quality. Bond and preferred share issuers are rated by independent firms. Investors, rightly or
wrongly, use this type of information and seasoned judgment in trading these securities in the secondary market. In the United States and Canada, bonds issued or guaranteed by the federal government are the most creditworthy securities. Over a period of time, the yields of other securities will develop their spreads when compared against the treasury or federal bonds. The extra yield over a comparable treasury (i.e., the same maturity and currency) is the implied credit risk determined by the marketplace.

The quotations in Table B1 illustrate how the bond market quantifies credit risk by demanding a higher yield from issuers other than the federal government. Notice that credit risk varies with the issuer as well as the term of maturity.

<table>
<thead>
<tr>
<th>Issuer</th>
<th>Maturity in Years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 Years</td>
</tr>
<tr>
<td>Federal Government</td>
<td>5.33%</td>
</tr>
<tr>
<td>A Utility</td>
<td>5.73%</td>
</tr>
<tr>
<td>A Retailer</td>
<td>6.49%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Issuer</th>
<th>Implied Credit Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 Years</td>
</tr>
<tr>
<td>A Utility</td>
<td>0.40%</td>
</tr>
<tr>
<td>A Retailer</td>
<td>1.16%</td>
</tr>
</tbody>
</table>

In the hypothetical bond portfolio discussed in Section 5, the credit risk (in terms of extra yield) of each bond is weighted by its market value to arrive at an average credit risk. As the hypothetical bond portfolio is composed mostly of federal government bonds and short-term corporate bonds, the implied credit risk is only 0.10 percent.

**Interest Rate Risk (Mismatching Asset/Liability Risk)**

For this section only bonds and T-bills are used as investments. A well-known result in the theory of finance is that a bond portfolio's market value changes inversely proportional to the product of its duration and the change in interest rate. [For more on duration see, for example, Panjer (1998, Section 3.5) or Boyle (1992, Chapter 3).] For example, if a bond portfolio has duration of three years, its value would
rise about 3 percent for a 100 basis point decrease in interest rate. If the expected claims payments should have an identical duration, its present value also will rise 3 percent when the discount rate decreases 100 basis points. When the asset and liability duration are about the same and the yield curve is normal (i.e., long-term bonds yield more than short-term ones), the assets are said to be *immunized* against the interest rate risk. In practice yield curves may become inverted (i.e., short-term bond yields exceed long-term ones). Fortunately, the yield curve seldom remains inverted for long periods. For the remainder of this section, the yield curve is assumed to be normal. Risk to the insurer's surplus arises when there is a mismatch of asset and liability cash flow.

Let us assume market value of assets is equal to present value of claims at the current market yield and both are equal to 1.0. When liability duration \( D_L \) exceeds asset duration \( D_A \), any decrease in interest rate will diminish the surplus of the insurer. Conversely, when asset duration \( D_A \) exceeds liability duration \( D_L \), any increase in interest rate will diminish the surplus of the insurer. For every 100 basis points (bp) change in interest rate, the impact on the insurer's surplus is approximately as follows:

<table>
<thead>
<tr>
<th>Interest Rate Change</th>
<th>( D_L &gt; D_A )</th>
<th>( D_A &gt; D_L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-100 bp</td>
<td>(-(D_L - D_A))</td>
<td>Favorable Effect</td>
</tr>
<tr>
<td>+100 bp</td>
<td>Favorable Effect</td>
<td>(-(D_A - D_L))</td>
</tr>
</tbody>
</table>

The risk (adverse effect) in both cases is approximately the absolute value of \( D_A - D_L \). Because the liability duration is \( D_L \), one can increase the discounted liability value by approximately \(|D_A - D_L| / D_L\) if the discount rate is reduced by \(|D_A - D_L| / D_L\). In so doing, 100 bp interest rate risk can be absorbed into the discounted liability estimate. If the anticipated change in interest rate is \( t \) bp, the discount rate should be reduced by \((|D_A - D_L| / D_L) \times t / 100\).

Finally, if asset value is higher than liability value, one needs to cover only a portion of the assets for the interest rate risk (i.e., present value of liability/market value of assets).

Therefore, interest rate risk can be quantified approximately as the absolute value of:

\[
\text{Interest Rate Risk} = \text{COV\%} \times \left( \frac{D_A - D_L}{D_L} \right) \times \Delta\text{INT}
\]

where \( \Delta\text{INT} \) is the anticipated change in interest rate, and
Cheng: Automobile Warranty

\[ \text{COV\%} = \frac{\text{Present Value of Future Claims and Expenses}}{\text{Market Value of Investments Used in Calculations}}. \]

Tables B2 and B3 show the results of our formula in two situations:

### Table B2

**Actual Vs. Estimated Interest Rate Risk**  
**When** \( D_A = 1.000 \) **and** \( D_L = 1.941 \)

<table>
<thead>
<tr>
<th>( \Delta \text{INT (in %)} )</th>
<th>INT MV of Assets</th>
<th>PV of Liabs</th>
<th>Actual (1)</th>
<th>Approx (2)</th>
<th>AbsEr</th>
<th>RelEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>83.115</td>
<td>83.115</td>
<td>0.682</td>
<td>0.694</td>
<td>0.012</td>
</tr>
<tr>
<td>-1</td>
<td>9</td>
<td>83.877</td>
<td>84.559</td>
<td>1.398</td>
<td>1.400</td>
<td>0.002</td>
</tr>
<tr>
<td>-2</td>
<td>8</td>
<td>84.654</td>
<td>86.052</td>
<td>2.149</td>
<td>2.116</td>
<td>-0.033</td>
</tr>
<tr>
<td>-3</td>
<td>7</td>
<td>85.445</td>
<td>87.594</td>
<td>2.938</td>
<td>2.844</td>
<td>-0.094</td>
</tr>
<tr>
<td>-4</td>
<td>6</td>
<td>86.251</td>
<td>89.189</td>
<td>2.938</td>
<td>2.844</td>
<td>-0.094</td>
</tr>
</tbody>
</table>

**Notes:** \( \Delta \text{INT} = \text{Change in Interest Rate (in \%)}; \) INT = Interest Rate; MV = Market Value; PV of Liabs = Present Value of Liabilities; Approx = Approximation by Formula; AbsEr = Absolute Error of Formula = \( (\text{Approx} - \text{Actual}) \); RelEr = Relative Error of Formula = \( \frac{(\text{Approx} - \text{Actual})}{\text{Actual}} \).

The important information used in constructing these tables is as follows:

(a) **Assets Available from Extended Warranty Premiums** $100;
(b) **Present Value of Future Claims and Expenses** $50;
(c) **Duration of Bonds** 4.0 years;
(d) **Duration of Liability** 2.5 years;
(e) **Interest Rate Risk** \( \frac{50}{100} \times \left( \frac{4.0 - 2.5}{2.5} \right) \times 100\text{bp} \) 30 bp.

**Payment Pattern Risk**

As faster payment means shorter liability duration and a smaller discount, the same effect can be achieved by decreasing the (liability) discount rate. In practice, it is more convenient to lower the discount rate than changing the payment pattern.
Table B3
Actual Vs. Estimated Interest Rate Risk
When $D_A = 1.000$ and $D_L = 0.500$

<table>
<thead>
<tr>
<th>$\Delta INT$ (in %)</th>
<th>INT</th>
<th>MV of Assets</th>
<th>PV of Liabs</th>
<th>Actual (1)</th>
<th>Approx (2)</th>
<th>AbsEr</th>
<th>RelEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>95.345</td>
<td>95.346</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td>11</td>
<td>94.486</td>
<td>94.916</td>
<td>0.429</td>
<td>0.436</td>
<td>0.007</td>
<td>0.016</td>
</tr>
<tr>
<td>+2</td>
<td>12</td>
<td>93.643</td>
<td>94.491</td>
<td>0.847</td>
<td>0.879</td>
<td>0.032</td>
<td>0.038</td>
</tr>
<tr>
<td>+3</td>
<td>13</td>
<td>92.814</td>
<td>94.072</td>
<td>1.257</td>
<td>1.327</td>
<td>0.070</td>
<td>0.056</td>
</tr>
<tr>
<td>+4</td>
<td>14</td>
<td>92.000</td>
<td>93.659</td>
<td>1.658</td>
<td>1.782</td>
<td>0.124</td>
<td>0.075</td>
</tr>
</tbody>
</table>

Notes: $\Delta INT = $ Change in Interest Rate (in %); INT = Interest Rate; MV = Market Value; PV of Liabs = Present Value of Liabilities; Approx = Approximation by Formula; AbsEr = Absolute Error of Formula = (2) - (1); RelEr = Relative Error of Formula = ((2) - (1))/1.

Suppose:

- Market Yield of Portfolio 5.75%;
- Liability Duration 4 years;
- Average Discount Factor 0.80;
- Amount of Discount $1 - (1.0575)^{-4} = 20\%$.

If payment pattern is assumed to be three months faster, then,

- Amount of Discount $1 - (1.0575)^{-3.75} = 19\%$;
- Average Discount Factor 0.81.

The implied discount rate is 5.38 percent because $(1.0538)^{-4} = 0.811$. In this case, the risk margin for the payment pattern risk is 5.75 percent less 5.38 percent or 37 basis points. In Section 5, the risk margin is rounded as 35 basis points.
Table B4 shows the relationship between shorter duration and implied discount rate:

<table>
<thead>
<tr>
<th>Average Discount</th>
<th>Duration</th>
<th>Implied Discount Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.800</td>
<td>4.00 years</td>
<td>5.75%</td>
</tr>
<tr>
<td>0.811</td>
<td>3.75 years</td>
<td>5.38%</td>
</tr>
<tr>
<td>0.822</td>
<td>3.50 years</td>
<td>5.01%</td>
</tr>
<tr>
<td>0.834</td>
<td>3.25 years</td>
<td>4.65%</td>
</tr>
<tr>
<td>0.846</td>
<td>3.00 years</td>
<td>4.28%</td>
</tr>
<tr>
<td>0.857</td>
<td>2.75 years</td>
<td>3.92%</td>
</tr>
<tr>
<td>0.870</td>
<td>2.50 years</td>
<td>3.56%</td>
</tr>
</tbody>
</table>

References


Can Utility-Maximization Models Assist With Retirement Planning?
Zaki Khorasanee*

Abstract

Utility-maximization models for optimizing portfolio choices can be subdivided into two classes: those based on maximizing the expected utility of lifetime consumption and those based on maximizing the expected utility of retirement wealth. It is argued that the first type of model, which optimizes both saving and investment decisions, is difficult to apply in practice because of inadequate (or unreliable) information about individual preferences. Although the second type of model only optimizes investment decisions, it is of greater practical value because fewer data on individual preferences are required. The second type of model is used to derive formulae for the optimal portfolio choice at any duration from retirement, assuming that risky investment returns follow a geometric Brownian motion and that the utility function is of the hyperbolic absolute risk aversion (HARA) class. It is shown that individuals who expect to make further contributions to their fund should switch into less risky portfolios on nearing retirement.

Key words and phrases: utility function, portfolio choice, multi-period problem


Mr. Khorasanee's address is: Faculty of Actuarial Science and Statistics, Cass Business School, 106 Bunhill Row, London EC1Y 8TZ, UNITED KINGDOM. Internet address: mzk@city.ac.uk
1 Introduction

The growing popularity of defined-contribution pension plans has created a need for practical methods of advising the members of such plans on their portfolio choices. The mean-variance model for portfolio choice developed by Markowitz (1952) is a special case of a more general multi-period approach based on the maximization of expected utility. The utility-maximization problem can be formulated as one of two models:

- Maximization of the expected utility of lifetime consumption with due allowance for the bequest motive; or
- Maximization of the expected utility of terminal wealth (e.g., at retirement).

Merton (1969, 1971) develops these models in continuous time and derives closed-form solutions for certain classes of utility functions. Although these results are useful as a description of the kinds of behavior we might expect from individuals in a hypothetical equilibrium scenario, it is an open question whether utility-maximization models can be used in a normative way, i.e., as a tool for financial professionals to help individuals with their saving and investing decisions.

The normative use of utility-maximization models for members of defined-contribution pension plans is considered in some detail by Thomson (1998, 2002), who focuses on the second type of model based on the utility of retirement wealth. The utility functions of 49 individuals1 were derived from answers to a standardized questionnaire, and discrete-time dynamic programming was used to obtain optimal portfolio choices for each individual, using a vector autoregressive model for investment returns from different asset-types. As Thomson uses a fairly complex parametric form for the utility function, no simple closed-form solution emerges for the optimal portfolio at any duration from retirement.

In this article, we examine the utility of the lifetime consumption model and conclude that it is unlikely that it could be used in a practical manner to advise individuals on consumption and portfolio choices. Although maximizing the expected utility of retirement wealth is a less generalized approach, we argue that this is how the problem should be formulated in a defined-contribution pension plan when the rate of saving is assumed to be predetermined. We then use discrete-time dynamic

---

1 They were the parents (or other relatives) of South African university students.
programming to derive Merton’s solution for a lump sum investment and extend it to cover the more realistic situation in which the plan member is investing future contributions as well as an initial fund. A graphical presentation of these results, which would allow an individual to optimize his/her portfolio over the period up to retirement, is presented.

2 Utility of the Lifetime Consumption Model

Financial markets allow individuals to redistribute consumption over their lives in order to increase their overall satisfaction. If we add increments to the income of an individual in any single time period, each successive increment will be used to satisfy wants that are less urgently felt. This simple intuition gives rise to the principle of diminishing marginal utility, as discovered in the late nineteenth century by neo-classical economists such as Menger (1871). It follows that an individual can increase the utility of lifetime consumption by transferring wealth from high-income periods to low-income periods. The most effective way of doing this is to save when income is high and to borrow (or run down savings) when income is low.

Merton (1969) describes a model in which individuals can invest in a single risky asset and combine this with an arbitrary level of borrowing or lending at a constant risk-free rate. Although it may seem unduly restrictive to allow only one risky asset, this approach is justified by the separation principle of portfolio theory, which states that the set of efficient portfolios for investors who can borrow or lend at the risk-free rate contains a unique sub-portfolio of risky assets; see, for example, Cuthbertson (1998) for a simple derivation of this principle. The amount of risk-free borrowing or lending in the optimal portfolio depends on the risk tolerance of the investor, but the sub-portfolio of risky assets is the same for all investors. Hence, the single risky asset in Merton’s model can be taken as the optimal sub-portfolio of risky assets. It should be noted, however, that the separation principle is only valid when all the assets in the portfolio are marketable.

In Merton’s model, it is assumed that an individual with some initial wealth invests and consumes this wealth over a fixed lifespan, leaving a bequest for descendants. The more general form of the model allows for future earnings as well as investment gains. The aim of the model is to determine optimal values for both:

- The amount of wealth that is consumed (i.e., spent on goods and services) at any duration; and
• The proportion of wealth invested in the risky asset at any duration.

This is achieved by maximizing the expected value of a function that depends both on the utility of lifetime consumption and the amount of the bequest.

2.1 Mathematical Description of the Model

For ease of explanation, we present the model in a discrete-time framework. The remaining lifespan of the individual is certain and is divided into N sub-intervals, each of duration \( \Delta t \), so that sub-interval \( k + 1 \) is \([k\Delta t, (k + 1)\Delta t)\), for \( k = 0, 1, \ldots, N - 1 \). The following variables are defined:

- \( W_k \) = Total wealth of the individual at the start of sub-interval \( k + 1 \);
- \( S_k \) = Salary payment received at the start of sub-interval \( k + 1 \);
- \( G_k \) = Wealth consumed at the start of sub-interval \( k + 1 \);
- \( X_k \) = Proportion of wealth invested in the risky asset at start of sub-interval \( k + 1 \);
- \( \delta_k \) = Random force of return on risky asset over sub-interval \( k + 1 \);
- \( \rho = \) Constant risk-free force of return over each sub-interval.

The variables \( \delta_k \) and \( \rho \) are small forces of growth measured over the duration \( \Delta t \). The former is a random variable that depends on the stochastic process used to model the return on the risky asset, whereas the latter is given by:

\[
\rho = r \Delta t
\]

where \( r \) is the annual risk-free force of interest.

The total wealth of the individual must change over each sub-interval as follows:

\[
W_{k+1} = (1 - X_k)(W_k + S_k - G_k)e^\rho + X_k(W_k + S_k - G_k)e^{\delta_k}
\]  

(1)

where \( W_k, G_k, S_k \) are non-negative for \( k = 0, 1, \ldots, N - 1 \). Equation (1) is known as the individual’s budget constraint, as future consumption is constrained by initial wealth and future earnings.
At the end of the \( n \)th sub-interval, \( n = 1, 2, \ldots, N - 1 \), it is assumed that the individual wishes to maximize \( E_n[I_n] \) where:

\[
I_n = \sum_{k=n}^{N-1} U(G_k)e^{-k\theta} + B(W_N)
\]

\( E_n = \) Expected value operator given the information available at the end of the \( n \)th sub-interval;

\( U(\cdot) = \) Utility function for consumption at any duration;

\( B(\cdot) = \) Bequest function giving the utility of wealth at death; and

\( \theta = \) A parameter reflecting the subjective time-preference for consumption.\(^2\)

Although the proportion invested in the risky asset, \( X_n \), does not appear explicitly in equation (2), it is clear from the budget constraint that the value chosen for \( X_n \) will affect the distribution of future wealth and hence the expected value of \( I_n \). Thus, we must find the optimal values \( G_n^* \) and \( X_n^* \) of \( G_n \) and \( X_n \), respectively, that maximize \( E_n[I_n] \). These optimal values will depend on the current amount of wealth, future salary payments, and the length of the remaining lifespan. We can represent them as functions of the following form:

\[
G_n^* = G(W_n, S_n, S_{n+1}, \ldots, S_{N-1}, N-n)
\]

\[
X_n^* = X(W_n, S_n, S_{n+1}, \ldots, S_{N-1}, N-n).
\]

If we assume that the individual can revise the consumption and portfolio choices at the start of each remaining time interval, the optimization problem is not straightforward. We must find the values of \( G_n \) and \( X_n \) that maximize the expected value of \( I_n \), given that the individual will apply the same optimizing procedure at the start of each future time interval. Moreover, we cannot predict what the optimal future values of \( G_k \) and \( X_k \) will be (for \( k > n \)), because they will depend on future wealth. As part of the wealth is being invested in a risky asset, the future wealth at any duration will be a random variable; thus, the optimal future values of \( G_k \) and \( X_k \) must also be random.

The problem outlined above is referred to as a multi-period problem in the financial literature, e.g., Mossin (1968), and its solution is based on an algorithm developed by Bellman (1959). This algorithm is applied in Section 3, where results are obtained for the model based on maximizing the expected utility of retirement wealth.

\(^2\)For more on subjective time-preferences, see Appendix A3.
2.2 Closed-Form Solution

Merton approached the problem of maximization of the expected utility of lifetime consumption in continuous time, showing that closed-form solutions for $G^*_n$ and $X^*_n$ exist under the following conditions:

1. The stochastic process for the return on the risky asset is of the form:

$$\delta_k \sim N((\mu - \frac{1}{2}\sigma^2)\Delta t, \sigma^2\Delta t);$$

2. The utility function for consumption $G$ is of the form:

$$U(G) = \frac{(G - G_{\text{min}})^{1-\gamma}}{1-\gamma}, \quad G \geq G_{\text{min}}$$

(3)

where $G_{\text{min}}$ and $\gamma$ are positive constants; and

3. The bequest function is of a similar form to the utility function or zero.

The first condition assumes that investment returns on the risky asset follow a geometric Brownian motion. The second condition requires that the utility function belongs to the hyperbolic absolute risk aversion (HARA) class. The parameter $G_{\text{min}}$ can be thought of as the minimum level of consumption required for subsistence, at which point the risk tolerance of the individual is zero, and $\gamma$ is the limiting value of the individual's relative risk aversion as $G \to \infty$.

If we set $G_{\text{min}} = 0$ we obtain the sub-class of iso-elastic utility functions, for which the solution for $X^*_n$ has a simple form. In the case of an individual with no future earnings (i.e., $S_k = 0$ for $k > n$), it can be shown that:

$$X^*_n = \left(\frac{\mu - \gamma}{\gamma \sigma^2}\right) e^\rho.$$  

(4)

As all the parameters on the right side of equation (4) are constants, the same proportion of accumulated wealth should be invested in the risky asset at all points in the lifespan. This appears to be a refutation of lifestyle investment strategies, (Booth and Yakoubov, 2000) where portfolios are progressively switched into less risky assets as the individual ages, but the result only applies when there are no future earnings.

---

In the continuous time limit given by Merton $e^\rho - 1$. 

If the individual does expect to receive future earnings, the same proportion of the total wealth (i.e., the sum of the accumulated wealth and the present value of future earnings) should be invested in the risky asset. Thus, the proportion of accumulated wealth that should be invested in the risky asset is given by:

\[
X^*_n = \left( \frac{\mu - r}{\gamma \sigma^2} \right) e^{p} \left( \frac{W_n - G^*_n + \sum_{k=n}^{N-1} S_k e^{-(k-n)p}}{W_n - G^*_n + S_n} \right).
\]

Equation (5) indicates that young workers, for whom the capitalized value of future earnings will be relatively large, should invest a higher proportion of their accumulated wealth in risky assets. It is probable that the optimal proportion will exceed one for some young workers, implying that such individuals should borrow money in order to invest in risky assets expected to provide a higher return than the interest rate on their loans. In Section 3, analogous results to those presented above will be derived for a model based on maximizing the expected utility of retirement wealth.

### 2.3 Practical Application of the Model

A powerful feature of the utility of lifetime consumption model is that portfolio and consumption choices are optimized together. In theory, the model could be used to advise individuals on how much to contribute to a retirement fund as well as on their portfolio choices. In order to use the model in this way, however, we would have to estimate various items for the individual, such as the parameter for the subjective rate of time preference, which may be difficult to do in practice. A further problem with the model presented above is its assumption that individuals save only to increase their future consumption or make bequests. The leisure-motive is ignored.

A fuller discussion of these problems is given in the appendix, which concludes that it would be difficult to provide advice to individuals using a model based on maximizing the expected utility of lifetime consumption. For this reason, we derive the main results of this article using the model based on maximizing the expected utility of retirement wealth.
3 Utility of the Retirement Wealth Model

We now consider the model based on maximizing the expected utility of retirement wealth. Unlike the previous model, the income saved during future periods is assumed to be predetermined, so the only decision left for the individual is how to adjust the investment portfolio over the period up to retirement.

The main advantage of this simpler model is that we no longer need to allow for the subjective rate of time-preference, as we are only interested in the utility of the projected wealth at a fixed point in time. The disadvantage of this approach is that we cannot allow for adjustments to the rate of saving that may be desired in light of realized investment returns.

There are two plausible justifications for ignoring variations in the future rate of saving. First, if we are applying the model to a defined-contribution pension plan, the scope for varying the future contribution rate may be limited. Second, the individual's own retirement planning is likely to be based on some assumed rate of saving until a targeted retirement age, so a utility-maximization exercise based on this plan is likely to be of practical help. It follows that the question we are seeking to answer for any individual is:

The Question: Given a particular rate of saving and a particular age of retirement, what is my optimal investment policy?

A drawback of the utility of lifetime consumption model is its failure to allow for the leisure motive. Is a model based on maximizing the utility of retirement wealth any better in this regard? The answer is that the leisure motive is implicitly a part of this model because the individual can choose his/her retirement age, which may be below the normal retirement age of his/her occupation. This is clearly an imperfect method of allowing for the leisure motive, as there is no attempt to optimize the retirement age in light of the actual circumstances of the individual at future ages. Given the near impossibility of anticipating what the individual preference for leisure over work will be at any future age, it may be the only practical approach.

3.1 Mathematical Description of the Model

In this model it is the remaining period until retirement that is divided into $N$ sub-intervals, each of duration $\Delta t$. The required variables

\[ N \quad \text{(number of sub-intervals)} \]
\[ \Delta t \quad \text{(duration of each sub-interval)} \]

4In the U.K., for example, most employer-sponsored DC plans do not allow employees to take extra salary in lieu of pension benefits.
are as defined in Section 2.1, except that the consumption and salary cash flows, $S_k$ and $G_k$, are replaced with a single cash flow equal to the contribution made to the retirement fund. Hence, we define $C_k$ as the contribution to retirement fund at the start of sub-interval $k + 1$. The budget-constraint equation is now given by:

$$W_{k+1} = (1 - X_k)(W_k + C_k)e^\rho + X_k(W_k + C_k)e^{\delta_k}$$

(6)

where $W_k$ and $C_k$ are non-negative for all possible values of $k$.

The aim of the model is to find the value of $X_n$ that maximizes:

$$\mathbb{E}_n[U(W_N)]$$

where $U(\cdot)$ is the utility function for retirement wealth. This is again a multi-period problem, as we must allow for further utility-maximizing adjustments to the value of $X_k$ over the period up to retirement (for $k > n$). It is useful to begin by obtaining a solution for the single period case, however, as this can later be applied to the multi-period problem.

### 3.2 Single-Period Problem

We now obtain the optimal portfolio for an individual investing a lump sum over a single small time interval of duration $\Delta t$. We assume that the return on the risky asset follows a geometric Brownian motion, thus:

$$\delta_k \sim N((\mu - \frac{1}{2}\sigma^2)\Delta t, \sigma^2\Delta t).$$

For a lump sum investment, there is no contribution to the retirement fund. The budget constraint becomes:

$$W_{k+1} = (1 - X_k)W_k e^\rho + X_k W_k e^{\delta_k}$$

which can be re-written as:

$$W_{k+1} = W_k e^\rho + X_k W_k (e^{\delta_k} - e^\rho).$$

The $(e^{\delta_k} - e^\rho)$ term is the risk premium on the risky asset, which is a small number over the small duration $\Delta t$.

The utility of $W_{k+1}$ can approximated by a Taylor expansion about $W_k e^\rho$:
\[ U(W_{k+1}) = U(W_k e^\rho) + X_k W_k (e^{\delta_k} - e^\rho) U'(W_k e^\rho) \]
\[ + \frac{1}{2} X_k^2 W_k^2 (e^{\delta_k} - e^\rho)^2 U''(W_k e^\rho) + \ldots. \]  

(7)

Ignoring terms of the order of \((\Delta t)^2\), we obtain:

\[ e^\rho = e^{r \Delta t} \approx 1 + r \Delta t \]

\[ \mathbb{E}_k [e^{\delta_k}] = \exp((\mu - \frac{1}{2} \sigma^2) \Delta t + \frac{1}{2} \sigma^2 \Delta t) \approx 1 + \mu \Delta t \]

\[ \mathbb{E}_k [e^{2\delta_k}] = \exp(2(\mu - \frac{1}{2} \sigma^2) \Delta t + \frac{4}{2} \sigma^2 \Delta t) \approx 1 + (2 \mu + \sigma^2) \Delta t. \]

We now apply the \(\mathbb{E}_k[\cdot]\) operator to both sides of equation (7), inserting the relationships given above into the right side. Ignoring terms of the order of \((\Delta t)^2\), we obtain:

\[ \mathbb{E}_k[U(W_{k+1})] \approx U(W_k e^\rho) + X_k W_k (\mu - r) U'(W_k e^\rho) \Delta t \]
\[ + \frac{1}{2} X_k^2 W_k^2 \sigma^2 U''(W_k e^\rho) \Delta t. \]  

(8)

The right side of equation (8) is quadratic in \(X_k\) and has a global maximum provided that:

\[ U''(W_k e^\rho) < 0. \]

The above inequality holds for risk-averse investors.

To find the value of \(X_k\) that maximizes the expected utility of wealth, we take the partial derivative of equation (8) with respect to \(X_k\) and set it equal to zero. The optimal proportion invested in the risky asset is then given by:

\[ X_k^* = \left( \frac{\mu - r}{\sigma^2} \right) \left( \frac{U'(W_k e^\rho)}{W_k U''(W_k e^\rho)} \right). \]  

(9)

The continuous time limit of equation (9) is given by setting \(e^\rho = 1\) and is called the Merton ratio by Panjer et al., (1998). For an iso-elastic utility function, it is easy to show that equation (9) is identical to Merton's closed-form solution for the utility of lifetime consumption model as given in equation (4).

Equation (9) gives us a useful way of interpreting two properties of utility functions known as absolute risk aversion and relative risk aversion, which are defined as follows:
Absolute risk aversion = \(-\frac{U''(W)}{U'(W)}\)

Relative risk aversion = \(-\frac{WU''(W)}{U'(W)}\).

If we apply these definitions in the continuous time limit of equation (9) (when \(e^\rho = 1\)), we obtain:

\[ X_k^* W_k = \left( \frac{\mu - r}{\sigma^2} \right) / \text{Absolute risk aversion} \]
\[ X_k^* = \left( \frac{\mu - r}{\sigma^2} \right) / \text{Relative risk aversion}. \]

It follows that an investor with a utility function exhibiting constant absolute risk aversion would be expected to invest the same amount of wealth in the risky asset, whereas an investor with a utility function exhibiting constant relative risk aversion would be expected to invest the same fraction of wealth in the risky asset.

The above results have been derived for a lump-sum investment made over a single time-period. It remains to be seen whether similar results can be derived for the multi-period case, with and without future contributions.

3.3 Multi-Period Problem for a Lump Sum Investment

We now consider the multi-period problem for a lump sum investment. Equation (8) can be applied to the time interval before retirement as follows:

\[ E_{N-1}[U(W_N)] = U(W_{N-1}e^\rho) + X_{N-1}W_{N-1}(\mu - r)\Delta t U'(W_{N-1}e^\rho) \]
\[ + \frac{1}{2} X_{N-1}^2 W_{N-1}^2 \sigma^2 \Delta t U''(W_{N-1}e^\rho). \]

From equation (9), we can deduce that the optimal value of \(X_{N-1}\) is given by:

\[ X_{N-1}^* = \left( \frac{\mu - r}{\sigma^2} \right) \left( \frac{-U'(W_{N-1}e^\rho)}{W_{N-1}U''(W_{N-1}e^\rho)} \right). \]

If we substitute the optimal value of \(X_{N-1}\) into the Taylor expansion for \(E_{N-1}[U(W_N)]\), we obtain the following expression for the maximum value of \(E_{N-1}[U(W_N)]\):
3.3.1 Restricting the Choice of Utility Function

We now observe that the multi-period problem is greatly simplified for utility functions satisfying the following relationship:

\[ \frac{[U'(W)]^2}{U''(W)} = \lambda U(W) \]  

where \( \lambda \) is a constant. If the above relationship holds, equation (10) reduces to:

\[ \mathbb{E}_{N-1}^*[U(W_N)] = \tilde{\lambda} U(W_{N-1}e^\rho) \]  

where \( \tilde{\lambda} \) is another constant.

Thus, the maximum expected utility of retirement wealth at the start of the final time interval has a simple form; it is proportional to the utility of the retirement wealth that would be obtained by investing in the risk-free asset. But this is only true for utility functions satisfying the relationship given above in equation (11). It is not difficult to show that the HARA class of utility functions, referred to in equation (3), meet this requirement.

3.3.2 Moving Back One Period

If we now consider the optimal portfolio choice at the start of the penultimate time interval, the law of iterated expectations allows us to express the maximum expected utility of the retirement wealth as:

\[ \mathbb{E}_{N-2}^*[U(W_N)] = \mathbb{E}_{N-2}^*[\mathbb{E}_{N-1}^*[U(W_N)]] \]

For HARA utility functions we can use equation (12) to substitute for \( \mathbb{E}_{N-1}^*[U(W_N)] \), which gives:

\[ \mathbb{E}_{N-2}^*[U(W_N)] = \tilde{\lambda}\mathbb{E}_{N-2}^*[U(W_{N-1}e^\rho)] \]

Thus, the optimal portfolio choice at the start of the penultimate time interval is obtained by finding the value of \( X_{N-2} \) that maximizes the value of \( \mathbb{E}_{N-2}^*[U(W_{N-1}e^\rho)] \).

On multiplying through the budget constraint equation for the penultimate sub-interval by \( e^\rho \), we obtain the following formula:
\[ W_{N-1} e^\rho = W_{N-2} e^{2\rho} + X_{N-2} W_{N-2} e^\rho (e^{\delta_{N-2}} - e^\rho). \]

This expression leads to a Taylor expansion for \( U(W_{N-1} e^\rho) \). Neglecting terms of higher order than second gives:

\[
U(W_{N-1} e^\rho) = U(W_{N-2} e^{2\rho}) + X_{N-2} W_{N-2} e^\rho (e^{\delta_{N-2}} - e^\rho) U'(W_{N-2} e^{2\rho}) + \frac{1}{2} X_{N-2}^2 W_{N-2}^2 e^{2\rho} (e^{\delta_{N-2}} - e^\rho)^2 U''(W_{N-2} e^{2\rho}).
\]

We now follow the same steps presented in Section 3.2 to obtain the following expressions for \( \mathbb{E}_{N-2}[U(W_{N-1} e^\rho)] \) and \( X_{N-2}^* \):

\[
\mathbb{E}_{N-2}[U(W_{N-1} e^\rho)] = U(W_{N-2} e^{2\rho}) + X_{N-2} W_{N-2} e^\rho (\mu - \sigma) \Delta t U'(W_{N-2} e^{2\rho}) + \frac{1}{2} X_{N-2}^2 W_{N-2}^2 e^{2\rho} \sigma^2 \Delta t U''(W_{N-2} e^{2\rho})
\]

\[
X_{N-2}^* = \left( \frac{\mu - \sigma}{\sigma^2} \right) \left( \frac{-U''(W_{N-2} e^{2\rho})}{W_{N-2} U''(W_{N-2} e^{2\rho})} \right) e^{-\rho}.
\]

On comparing the expressions for \( X_{N-2}^* \) and \( X_{N-1}^* \), we see that although the derivatives of the utility function have different arguments, both are equal to the current wealth multiplied by the risk-free return compounded up to retirement. The only other difference is that the expression for \( X_{N-2}^* \) is discounted by the risk-free interest rate for a single period.

### 3.3.3 The General Solution

It is not difficult to see the pattern that will emerge if we continue to move backwards in time, period by period. As long as we are using a utility function of the HARA class, an expression of the following form will apply at the end of the \( n \)th sub-interval:

\[
\mathbb{E}_{n}^*[U(W_N)] = \bar{\lambda} \mathbb{E}_{n+1}^*[U(W_{n+1} e^{\rho(N-n-1)})].
\]

The Taylor expansion for \( U(W_{n+1} e^{\rho(N-n-1)}) \) is derived using the budget constraint as follows:

\[
W_{n+1} e^{\rho(N-n-1)} = W_n e^{\rho(N-n)} + X_n W_n e^{\rho(N-n-1)} (e^{\delta_n} - e^\rho).
\]

And the optimal proportion invested in the risky asset will be:
\[ X^*_n = \left( \frac{\mu - r}{\sigma^2} \right) \left( -U'(W_n e^{\rho(N-n)}) \right) e^{-\rho(N-n-1)}. \] (13)

On substituting the generic HARA utility function \( U(W) \), i.e.,

\[ U(W) = \frac{(W - A)^{1-\gamma}}{(1 - \gamma)} \quad W \geq A \] (14)

where \( A \) can be interpreted as the minimum retirement wealth required for subsistence, we obtain:

\[ X^*_n = \left( \frac{\mu - r}{\sigma^2} \right) \left( W_n - Ae^{-\rho(N-n)} \right) e^\rho. \] (15)

Equation (15) indicates that the amount of wealth that should be invested in the risky asset is proportional to excess of the accumulated wealth over that amount that can guarantee the subsistence wealth at retirement. Thus, the individual should follow a strategy in which the subsistence wealth is guaranteed by investing a proportion of the fund at the risk-free rate and the remainder of the fund is split between the risky and risk-free asset, according to the Merton ratio. Samuelson (1989) observes that this result implies that the proportion invested in the risky asset will decline nearing retirement if the accumulated wealth is fixed over time. The accumulated wealth is likely to increase over time, often at a faster rate of growth than the risk-free rate, however, so the above result is not really an argument for lifestyle strategies.

If we set \( A = 0 \) we get the optimal proportion for an iso-elastic utility function, which is identical to Merton's result for the utility of lifetime consumption model, as given in Section 2.2. The single-period solution of equation (9) also gives this result for an iso-elastic utility function, indicating that the short-term and long-term problems have the same solution for this type of utility function.

### 3.4 Multi-Period Problem for a Lump Sum and Future Contributions

The multi-period solution of equation (15) does not provide a strong case for investing in a less risky portfolio on nearing retirement, and the solution for an iso-elastic utility function supports a policy of investing the same fraction of wealth in risky assets at all durations from retirement. The problem we have considered, however, is not a realistic one for most members of defined-contribution pension plans, as no allowance has been made for future contributions. We shall show
that there is a strong case for lifestyle strategies when the individual expects to make further contributions to the retirement fund. Equation (6) gives the general form of the budget constraint, which in the time interval before retirement can be written as:

$$W_N = (W_{N-1} + C_{N-1})e^\rho + X_{N-1}(W_{N-1} + C_{N-1})e^{\delta N-1} - e^\rho.$$ 

If we now obtain a Taylor expansion for $U(W_N)$ and follow the same steps as given in Section 3.2, the only change in the expression for optimal equity proportion at the start of the final time interval is that $W_{N-I}$ is replaced by $(W_{N-1} + C_{N-1})$, hence:

$$x^*_{N-1} = \left(\frac{\mu - r}{\sigma^2}\right) \left(\frac{-U'(W_{N-1} + C_{N-1})e^\rho}{U''(W_{N-1} + C_{N-1})e^\rho}\right) \left(\frac{1}{W_{N-1} + C_{N-1}}\right).$$

For a utility function of the HARA class, the expression for the maximum value of the expected retirement wealth at the start of the final time interval becomes:

$$E_{N-1}[U(W_N)] = \tilde{\lambda} U((W_{N-1} + C_{N-1})e^\rho).$$

To obtain a Taylor expansion for $U((W_{N-1} + C_{N-1})e^\rho)$ the budget constraint for the penultimate time interval needs to be expressed in the following form:

$$(W_{N-1} + C_{N-1})e^\rho = (W_{N-2}e^{2\rho} + C_{N-2}e^{2\rho} + C_{N-1}e^\rho) + X_{N-2}(W_{N-2} + C_{N-2})e^\rho(e^{\delta N-1} - e^\rho).$$

Hence, the Taylor expansion for the penultimate time interval will be taken about the first term in brackets on the right side of the above equation. This term is equal to the retirement wealth that could be secured by investing wholly in the risk-free asset, allowing for future contributions as well as the current fund. This leads to the following expression for the optimal portfolio choice at the start of the penultimate sub-interval:

$$x^*_{N-2} = \left(\frac{\mu - r}{\sigma^2}\right) \left(\frac{-U'(W_{N-2}e^{2\rho} + C_{N-2}e^{2\rho} + C_{N-1}e^\rho)}{U''(W_{N-2}e^{2\rho} + C_{N-2}e^{2\rho} + C_{N-1}e^\rho)}\right) \left(\frac{e^{-\rho}}{W_{N-2} + C_{N-2}}\right).$$
3.4.1 The General Solution

The pattern emerging is now clear: for any earlier time interval, the derivatives of the utility function will have an argument equal to the projected retirement wealth using the risk-free interest rate. For each period moved backward, we must discount the expression by the risk-free interest rate for a single period. It follows that the general expression for the optimal portfolio choice is:

\[ X_n^* = \left( \frac{\mu - r}{\sigma^2} \right) \left( -U'(W_n e^{\rho(N-n)} + \sum_{k=n}^{N-1} C_k e^{\rho(N-k)}) \right) \frac{e^{-(N-n-1)}}{W_n + C_n} \]  

(16)

On substituting the generic form for HARA utility functions, as given in equation (14), we obtain:

\[ X_n^* = \left( \frac{\mu - r}{\gamma \sigma^2} \right) \left( W_n + \sum_{k=n}^{N-1} C_k e^{-\rho(k-n)} - A e^{-\rho(N-n)} \right) W_n + C_n \]  

(17)

The amount invested in the risky asset is proportional to the excess of the total wealth over the amount required to guarantee subsistence at retirement, where total wealth includes both the accumulated fund and the present value of future contributions. Allowing for the capitalized value of future contributions in this way is analogous to allowing for the capitalized value of future earnings in the utility of consumption model, as described in Section (2).

Equation (17) shows that when future contributions are expected the case for a lifestyle strategy is strong: workers should invest a higher proportion of their accumulated fund in risky assets when the capitalized value of their future contributions is high, i.e., when they are young.

4 Graphical Presentation of Results

We now demonstrate how the solution for the optimal portfolio choice, as given by equation (17), can be presented graphically. We start by making the assumption that contributions to the fund occur at a uniform rate, so that equation (17) becomes:
where $C$ is the fixed contribution at the start of each time interval.

For the purpose of our graphical presentation we shall use the continuous time limit of equation (18), in which the variables will be reexpressed in terms of the duration from retirement, $T$, and a continuous rate of contribution, $\tilde{C}$. If we allow the length of each time interval, $\Delta t$, to tend to zero so that $(N - n)\Delta t = T$ and $\tilde{C} = \frac{\tilde{C}}{T}$ we find that equation (18) converges to:

$$X_n^* = \left(\frac{\mu - r}{y\sigma^2}\right) \left(\frac{W_n + C\tilde{a}_{N-n} \rho - Ae^{-\rho(N-n)}}{W_n + C} \right) e^\rho$$

(18)

The age-dependent variables on the right side of equation (19) are $(W_T, T)$. If we assume that the other parameters are constants for any one individual, $X_T^*$ is effectively a function of these two variables.

### 4.1 Portfolio Isoquants

Equation (19) can be represented graphically by plotting curves in the $(W_T, T)$ plane for which the optimal proportion $X_T^*$ is a constant. Each of these curves will be referred to as an isoquant.

Let $T_0$ be the unique solution to the equation

$$\tilde{C} = \frac{\tilde{C}}{T} - Ae^{-rT_0} = 0.$$  

On solving for $T_0$, we get

$$T_0 = \frac{1}{r} \ln \left(1 + \frac{Ar}{C}\right).$$

(20)

At this duration $X_T^*$ is independent of the accumulated wealth $W_T$. The portfolio isoquant at duration $T_0$ is a vertical line in the $(W_T, T)$ plane, and the optimal proportion invested in the risky asset at this duration is given by:

$$X_{T_0}^* = \left(\frac{\mu - r}{y\sigma^2}\right).$$

Thus, the duration $T_0$ is the one at which the individual will always invest the same proportion of accumulated wealth in the risky asset, this proportion being equivalent to the continuous time limit of the
Merton ratio given in Section 2.2. We shall show that the other isoquants are curves that intersect at the point: $W_T = 0, T = T_0$.\(^5\)

### 4.2 Choice of Parameter Values

We shall work in inflation-adjusted currency units, so that the uniform contribution rate, $\tilde{C}$, is a contribution that rises in line with inflation. The subsistence retirement wealth, $A$, is expressed in terms of today's dollars. Contributions rising with inflation are more representative of a typical retirement plan than fixed nominal contributions, and it is easier and more natural to estimate the retirement wealth required for subsistence in terms of current dollars.

The investment-related parameters are:

- The expected real return on the risky asset (which equals $e^\mu - 1$);
- The standard deviation of the real force of return on the risky asset (which equals $\sigma$); and
- The real risk-free return (which equals $r^r - 1$).

For the purpose of our illustration we shall take the risky asset as a representative portfolio of U.S. equities and the risk-free asset as U.S. Treasury bills. Annual data for the gross returns on each of these assets, deflated by the Consumer Price Index, are given in the *Barclays Capital Equity-Gilt Study 2001*.\(^6\) The following parameter estimates were obtained from these data over the 40 consecutive calendar years from 1961 to 2000:

$$
\mu = 0.068, \quad \sigma = 0.17, \quad r = 0.015.
$$

The other parameters in equation (19) are specific to the individual. They are:

- $A$, the subsistence retirement wealth;
- $\tilde{C}$, the annual rate of contribution; and
- $\gamma$, the limiting value of the individual's relative risk aversion.

---

\(^5\)As there is no wealth to invest at $W_T = 0$, it does not matter that the isoquants intersect there.

\(^6\)Barclays Capital is a U.K. investment bank. The source for its U.S. equity returns is an index of historic stock prices supplied by the University of Chicago Graduate School of Business.
The parameter $A$ can be removed by setting $A = 1$, in which case both the accumulated wealth and the annual contribution are measured relative to the subsistence retirement wealth. Reasonable values for $\hat{C}$ may lie in the range 0.025 to 0.05, so that the individual is saving at a rate that reasonably could assure the required subsistence wealth over a typical working life of 40 years. Panjer et al., (1998) quote Constantinides (1990) in which a value of 2 is recommended for the relative risk aversion of a typical investor, whereas Kapur and Orszag (1998) assume a value of 1.25 for an iso-elastic utility function. As our parameter $y$ gives the lower limit of the individual's relative risk aversion (as wealth tends to infinity), a reasonable range of values might be from 1.0 to 1.5.

4.3 Comments on Figures 1-4

Figures 1-4 were obtained by solving equation (19) for the fixed values of $X_T$ corresponding to each portfolio isoquant. As mentioned in Section 4.1, the isoquants meet at a fixed point on the horizontal axis, $W_T = 0$. The duration from retirement at this point, $T_0$, is as given by equation (20).

Each graph corresponds to a particular rate of contribution, $\hat{C}$, and a particular risk aversion parameter, $y$. It follows that any graph would have to be tailored to the circumstances of a particular individual. At any time, we can plot the position of an individual on the graph, as defined by the duration from retirement ($T$) and the market value of the accumulated fund ($W_T$). If this point lies between two isoquants, the optimal equity proportion lies between the proportions corresponding to each isoquant. The precise value of this optimal proportion is given by equation (19).

As the duration from retirement reduces, we would expect the accumulated wealth of most individuals to increase. Such individuals will map a line on each graph that slopes upwards from the right. The desirability of investing a greater proportion of wealth in the risk-free asset on nearing retirement is immediately apparent from the graphs. As the duration from retirement reduces, the individual passes through isoquants for which the optimal equity proportion gets smaller and smaller. These graphs suggest that the case for lifestyle investment strategies is a powerful one.
Figure 1
Portfolio Isoquants ($C = 0.025$ and $y = 1.5$)

Figure 2
Portfolio Isoquants ($C = 0.025$ and $y = 1.0$)
Figure 3
Portfolio Isoquants ($C = 0.05$ and $y = 1.5$)

Figure 4
Portfolio Isoquants ($C = 0.05$ and $y = 1.0$)
The isoquants at the furthest durations from retirement are for optimal equity proportions greater than one and have positive gradients. An optimal proportion greater than one implies that the individual should borrow at the risk-free rate to increase his/her exposure to equities; the positive gradients imply that the exposure to equities should be reduced as the accumulated wealth increases.

When the duration from retirement falls below \( T_0 \) the isoquants have negative gradients, which implies that the optimal equity proportion increases with wealth at any fixed duration from retirement. The final isoquant is always for \( X_T = 0 \). At this isoquant, the fund should be invested entirely in risk-free assets, because the projected retirement wealth is only just sufficient to guarantee subsistence. The model is indeterminate in the section of the graph below this isoquant, as the individual has passed beyond the point at which his/her relative risk aversion is infinite.

5 Summary and Conclusions

The two fundamental questions for individuals who are accumulating savings over their working lives are:

- How much should I save at any given time?
- Where should my accumulated savings be invested?

Models based on maximizing the expected utility of lifetime consumption theoretically can deal with both questions simultaneously. Such models, however, require individuals to supply comprehensive data on their future preferences for consumption and leisure. It seems unlikely that anyone would be able to provide such information. The evidence suggests that people who engage in long-term financial planning do so with the aim of accumulating sufficient wealth to provide for their future needs at a chosen target retirement age; see Uccello (2001).

Models based on maximizing the expected utility of retirement wealth do not require as much information about future preferences and are more in tune with the kind of long-term financial plans that people actually make. They are therefore more likely to be of practical value, even though we can only use them to optimize portfolio choices (and not saving decisions). Such a model is used to derive a formula for the optimal proportion of accumulated wealth that should be invested in equities. We show that individuals who expect to pay future contributions to their retirement fund generally should reduce the equity
content of their fund over time. Hence, lifestyle investment strategies for defined-contribution pension plans appear to be justified.

The information provided by this model can be presented graphically in the form of portfolio isoquants. Each graph consists of a series of curves mapping points in the plane of accumulated wealth against duration from retirement, and each curve consists of those points at which the optimal equity proportion is a constant. These graphs show that individuals who are far from retirement (i.e., those close to the start of their working lives) should borrow money to increase their equity proportion above one. We also find that the optimal equity proportion reduces with wealth at long durations from retirement and increases with wealth at short durations from retirement.

There are limitations to our model, however. The risky asset returns are assumed to follow a geometric Brownian motion. As these returns are independent, the variance of the projected fund increases more quickly than in alternative stochastic models that incorporate some element of mean reversion. Thus, the model presented in this article might tend to understate the long-term case for equity investment. On the other hand, a geometric Brownian motion ignores the possibility of sudden changes in equity prices (e.g., the equity market crash of 1987), which tends to understate the short-term risks of equity investment. Given the parameter uncertainty inherent in any model, it is not clear whether much would be gained by using a more complex stochastic model. The possibility of an equity market crash should certainly be kept in mind, however, when interpreting the results of the model at durations close to retirement.

Throughout this article we have assumed that the problem of portfolio choice can be reduced to the subdivision of an accumulated fund between a risky and a risk-free asset. This simplification depends on the separation principle of portfolio theory, which states that the optimal portfolio of any individual who can borrow or lend at the risk-free rate contains a unique sub-portfolio of risky assets. The separation principle assumes that all the available risky assets are marketable, which is not the case for most individuals: significant non-marketable assets might include domestic property and defined-benefit pension assets (e.g., from a social security scheme). An important area of further work, therefore, would be to examine the effect of illiquid assets on portfolio choices in defined-contribution pension plans.
References


Appendix: Review of the Utility of the Lifetime Consumption Model

Applying the utility of consumption model as a normative tool requires a method for obtaining each of the following items for the individual we are seeking to advise:

- The utility function, $U(\cdot)$;
- The bequest function, $B(\cdot)$; and
- The subjective rate of time-preference, $\theta$.

A1: Utility Function

A method for obtaining the utility function of any individual is described by Bowers et al., (1997). Essentially, this involves asking the individual what minimum amount of consumption he/she would accept with certainty in preference to a lottery where the amount of consumption will be either of two values with equal probability. By asking this question for lotteries offering different levels of consumption, the utility function can be constructed piecewise. Alternatively, this approach could be used to determine the subjective parameter values of a standard type of utility function (e.g., the HARA class mentioned above).

The model assumes that the utility function will remain unchanged throughout the lifetime of the individual. While it is unlikely that such an assumption is generally correct, it may not be too far from the truth if consumption is measured in inflation-adjusted dollars so that one unit of future consumption will purchase the same basket of goods now and in the future.\(^7\)

A2: Bequest Function

The bequest function is also subjective—it represents the utility that the individual attaches to wealth inherited by next of kin (or other beneficiaries of the estate). To derive the bequest function, we need to determine how much consumption the individual would be prepared to sacrifice for a given increase in the bequest. A way of approaching

---

\(^7\)This makes no allowance for the fact the range of goods available for consumption in the future may differ from that available today, which would make any quantitative comparison of intertemporal utilities difficult. If technology continues to improve the quality and range of goods, one might expect the marginal utility of each inflation-adjusted dollar to increase over time.
this problem is to ask the individual to imagine a scenario in which total lifetime wealth is fixed, all saving or borrowing is at the risk-free rate, and any uniform rate of consumption consistent with a non-negative bequest may be chosen.

It follows that the function we are seeking to maximize can be written as:

$$I_0(G) = U(G)\tilde{a}_{\bar{N}}|\theta + B(W_N)$$

and the lifetime budget constraint is given by:

$$G\tilde{a}_{\bar{N}}|\theta + W_N e^{-\rho N} = TW_0$$

where $TW_0$ is the present value of total lifetime wealth (assumed to be fixed).

For any fixed value of $TW_0$ we can evaluate possible combination of $G$ and $W_N$ and ask the individual to select the preferred combination $(G^*, W_N^*)$. Eliminating $W_N$ between the previous two equations gives:

$$I_0(G) = U(G)\tilde{a}_{\bar{N}}|\theta - B(TW_0 e^{\rho N} - G\tilde{s}_{\bar{N}}|\rho).$$

It can be inferred that $I_0(G)$ has its maximum value for the preferred consumption, $G^*$, so that:

$$I'_0(G^*) = U'(G^*)\tilde{a}_{\bar{N}}|\theta - \tilde{s}_{\bar{N}}|\rho B'(TW_0 e^{\rho N} - G^*\tilde{s}_{\bar{N}}|\rho) = 0.$$ 

By asking the individual to choose preferred combinations of $G$ and $W_N$ for different values of $TW_0$, the above equation can be used to derive a suitable bequest function, assuming the utility function is already known.

A3: Subjective Rate of Time Preference

Last, we require a method of deriving $\theta$, the subjective rate of time preference. This discount rate is intended to allow for the fact that individuals generally prefer to have goods now rather than goods later. As a result, they only will postpone buying extra goods if they later can buy more goods from the money they have saved. Economists have used this concept to explain the phenomenon of interest; see, e.g., von Mises (1949).

To arrive at a method of estimating this parameter, we again assume that the individual has a fixed amount of lifetime wealth that can be reallocated over time by borrowing or lending at the risk-free rate of interest. We further assume that the individual already has assigned a
portion of this wealth for the bequest, so that the only remaining decision is how to spend the wealth available for consumption. It follows that the function we are seeking to maximize simplifies to:

$$I_0 = \sum_{k=0}^{N-1} U(G_k) \exp(-k\theta)$$

subject to the budget constraint:

$$\sum_{k=0}^{N-1} G_k \exp(-k\rho) = GW_0$$

where $GW_0$ is the present value of the wealth available for consumption.

The preferred values of $G_k$ for any fixed value of $GW_0$ must maximize $I_0$ subject to the budget constraint. Using the method of Lagrange multipliers to maximize $I_0$ gives:

$$U'(G_k^*) e^{-k\rho} = \lambda e^{-k\rho},$$

where $\lambda$ is the Lagrange multiplier. When $k = 0$ this becomes:

$$U'(G_0^*) = \lambda.$$  

Eliminating the parameter $\lambda$ between these two equations gives:

$$e^\theta = e^\rho \left( \frac{U'(G_k^*)}{U'(G_0^*)} \right)^{1/k}.$$ 

If the individual prefers consumption to be uniformly distributed over time, then $\theta = \rho$. For any given utility function we could derive some other pattern of consumption that would give the same value of $\theta$ for all values of $k$. The above equation suggests that the value of $\theta$ will not generally be independent of duration, however, which is contrary to the assumption of the model. It seems probable that many individuals will have a term-dependent discount rate because their preferred distribution of consumption involves patterns of spending that will vary over their remaining lifespan.

Another problem concerning the estimation of $\theta$ is the assumption that a single rate of discount can be applied to the utility of total consumption in any time interval. Strictly, we can only infer the discount rate for the marginal utility of consumption at different durations. This

---

8For an iso-elastic utility function the preferred amount of consumption would have to change over time at a fixed compound rate.
point is made clear by assuming a more general form for the utility of consumption over each time interval, so that:

$$I_0 = \sum_{k=0}^{N-1} U_k(G_k).$$

In the existing model we have $$U_k(G_k) = U(G_k)e^{-k\theta},$$ but suppose we instead had assumed:

$$U_k(G_k) = \int_{G_{min}}^{G_k} U'(z) \exp(-k\theta(z))dz$$

where $$G_{min}$$ is the minimum level of consumption required for subsistence and the discount rate $$\theta$$ now depends on the level of consumption. If we again apply the method of Lagrange multipliers to maximize $$I_0$$ for our new utility function we obtain:

$$\frac{\partial U_k}{\partial G_k} \bigg|_{\theta(G_k^*)} = U'(G_k^*) \exp(-k\theta(G_k^*)) = \lambda e^{-k\rho}.$$

This is the same expression as before, with $$\theta(G_k^*)$$ replacing $$\theta$$. Thus, if we derive a value of $$\theta$$ from the preferred distribution of consumption for any given total wealth, what we obtain is a discount rate for the marginal utility of consumption. If we change the total wealth available for consumption and ask the individual to select new values of $$G_k^*$$, we cannot be certain that we will obtain the same discount rate for any given duration. This will only be so if the discount rate is independent of $$G_k$$, as assumed by the model.

Is it necessary to assume that the discount rate is a function of consumption? Consider the purchase of a durable good with a useful life of $$T$$ periods. By delaying the purchase of this good for one period, I sacrifice the use of the good in period 1 for the use of the good in period $$T+1$$. If I prefer to use this good sooner rather than later, my subjective rate of discount will be an increasing function of $$T$$. Now as durable goods have a wide range of useful lives, we can infer that different discount rates will apply to different goods. This suggests that the subjective rate of time preference might vary with the level of consumption in a complex manner that depends on the ordering of preferences for different goods at different times.

---

9This reasoning is consistent with the observation that people assume loans to purchase goods with long useful lives, such as motor vehicles, but are less inclined to borrow money for short-term expenditures.
The above considerations suggest that use of a constant subjective rate of time preference is an oversimplification that probably only can be justified for individuals who are prepared to accept that their planned future consumption always should be uniformly distributed over time. Allowing for more complex patterns of consumption results in a term-dependent discount rate that probably also varies with the amount of consumption at any duration. As well as complicating the solution of the model, a serious difficulty would arise in attempting to deduce this subjective discount function for any individual: it seems highly unlikely that people are sufficiently knowledgeable about their own preferences to give reliable answers to the many hypothetical questions that would be necessary.

A4: The Disutility of Work

The model we are considering assumes that the motive for saving is either to increase future consumption or to provide a bequest. This ignores the disutility of work: an important reason for saving might be to reduce the amount of future work required to obtain a desired level of consumption combined with a desired amount of bequest. The disutility of work (or leisure motive) is important enough to be recognized in economic textbooks as a critical component in any model of the labor market; see, for example, Begg, Fischer, and Dornbusch (2000) pp. 183–186.

The practical significance of the leisure motive also is illustrated in the service tables used by pension actuaries, where the sum of the decrements for voluntary early retirement is typically greater than the decrement at the normal retirement age.

Attempts have made to incorporate the leisure motive into models of consumer choice involving utility functions. For example, Debreu (1959) envisages a utility function for the entire consumption plan of an individual. This plan consists of the number of goods of a specific type, bought (or sold) at a specific time and location, throughout the lifespan of the individual. Goods bought are treated as positive numbers (inputs), and goods sold are treated as negative numbers (outputs). As the most important type of output for most individuals will be the sale of their labor, this generalized utility function does implicitly allow for the disutility of work. For the purpose of the model under consideration, we might replace the utility function for consumption with a utility function of the form:

$$U_k(G_k, -H_k).$$
where $H_k$ is the number of hours worked in the $k^{\text{th}}$ time interval.

In this revised model, $H_k$ would be a third variable to be optimized, along with $G_k$ and $X_k$. Moreover, the future salary of the individual, $S_k$, also would be a random variable equal to $H_k$ multiplied by the projected hourly rate of pay. This would make the model more difficult to solve, but a more immediate question is whether a utility function of the form shown above could be derived for any individual.

The first observation to make about the suggested consumption leisure utility function is that we cannot realistically expect it to remain the same over the lifespan of the individual. The disutility of work increases with age because working becomes more onerous. At some age most people become incapable of work irrespective of their personal preference for leisure. Thus, we must specify a function of the form $U_k(G_k, -H_k)$ that changes over the lifespan of the individual in some manner to be determined. The derivation of this function for any individual would have to allow for the following facts:

- The disutility of work is affected by factors such as state of health, job satisfaction, and the opportunity for meaningful activities outside work; and

- The hourly rate of pay has a critical impact on the consumption/leisure trade-off, as a higher rate of pay will allow more leisure without any sacrifice of consumption.

Although an individual should be able to allow for the above factors in making current choices between leisure and work, it would be impossible to expect an individual to predict how these factors will affect future choices. The rate of pay that the individual will be able to obtain will depend on his/her physical and mental capacity for work, which will begin to deteriorate at an uncertain future age and will fall to zero when the individual is no longer capable of working. It is also unlikely that any individual could predict the comparative satisfaction that will be derived from work and leisure activities many years into the future. Thus, although the disutility of work is an important factor influencing the choice between consumption and saving, it is difficult to incorporate into a quantitative model based on the maximizing lifetime utility.

**A5: Allowing for Mortality**

An unrealistic feature of the utility of consumption model is the assumption of a predetermined lifespan. Kapur and Orszag (1999) show that this defect can be remedied by allowing for survival probabilities
in projecting the utility of future consumption. They apply this method specifically to retired individuals with no bequest motive and no future earnings, so that the function to be maximized becomes:

\[ I_n = \sum_{k=n}^{\infty} U(G_k) e^{-k\theta} \frac{l_k}{l_n} \]

where the \( l_k \)s are taken from a suitable life table.

They assume that such individuals would divide their wealth between a risky asset and the purchase of whole-life annuities. If the whole-life annuities are priced using the risk-free interest rate, the modified budget constraint becomes:

\[ W_{k+1} = (1 - X_k) (W_k - G_k) e^{\rho + q_k} + X_k (W_k - G_k) e^{\delta_k} \]

where \( q_k \) is the (non-random) force of mortality over sub-interval \( k + 1 \).

Under the same conditions as stated above for Merton's closed-form solution for an iso-elastic utility function, the formula for the optimal proportion of wealth invested in the risky asset becomes:

\[ X_{n}^* = \left( \frac{\mu - r - q_n}{\gamma \sigma^2} \right) e^\rho. \]

This is similar to the result for a fixed lifespan, the only difference being that the risk-free rate \( r \) is replaced with \( r + q_k \). The implication of this result is that retired individuals should progressively switch their wealth into whole-life annuities as they grow older and disinvest in risky assets on reaching the age at which the force of mortality\(^{10}\) is greater than the risk premium on these assets.

A6: Conclusion

The utility of lifetime consumption model, as described by Merton (1969, 1971), enables us to find optimal values for how much individuals should save (or borrow) at different points in their lifespan and how their accumulated wealth should be split between risky and risk-free assets. By optimizing both consumption and portfolio choices, the model accounts for individuals who might wish to save more (or less) if past investment returns have been worse (or better) than expected. The model also allows for the desire of the individual to make bequests.

\(^{10}\)This is the force of mortality used to price annuities rather than the member's subjective force of mortality.
In applying the model as a decision-making tool in advising individuals on how to optimize their portfolio and consumption choices, various subjective items must be derived for the individual concerned. These are the utility function, the bequest function, and the subjective rate of time-preference. While the first two items might reasonably be estimated by asking the individual suitable hypothetical questions, such an approach may not be feasible for the subjective discount rate. The assumption of a constant discount rate may be a flaw in the model; it seems possible that the discount rate will depend on both the duration and the amount of consumption.

The generalized form of the model allows for future earnings from work as well as investment gains, but this leads to another problem. It is wrong to assume that workers accumulate savings purely to increase their future consumption or the size of their bequests. A powerful motive for saving is to substitute leisure for work, often by retiring before the normal retirement age of an occupation. Although utility functions that allow for the disutility of work have been proposed, the form of any such function is likely to change significantly over the lifespan of the individual. It seems unlikely that we could find a reliable method of deriving a worker's consumption-leisure utility function many years into the future.

Although the model assumes a fixed lifespan, it can be modified to allow for survival probabilities taken from an actuarial life table. As mortality increases with age, the impact of this modification on portfolio and consumption choices also will increase with age; hence, this form of the model is likely to be of most practical use for retired individuals. Although the issue of the leisure motive does not arise for retired people (as they have given up work by definition), the problems associated with estimating the subjective rate of time preference remain.

References


Dynamic Funding and Investment Strategy for Defined Benefit Pension Schemes: A Model Incorporating Asset-Liability Matching Criteria

Shih-Chieh Chang,* Cheng-Hsien Tsai,† Chia-Jung Tien,‡ and Chang-Ye Tu§

Abstract

This paper studies the dynamic funding policy and investment strategy for defined benefit pension plans using one of the most comprehensive dynamic...
pension models to date. The model includes three investable assets: one risk-free and two risky. The optimal plan decisions are formulated as a stochastic control problem that is solved using dynamic programming. The objective function uses performance measures to take into account the stability and solvency of the plan. The model is then applied to a Taiwanese pension.

Key words and phrases: optimal contribution, asset allocation, dynamic programming, performance measure

1 Introduction

Although most pension liabilities are long-term in nature, traditional defined benefit pension plan management is based on one-period assumptions. The pension plan manager seeks an optimal investment decision for the next period, based on the plan's current experience, current market conditions, and expectations about future contributions, returns, and risks. Such a short-sighted mechanism has two drawbacks: (i) the accumulation of a sequence of single-period optimal decisions across each of \( n \) periods may not be optimal for the \( n \) periods taken as a whole; and (ii) single-period decisions have difficulties in dealing with the investment and funding sides of a pension plan because the interaction between investments and funding appears only in the multi-period setting.

An important tool that can be used to assist plan managers in developing optimal funding policies over many periods is stochastic optimal control theory. This theory can be used to solve long-term financial planning problems through global optimization across periods instead of local optimization within a period.

Control theory has been developed by engineers since the 1930s. Its applications to economics emerged in the 1950s. [See Petit (1990) for more on this.] Several authors, including Samuelson (1969), Merton (1971, 1990), Brennan and Schwartz (1982), Karatzas et al., (1986), Brennan, Schwartz, and Lagnado (1997), Boyle and Yang (1997), Brennan and Schwartz (1998), and Sorensen (1999), have studied optimal consumption and investment problems using control theory. Although the popularity of stochastic control theory was hindered by its inher-

---

1 Traditional pension management usually employs a mean-variance approach. Sharpe (1991) describes the mean-variance approach as a highly parsimonious characterization of investors' goals, employing a myopic view (i.e., one period at a time) and focusing on only two aspects of the probability distribution of possible returns over that period.
Chang: Dynamic Funding and Investment Strategy

ent complexity, it is becoming more popular today due to the ready availability of high-speed computers.

The application of control theory to pension plan management started with O'Brien (1986, 1987) who constructed a stochastic model for the pension plan and studied the optimal funding policies for target funding ratios. Cairns (1995, 1996, 2000) introduced asset allocation into the control process to study the optimal funding and investment strategies needed to minimize certain quadratic loss functions. Chang (1999, 2000) applied their methods to a real pension plan in Taiwan using service tables and stochastic asset returns to numerically solve for optimal funding policies over various time horizons. Applications of control theory to other actuarial problems can be seen in Runggaldier (1998) and Schäl (1998). Runggaldier reviews the concepts and solution methods, while Schäl focuses on the dynamic programming for piecewise deterministic Markov processes.

In this paper we construct one of the most comprehensive dynamic models of a pension plan to date, numerically solve the stochastic control problem, and provide illustrations of the optimal investment and funding strategies. Compared to Cairns (1995, 1996, 2000), we have a richer set of liability dynamics, and we have included risk-free as well as risky assets. Compared with Chang (1999, 2000), we consider not only funding policies, but also asset allocations. In addition we consider more risk factors for invested assets. Furthermore, our performance measures (also called loss functions) take into account the stability of contributions and the security (the funding ratio) of the pension plan.

The features of our models are summarized as follows:

1. The dynamics of the plan's demography can be explicitly incorporated into investment decisions under different evaluation time horizons.

2. The optimal funding and investment strategies of the plan can be formalized with specific risk performance measures through a computerized system.

3. The contribution risk and solvency risk associated with any funding policy and investment strategy can be assessed given any evaluation horizon.

The paper is organized as follows. Section 2 describes the model (i.e., the basic framework, the dynamics of invested assets, and performance measures used) of the proposed dynamic optimization scheme.
Section 3 develops the solution to the optimal equation. Section 4 provides a practical example to illustrate the usefulness of the theory presented. A summary and closing comments are given in Section 5.

2 The Model

In this section, we formulate the funding and investment decisions of pension funds as an stochastic optimal control problem. These decisions are modeled through a continuous-time stochastic process over a specific evaluation time horizon.

2.1 The Basic Framework

The following notation is used for the various stochastic processes used in the paper:

\[ T = \text{Management's planning horizon}; \]
\[ \mathcal{F}_t = \text{Plan's history up to time } t; \]
\[ F(t) = \text{Total assets of the pension plan at time } t \text{ excluding any contributions made at time } t; \]
\[ d\delta(t, F) = \text{Rate of investment return in } (t, t + dt); \]
\[ C(t) = \text{Contributions at time } t; \]
\[ B(t) = \text{Retirement benefit payment rate at time } t; \]
\[ \sigma_B = \text{Volatility of } B(t); \]
\[ Z_i(t) = \text{Wiener processes } (i \in \{NC, B, W\} \text{ at time } t; \]
\[ AL(t) = \text{Total plan accrued liabilities at time } t; \]
\[ r' = \text{Valuation rate for accrued liabilities}; \]
\[ NC(t) = \text{Normal cost rate at time } t; \]
\[ W(t) = \text{Reduction rate in retirement liability at time } t \text{ due to withdrawal payment}; \]

Throughout this paper we assume that all stochastic processes are defined on appropriate probability spaces.

Because death benefits are not currently included in the Taiwanese Labor Standard Law and withdrawal benefits are not paid from the accumulated pension fund, we consider only the retirement benefits payments and the reduced withdrawal liability in this paper.
Chang: Dynamic Funding and Investment Strategy

\( \sigma_{NC} = \) Volatility of \( NC(t) \); and

\( \sigma_{W} = \) Volatility of \( W(t) \).

The \( \sigma \)s are assumed to be constants. The term rate, as used with respect to \( C(t), NC(t), B(t), \) and \( W(t) \), refers to the amount paid in an infinitesimal time interval. For example, \( C(t)dt \) is the amount contributed in \((t, t + dt)\).

The funding level \( F(t) \) and accrued liabilities \( AL(t) \) are described by the following stochastic differential equations:

\[
\begin{align*}
\mathrm{d}F(t) &= F(t)\mathrm{d}\delta(t, F) + C(t)\mathrm{d}t - B(t)\mathrm{d}t + \sigma_B \mathrm{d}Z_B(t), \\
\mathrm{d}AL(t) &= (AL(t)r' + NC(t) - B(t) - W(t))\mathrm{d}t + \sigma_{NC} \mathrm{d}Z_{NC}(t) \\
&\quad + \sigma_B \mathrm{d}Z_B(t) + \sigma_W \mathrm{d}Z_W(t).
\end{align*}
\]  

(1)

(2)

2.2 The Dynamics of Invested Assets

We assume three types of assets are available to the pension plan: cash, stocks, and consol bonds.\(^4\) The proportion of the pension funds invested in stocks, consol bonds, and cash at time \( t \) is denoted by \( P(t) = (a(t), b(t), c(t)) \), where

\[ a(t) + b(t) + c(t) = 1. \]

Following Brennan, Schwartz, and Lagnado (1997), we model the instantaneous rate of return on the stock portfolio \( \frac{\mathrm{d}S(t)}{S(t)} \), the short rate \( r(t) \), and the long rate \( l(t) \), as the following joint stochastic process:

\[
\begin{align*}
\frac{\mathrm{d}S(t)}{S(t)} &= \mu_S \mathrm{d}t + \sigma_S \mathrm{d}Z_S(t), \\
\mathrm{d}r(t) &= \mu_r \mathrm{d}t + \sigma_r \mathrm{d}Z_r(t), \\
\mathrm{d}l(t) &= \mu_l \mathrm{d}t + \sigma_l \mathrm{d}Z_l(t),
\end{align*}
\]  

(3)

(4)

(5)

where the subscripts \( S, r, l \) refer to stocks, short rate, and long rate, respectively, \( \mu_i \) and \( \sigma_i (i \in \{S, r, l\}) \) are constant parameters and \( \mathrm{d}Z_i (i \in \{S, r, l\}) \) are increments to Wiener processes.

We note that the price of a consol bond, \( B_c(t) \), is inversely proportional to its yield and the total return of a consol bond is the sum of

\(^4\)Consol bonds are bonds with infinite time to maturity, i.e., they never mature.
the yield and the price change. Also, from a simple application of Ito’s lemma, the instantaneous total return on the consol bond can be proved to be:

\[
\frac{dB_c(t)}{B_c(t)} + l(t) dt = \left( l(t) - \frac{\mu_l}{l(t)} + \frac{\sigma^2_t}{l(t)^2} \right) dt - \frac{\sigma_l}{l(t)} dZ_l(t). \tag{6}
\]

The investment return of the pension plan between time \( t \) and \( t + dt \), \( d\delta(t, F) \), can be formulated as:

\[
d\delta(t, F) = a(t) \frac{dS(t)}{S(t)} + b(t) \left( \frac{dB_c(t)}{B_c(t)} + l(t) dt \right) + (1 - a(t) - b(t)) r(t) dt. \tag{7}
\]

Hence, the instantaneous changes in the pension’s assets in equation (1) can be rewritten as

\[
dF(t) = F(t) \left[ a(t) \mu_S + b(t) \left( l - \frac{\mu_l}{l(t)} + \frac{\sigma^2_t}{l(t)^2} \right) + (1 - a(t) - b(t)) r(t) \right] dt
+ (C(t) - B(t)) dt + F(t) a(t) \sigma_S dZ_S(t) - F(t) b(t) \frac{\sigma_l}{l(t)} dZ_l(t)
+ \sigma_B dZ_B(t). \tag{8}
\]

The dynamics of the pension plan for the fund and the accrued liabilities can then be jointly written as:

\[
\begin{pmatrix} dF(t) \\ dAL(t) \end{pmatrix} = \begin{pmatrix} \mu_X(t) dt + \sigma_X(t) d\xi(t) \end{pmatrix}, \tag{9}
\]

where

\[
\mu_X(t) = \begin{pmatrix} a \mu_S + b(l - \mu_l/l + \sigma^2_t/l^2) + (1 - a - b) r \end{pmatrix} F + C - B, \quad \begin{pmatrix} \sigma_S \sigma_B \sigma_{NC} \sigma_W \end{pmatrix},
\]

and \( d\xi = (dZ_S, dZ_l, dZ_B, dZ_{NC}, dZ_W)^T \) is a five dimensional standard Wiener process with covariance matrix \([\sigma_{ij}]\). We adopt the notation \( \sigma_{S1} \) to denote the covariance between Wiener processes \( Z_S \) and \( Z_l \); other covariances are represented similarly.
Note: In the definition of $\mu_X(t)$ and $\sigma_X(t)$ above the function definitions are abbreviated by dropping (t) so that, for example, $F(t)$ is written as $F$ and $a(t)$ is written as $a$. When no confusion arises, (t) and subscripts $t$ or $T$ will be omitted. This convention is used throughout the rest of this paper.

2.3 Performance Measures

A good performance measure should consider the two most important factors in pension plan valuations: (i) the contribution rate risk (i.e., level of funding deficiency), which is the difference between normal costs and contributions; and (ii) the level of unfunded liabilities, which is the difference between accrued liabilities and assets. The level of funding deficiency affects the stability of the plan, while the level of unfunded liabilities affects the solvency of the plan. Following Haberman and Sung (1994), we design our performance measure (also called a loss function) to take into account the contribution rate and solvency risks to give:

$$L(t, X, \{C, P\}) = (NC(t) - C(t))^2 + k(\eta AL(t) - F(t))^2,$$

where $k$ is a constant chosen subjectively by the pension fund manager to adjust for the difference in size between the contribution rate risk and the solvency risk, and $\eta$ is the target funding ratio. The parameter $k$ reflects the relative importance of matching contributions with normal costs and matching plan assets with accrued liabilities.

3 The Optimal Equation

Assume first that the performance measure $L(t, X, \{C, P\})$ is discounted continuously by a constant rate $\rho$. If we let $B[X_T, T]$ be a function measuring the loss associated with the state of the pension plan at the end of period, then the problem of choosing the optimal asset allocation and funding policy for a fixed evaluation time horizon $T$ can be formulated as:

$$\inf \{ \int_0^T e^{-\rho t} L(t, X, \{C, P\}) dt + B[X_T, T] \},$$

subject to the asset and liability dynamics specified in equations (1) and (2), respectively. Furthermore, we have $C \geq 0$, $a$, $b$, $c \in R$, and $F(t) \geq 0$. Define
\[ J(s, X, \{C, P\}) = \mathbb{E}_{s, X} \left[ \int_{s}^{T} e^{-\rho t} L(t, X, \{C, P\}) dt + B[X_T, T] \bigg| \mathcal{F}_s \right] \quad (11) \]

where \( \mathbb{E}_{s, X} \) represents the expectation conditioned on being in the state \( X \) at time \( s \), given information \( \mathcal{F}_s \). As \( X \) is assumed to follow a time-homogeneous Markov process, only the information at \( s \) is needed (information before time \( s \) can be ignored).

Let us assume that there exist optimal strategies \( \{C^*, P^*\} \) that form a set of admissible control functions that minimize equation (11), i.e.,

\[ I(s, X) = \inf_{(C,P)} J(s, X, \{C, P\}) = J(s, X, \{C^*, P^*\}). \]

Then by the principle of optimality [Bellman (1957)], we can express the Bellman-Dreyfus fundamental equation of optimality as

\[ \inf_{(C,P)} \{ [l^2 (-B + C - (-1 + a + b) Fr + bFl + aF\mu_s - bF\mu_l) + bF\sigma_l^2] \frac{1}{l^2} I_F + (-B + NC - W + r' AL) I_{AL} + (\sigma_B^2 + \sigma_{NC}^2 + \sigma_W^2 + 2 \sigma_{NC}\sigma_W \sigma_{NC_W} + 2 \sigma_{NC}\sigma_B \sigma_{NC_B} + 2 \sigma_W \sigma_B \sigma_{WB}) \frac{1}{2} I_{AL,AL} + I_{F,F}(l^2 \sigma_B^2 + (\sigma_{NC}(a l \sigma_S \sigma_{NC,S} - b \sigma_I \sigma_{NC,I})) + \sigma_B(l^2 \sigma_{NC, B} + aFl \sigma_S \sigma_{SB} + lO_w \sigma_{WB} - bF \sigma_I \sigma_{IB})) + e^{-\rho s}( (NC - C)^2 + k(F - \eta AL)^2) + [l^2 \sigma_B^2 + (b^2 \sigma_l^2 + a l \sigma_S (a l \sigma_S - 2 b \sigma_I \sigma_{SL})) F^2 + 2Fl \sigma_B (a l \sigma_S \sigma_{SB} - b \sigma_I \sigma_{IB}) \frac{1}{2l^2} I_{F,F} \} + I_s = 0 \quad (12) \]

where the subscripts of the \( I \) denote partial derivatives, i.e., \( I_s = \partial I / \partial s \), \( I_{F,F} = \partial^2 I / \partial F^2 \), \( I_{F,AL} = \partial^2 I / \partial F \partial AL \), etc. Taking the partial derivatives with respect to \( C \) and \( P \) and using the boundary condition \( I(T, X) = B[X, T] \), we obtain from the first order condition that:

\[ C^* = NC - e^{\rho s} I_F / 2 \quad (13) \]

\[ a^* = \frac{1}{Fl I_{F,F} \sigma_S^2 \sigma_l (-1 + \sigma_S^2)} \left[ \left( I_{F,l} l(-r + l) - \mu_l \right) \sigma_S \sigma_{SI} + I_F \sigma_S \sigma_l^2 \sigma_{SI} + \left( I_{F,F} l(-r + \mu_S) + \sigma_S (I_{F,F} \sigma_B (\sigma_{SB} - \sigma_{SI} \sigma_{IB}) + I_{F,l} (\sigma_{NC} \sigma_{NC,S} + \sigma_{W} \sigma_{SW} - \sigma_{NC} \sigma_{NC,I} \sigma_{SI} + \sigma_B \sigma_{SB} - \sigma_{W} \sigma_{SI} \sigma_{WL} - \sigma_B \sigma_{SI} \sigma_{IB})) \right) \right] \quad (14) \]
The Hamilton-Jacobi-Bellman equation (Fleming and Rishel, 1975) is:

\[
0 = I_S + I_{AL}(-B + NC - W + ALr') + \frac{1}{2} I_{AL,AL}(\sigma_B^2 + \sigma_{NC}^2 + \sigma_W^2 + 2 \sigma_{NC} \sigma_W \sigma_{NC,W}) + 2 \sigma_W \sigma_B \sigma_{WB} + 2 \sigma_{NC} \sigma_B \sigma_{NC,B})
\]

\[
+ \frac{I_{F,F}^2}{2I_{F,F}} (\sigma_{NC}(\sigma_{NC,l} - \sigma_{NC,S} \sigma_{SI}) + \sigma_W(-\sigma_{SW} \sigma_{SI} + \sigma_{WI}) + \sigma_B(-\sigma_{SI} \sigma_{SB} + \sigma_{IB}))^2
\]

\[
- I_{F,AL}^2 (\sigma_{NC} \sigma_{NC,S} + \sigma_W \sigma_{SW} + \sigma_B \sigma_{SB})^2
\]

\[
+ \frac{I_{F,F} I_{F,AL} r - \mu_S (\sigma_{NC} \sigma_{NC,S} + \sigma_W \sigma_{SW} + \sigma_B \sigma_{SB})}{I_{F,F} \sigma_S}
\]

\[
- \frac{1}{2} I_{F,F} \sigma_B^2 (-1 + \sigma_{SB}^2) + \frac{I_{F,F} I_{F,AL} \sigma_S (l^2(-r + l) - l\mu_l + \sigma_l^2)}{l I_{F,F} \sigma_S \sigma_l (-1 + \sigma_{SI}^2)}
\]

\[
+ l(-r + \mu_S) \sigma_l \sigma_{SI} (\sigma_{NC}(-\sigma_{NC,l} + \sigma_{NC,S} \sigma_{SI}) + \sigma_W(\sigma_{SW} \sigma_{SI} - \sigma_{WI}) + \sigma_B(\sigma_{SI} \sigma_{SB} - \sigma_{IB}))
\]

\[
+ \frac{I_{F,AL} \sigma_B}{(-1 + \sigma_{SI}^2)} (\sigma_{NC}(-\sigma_{NC,l} + \sigma_{NC,S} \sigma_{SI}) + \sigma_W(\sigma_{SW} \sigma_{SI} - \sigma_{WI}) + \sigma_B(\sigma_{SI} \sigma_{SB} - \sigma_{IB}))
\]

\[
+ \sigma_B (\sigma_{SI} \sigma_{SB} - \sigma_{IB})(\sigma_{SI} \sigma_{SB} - \sigma_{WB}) + I_{F,AL} \sigma_B \sigma_{NC}(\sigma_{NC,B} - \sigma_{NC,S} \sigma_{SB})
\]

\[
+ I_{F,AL} \sigma_B(-\sigma_B(-1 + \sigma_{SB}^2) + \sigma_W(\sigma_{SW} \sigma_{SB} + \sigma_{WB}))
\]

\[
+ \frac{I_{F,F}^2 (\sigma_S l((-r - l)l + \mu_l) - \sigma_l^2) + l(r - \mu_S) \sigma_l \sigma_{SI}}{2l^2 I_{F,F} \sigma_S^2 \sigma_l^2 (-1 + \sigma_{SI}^2)}
\]

\[
- \frac{1}{4} e^{(r - \mu_S)} l^2_{F,1} (r - \mu_S)^2 + \frac{I_{F,F} \sigma_B^2 (-\sigma_{SI} \sigma_{SB} + \sigma_{IB})^2}{2(-1 + \sigma_{SI}^2)}
\]

\[
+ ke^{(r - \mu_S)} (F - \eta \sigma_{SI})^2 + \frac{I_{F,F} \sigma_S (l^2(-r + l) - l\mu_l + \sigma_l^2)}{l \sigma_S \sigma_l (-1 + \sigma_{SI}^2)}
\]

\[
+ l(-r + \mu_S) \sigma_l \sigma_{SI} (\sigma_{SI} \sigma_{SB} - \sigma_{IB})
\]

\[
- \frac{1}{I_{F}} I_{F}((B - NC - Fr)(r + (-\mu_S) \sigma_B \sigma_{SB})].
\]
The function \((C^*, P^*)\) is our candidate for the optimal strategy. Because \(I(s, X)\) is unknown, the description is incomplete. We therefore have to guess a solution for \(I(s, X)\) that has finite number of parameters, and then use the partial differential equation (16) to identify its parameters. Because our loss function is quadratic, our guess is:

\[
I[s, X] = X^t \Phi X + \Psi^t X + d(s), \quad 0 \leq s \leq T.
\]

where \(d(s)\) depends only on \(s\), and

\[
\Phi = \begin{pmatrix} a_1(s) & a_2(s) \\ a_2(s) & a_3(s) \end{pmatrix} \quad \text{and} \quad \Psi = \begin{pmatrix} b_1(s) \\ b_2(s) \end{pmatrix}.
\]

Substituting \(\Phi, \Psi, \text{ and } d(s)\) into the ordinary system of differential equations and noting that the optimal strategy must hold for all \((s, X)\), we can solve for the coefficient matrices \(\Phi, \Psi, \text{ and } d(s)\) in \(I[s, X]\) by checking the coefficients in equations for \(F^2, F_{AL}, AL^2, F, \text{ and } AL\), and the constant term. The boundary condition is \(I(T, X) = B[X, T] = 0\), i.e., \(a_1(T) = a_2(T) = a_3(T) = b_1(T) = b_2(T) = d(T) = 0\), because the plan manager who adopts this optimal strategy must be able to match contributions with the normal costs and match assets with accrued liabilities at the end of the evaluation period.

After inspecting the formulas for \(a, b, \text{ and } C\), we find that we need only to compute \(h, h,p, \text{ and } h,AL\) via the solution of \(a_1(s), a_2(s), \text{ and } b_1(s)\). The system of differential equations involving \(a_1(s), a_2(s), \text{ and } b_1(s)\) (with \(s\) removed for convenience) is as follows:

\[
0 = l^2(r - \mu_S)^2 \sigma_1^2 a_1 - 2l(r - \mu_S)\sigma_S \sigma_l(l^2(-r + l) - l\mu_l + \sigma_l^2)\sigma_{SI} a_1 + \sigma_1^2(l^2((r - l)l + \mu_l)^2 a_1 + \sigma_1^4 a_1 + l\sigma_l^2(-k e^{-\rho s})
\]

\[
-2((2r - l)l + \mu_l)a_1 + l\sigma_l^2(k e^{-\rho s})
\]

\[
+ 2r a_1 - e^{\rho s}a_1^2 + a_1')
\] (17)

\[
0 = 2(l^2(r - \mu_S)^2 \sigma_1^2 a_1 - 2l(r - \mu_S)\sigma_S \sigma_l(l^2(-r + l) - l\mu_l + \sigma_l^2)\sigma_{SI} a_2 + \sigma_1^2(l^2((r - l)l + \mu_l)^2 a_2 + \sigma_1^4 a_2 + l\sigma_l^2(k l e^{-\rho s})
\]

\[
+ (-2\mu_l + l(-3r + 2l - r' + e^{\rho s}a_1))a_2
\]

\[
- la_2' - l\sigma_l^2(k l e^{-\rho s}) + (r - r' - e^{\rho s}a_1)a_2 + a_2'))
\] (18)
\[ 0 = l^2 (r - \mu_S)^2 \sigma_1^2 b_1 + 2l (r - \mu_S) \sigma_S \sigma_1 (l \sigma_1 ((\sigma_{NC} (-\sigma_{NC,S} + \sigma_{NC,I} \sigma_{S1}) + \sigma_W (\sigma_{SW} + \sigma_{SW} \sigma_{WI})) a_2 - \sigma_B (\sigma_{SB} - \sigma_{SI} b_1) \\
+ (l (r - l) l + \mu_l - \sigma_1^2) \sigma_{SI} b_1) \\
+ \sigma_1^2 (2l^2 (l (-r + l) - \mu_l) \sigma_1 ((\sigma_{NC} (-\sigma_{NC,I} + \sigma_{NC,S} \sigma_{S1}) + \sigma_W (\sigma_{SW} \sigma_{SI} \sigma_{SI} - \sigma_{WI})) a_2 \\
+ \sigma_B (\sigma_{SI} \sigma_{SB} - \sigma_{IB})(a_1 + a_2)) + 2l \sigma_1^3 ((\sigma_{NC} (-\sigma_{NC,I} + \sigma_{NC,S} \sigma_{S1}) + \sigma_W (\sigma_{SW} \sigma_{SI} \sigma_{SI} - \sigma_{WI})) a_2 + \sigma_B (\sigma_{SI} \sigma_{SB} - \sigma_{IB})(a_1 + a_2)) \\
+ l^2 ((r - l) l + \mu_l)^2 b_1 + \sigma_1^4 b_1 \\
+ l \sigma_1^2 (-2 (B - NC + W) l (-1 + \sigma_S^2)) a_2 \\
+ (-2 \mu_l + l (-3 r + 2l + r \sigma_S^2)) b_1 - l (-1 + \sigma_S^2) a_1 (2B - 2NC \\
+ e^{\psi_S} b_1) + l (-1 + \sigma_S^2) b_1)). \] (19)

Numerical methods can now be used to solve the above system of differential equations for \( \Phi, \Psi \) and \( d \) in order to obtain the optimal strategy \( C^*, P^* \) for \( t \in [0, T] \). A Mathematica® subroutine that implements the sophisticated implicit Adams and Gear formulae with higher orders is used to solve the system of differential equations numerically.

4 An Illustration

We apply the results in Sections 2 and 3 to a defined benefit pension plan sponsored by a Taiwanese semi-conductor and electronic company. According to the Labor Standard Law enacted by the Taiwanese government in 1984, each employer is required to contribute from 2% to 15% of its employees' pensionable payroll to a government-managed trust fund. This trust fund is guaranteed a minimum return by the Taiwanese government. This mandatory pension plan is a defined benefit scheme in which a participant's retirement benefit is based on the participant's length of service and final salary. Although the pension plan has minimum guaranteed returns from the government, the plan is still subject to an insolvency risk because the contributions coupled with the investment returns may not be able to match the benefit payments.

The Taiwanese company's pension plan covers 2,768 members with initial assets of 254 million NT dollars and an accrued liability of 380 million NT dollars. We use the open group with size-constrained assumption to project the evolution of the plan's workforce (Winklevoss, 1993). The total number of employees is assumed to remain unchanged, i.e., each departing employee is immediately replaced by a new em-
ployee. The plan's service table is given in Tables A1 and A2 in the appendix. Table A3 shows the assumptions made for new entrants. The entry age normal cost method (Anderson, 1992) is used to determine accrued liabilities, normal costs, and benefit retirement benefits.

The retirement benefit, \( B(t) \), is formulated according to the current Labor Standard Law for the private pension plans in Taiwan. This law stipulates that the retirement benefit can only be paid as lump-sum payment to the retiree. The retirement benefit for a qualified plan member, \( B(t) \), can be written as the accumulated service credits multiple by the average final six-month salary. Each plan member receives two service credits each year of service for the first fifteen years of service and one service credit each year after fifteen years of service, up to a maximum of forty-five service credits.

Suppose an active plan member \( j \) entered the plan at age \( e_j \) and is currently age \( x_j \) at with annual salary of \( S^{(j)} \). If the annual salary growth rate is set at 3%, then the projected retirement benefit, \( \text{RBEN}^{(j)} \), can be written as

\[
\text{RBEN}^{(j)} = \begin{cases} 
2x_j S^{(j)} (1.03)^{60-x_j} & \text{if } 60 - e_j \leq 15 \\
S^{(j)} (1.03)^{60-x_j} \times \min(x_e + 15, 45) & \text{if } 60 - e_j \geq 15 
\end{cases}
\]

Let \( R_t \) denote the set of active plan members who retire at time \( t \) and \( W_t \) denote the set of active plan members who withdraw at time \( t \). The aggregated retirement benefit payments and reduced retirement liability payments at \( t \), respectively, for the plan are given by:

\[
B(t) = \sum_{j \in R_t} \text{RBEN}^{(j)} \quad \text{and} \\
W(t) = \sum_{j \in W_t} \text{RBEN}^{(j)}.
\]

For practical reasons, continuous time processes are simulated at weekly intervals. Salaries are assumed to increase 3% per annum and the valuation interest rate is \( \ln(1.03)/52 \) per week. Because we do not have the data to estimate the volatilities of normal costs and the retirement and withdrawal benefits, we first simulate 100 sets of NC, \( B \),

---

5 These are simplified service tables that give only the multiple decrement survival probability \( p_x^{(T)} \) (i.e., the probability that a person age \( x \) remains in the plan at age \( x + 1 \)) for males and females. The illustrated pension plan used by Chang and Cheng (2002) is different from the plan used in this study. They focus on a public pension plan (Tai-PERS), while we discuss private pension plans.
and $W$, then use the simulated standards as the volatilities in equation (2). With regard to the parameters associated with the loss function, the rate $\rho$ is assumed to be $\ln(1.06)/52$ per week. Two target funding ratios, 0.75 and 1, are used for comparisons, and we subjectively choose $k = 0.0001$.

Table 1

Simulated Paths $NC(t)$, $AL(t)$, $l(t)$ and $r(t)$

<table>
<thead>
<tr>
<th>$t$</th>
<th>$NC(t)$</th>
<th>$AL(t)$</th>
<th>$l(t)$</th>
<th>$r(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.947</td>
<td>381.0</td>
<td>0.0700</td>
<td>0.0350</td>
</tr>
<tr>
<td>25</td>
<td>0.958</td>
<td>388.0</td>
<td>0.0710</td>
<td>0.0370</td>
</tr>
<tr>
<td>50</td>
<td>0.973</td>
<td>393.0</td>
<td>0.0715</td>
<td>0.0360</td>
</tr>
<tr>
<td>75</td>
<td>0.986</td>
<td>402.0</td>
<td>0.0730</td>
<td>0.0340</td>
</tr>
<tr>
<td>100</td>
<td>0.995</td>
<td>409.0</td>
<td>0.0735</td>
<td>0.0320</td>
</tr>
<tr>
<td>125</td>
<td>1.010</td>
<td>417.0</td>
<td>0.0730</td>
<td>0.0330</td>
</tr>
<tr>
<td>150</td>
<td>1.020</td>
<td>421.0</td>
<td>0.0760</td>
<td>0.0310</td>
</tr>
<tr>
<td>175</td>
<td>1.030</td>
<td>431.0</td>
<td>0.0740</td>
<td>0.0300</td>
</tr>
<tr>
<td>200</td>
<td>1.040</td>
<td>439.0</td>
<td>0.0720</td>
<td>0.0290</td>
</tr>
<tr>
<td>225</td>
<td>1.048</td>
<td>442.0</td>
<td>0.0715</td>
<td>0.0310</td>
</tr>
<tr>
<td>250</td>
<td>1.058</td>
<td>449.0</td>
<td>0.0750</td>
<td>0.0300</td>
</tr>
<tr>
<td>275</td>
<td>1.060</td>
<td>453.0</td>
<td>0.0730</td>
<td>0.0300</td>
</tr>
<tr>
<td>300</td>
<td>1.072</td>
<td>462.0</td>
<td>0.0725</td>
<td>0.0315</td>
</tr>
<tr>
<td>325</td>
<td>1.073</td>
<td>470.0</td>
<td>0.0740</td>
<td>0.0305</td>
</tr>
<tr>
<td>350</td>
<td>1.080</td>
<td>475.0</td>
<td>0.0730</td>
<td>0.0330</td>
</tr>
<tr>
<td>375</td>
<td>1.083</td>
<td>480.0</td>
<td>0.0725</td>
<td>0.0320</td>
</tr>
<tr>
<td>400</td>
<td>1.092</td>
<td>490.0</td>
<td>0.0715</td>
<td>0.0300</td>
</tr>
<tr>
<td>425</td>
<td>1.099</td>
<td>500.0</td>
<td>0.0700</td>
<td>0.0320</td>
</tr>
<tr>
<td>450</td>
<td>1.110</td>
<td>508.0</td>
<td>0.0690</td>
<td>0.0340</td>
</tr>
<tr>
<td>475</td>
<td>1.120</td>
<td>518.0</td>
<td>0.0710</td>
<td>0.0330</td>
</tr>
<tr>
<td>500</td>
<td>1.160</td>
<td>526.0</td>
<td>0.0720</td>
<td>0.0310</td>
</tr>
<tr>
<td>520</td>
<td>1.200</td>
<td>540.0</td>
<td>0.0740</td>
<td>0.0320</td>
</tr>
</tbody>
</table>

Notes: $t$ is in weeks; $NC(t)$ and $AL(t)$ in 1,000,000 NT dollars.
The parameters for the dynamics of the assets are taken from Brennan, Schwartz, and Lagnado (1998):

\[
\frac{dS}{S} = 0.009992 \, dt + 0.041 \, dW_S, \\
dr = 0.000158 \, dt + 0.005472 \, dW_r, \\
dl = -0.0002236 \, dt + 0.00304 \, dW_l.
\]

The initial values for \( r \) and \( l \) are 3.5% and 7%, respectively.

The simulated paths of normal costs, accrued liabilities, short rate, and long rate are shown in Table 1.

We assume the following covariance (or correlation) matrix:\(^6\)

\[
\begin{bmatrix}
1 & \sigma_{S,NC} & \sigma_{W,NC} & \sigma_{R,NC} & \sigma_{I,NC} \\
\sigma_{NC,S} & 1 & \sigma_{WS} & \sigma_{RS} & \sigma_{IS} \\
\sigma_{NC,W} & \sigma_{SW} & 1 & \sigma_{BW} & \sigma_{IW} \\
\sigma_{NC,B} & \sigma_{SB} & \sigma_{WB} & 1 & \sigma_{IB} \\
\sigma_{NC,L} & \sigma_{SL} & \sigma_{WL} & \sigma_{BL} & 1
\end{bmatrix}
\]

\[
= \begin{bmatrix}
1 & 0.2 & -0.3 & 0.1 & 0.4 \\
0.2 & 1 & -0.25 & -0.5 & 0.3 \\
-0.3 & -0.25 & 1 & 0.1 & -0.2 \\
0.1 & -0.5 & 0.1 & 1 & 0.1 \\
0.4 & 0.3 & -0.2 & 0.1 & 1
\end{bmatrix}
\]

To illustrate the solved optimal control law, we first simulate a set of paths that includes a short rate path and a long rate path. These paths are then plugged into the system of differential equations given in equations (17), (18), and (19) to solve for the optimal strategies corresponding to the \((t, X)\). The solved optimal strategies are the optimal contributions per week and the corresponding optimal contribution rates\(^7\) under 5-year and 10-year evaluation periods and target funding ratios of \( \eta = 1 \) and \( \eta = 0.75 \). Tables 2, 3, and 4 show results for \( \eta = 1 \).

---

\(^6\) As we are using standard Wiener processes, the covariances and the correlations are the same.

\(^7\) Weekly contributions are measured in dollars, while weekly contribution rates are measured by the ratio of weekly contributions to salary.
The optimal weekly contributions and contribution rates are shown in Table 2 for $\eta = 1$. The optimal weekly contributions and contribution rates increase steadily with time. Such increases are reasonable because normal costs increase with the aging of the employees in the plan. Table 3 show the resulting fund when the pension plan adopts the optimal investment strategies under different evaluation periods (5 and 10 years) and a target funding ratio of 1. Table 4 shows the evolution of the optimal mix of stocks, bonds and cash under different evaluation periods.
periods and a target funding ratio of 1. A summary of the results is given in Table 5.

Table 3
Evolution of the Optimal Fund and Fund Ratios with $\eta = 1$
Under 10-Year and 5-Year Time Horizons

<table>
<thead>
<tr>
<th>Week</th>
<th>Fund ($F(t)$) 10-Years</th>
<th>Fund ($F(t)$) 5-Years</th>
<th>Fund Ratio 10-Years</th>
<th>Fund Ratio 5-Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.510</td>
<td>2.510</td>
<td>0.485</td>
<td>0.580</td>
</tr>
<tr>
<td>25</td>
<td>2.930</td>
<td>3.030</td>
<td>0.550</td>
<td>0.670</td>
</tr>
<tr>
<td>50</td>
<td>3.400</td>
<td>3.400</td>
<td>0.620</td>
<td>0.740</td>
</tr>
<tr>
<td>75</td>
<td>3.600</td>
<td>3.580</td>
<td>0.685</td>
<td>0.800</td>
</tr>
<tr>
<td>100</td>
<td>3.800</td>
<td>3.750</td>
<td>0.720</td>
<td>0.830</td>
</tr>
<tr>
<td>125</td>
<td>3.950</td>
<td>3.900</td>
<td>0.750</td>
<td>0.875</td>
</tr>
<tr>
<td>150</td>
<td>4.100</td>
<td>4.080</td>
<td>0.760</td>
<td>0.900</td>
</tr>
<tr>
<td>175</td>
<td>4.280</td>
<td>4.270</td>
<td>0.780</td>
<td>0.905</td>
</tr>
<tr>
<td>200</td>
<td>4.300</td>
<td>4.300</td>
<td>0.800</td>
<td>0.915</td>
</tr>
<tr>
<td>225</td>
<td>4.350</td>
<td>4.350</td>
<td>0.810</td>
<td>0.950</td>
</tr>
<tr>
<td>250</td>
<td>4.420</td>
<td>4.420</td>
<td>0.820</td>
<td>0.975</td>
</tr>
<tr>
<td>260</td>
<td>4.430</td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td>275</td>
<td>4.480</td>
<td></td>
<td>0.830</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>4.510</td>
<td></td>
<td>0.850</td>
<td></td>
</tr>
<tr>
<td>325</td>
<td>4.600</td>
<td></td>
<td>0.860</td>
<td></td>
</tr>
<tr>
<td>350</td>
<td>4.680</td>
<td></td>
<td>0.880</td>
<td></td>
</tr>
<tr>
<td>375</td>
<td>4.750</td>
<td></td>
<td>0.895</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>4.800</td>
<td></td>
<td>0.900</td>
<td></td>
</tr>
<tr>
<td>425</td>
<td>4.850</td>
<td></td>
<td>0.905</td>
<td></td>
</tr>
<tr>
<td>450</td>
<td>4.920</td>
<td></td>
<td>0.920</td>
<td></td>
</tr>
<tr>
<td>475</td>
<td>5.000</td>
<td></td>
<td>0.935</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>5.080</td>
<td></td>
<td>0.950</td>
<td></td>
</tr>
<tr>
<td>520</td>
<td>5.140</td>
<td></td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Notes: The fund is measured in 100,000,000 NT dollars.

To quickly achieve the target funding ratio, the plan manager has to take some unusual positions (such as large amounts of short or long positions) at the beginning. These extreme positions are mainly driven by the parameters of financial market processes and the choice of $k$. As the optimal control law is sensitive to the estimated or chosen pa-
rameters, the plan manager should pay close attention to the choice of parameters.

### Table 4
Proportions of Stocks, Bonds, and Cash Under 10-Year and 5-Year Time Horizons with $\eta = 1$

<table>
<thead>
<tr>
<th>Weeks</th>
<th>10-Year Horizon</th>
<th>5-Year Horizon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stocks</td>
<td>Bonds</td>
</tr>
<tr>
<td>0</td>
<td>-4.100</td>
<td>8.800</td>
</tr>
<tr>
<td>25</td>
<td>-2.750</td>
<td>4.800</td>
</tr>
<tr>
<td>50</td>
<td>-1.400</td>
<td>2.600</td>
</tr>
<tr>
<td>75</td>
<td>-0.650</td>
<td>1.600</td>
</tr>
<tr>
<td>100</td>
<td>-0.250</td>
<td>1.000</td>
</tr>
<tr>
<td>125</td>
<td>-0.200</td>
<td>0.500</td>
</tr>
<tr>
<td>150</td>
<td>-0.050</td>
<td>0.300</td>
</tr>
<tr>
<td>175</td>
<td>-0.010</td>
<td>0.150</td>
</tr>
<tr>
<td>200</td>
<td>0.000</td>
<td>0.050</td>
</tr>
<tr>
<td>225</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>250</td>
<td>0.010</td>
<td>0.002</td>
</tr>
<tr>
<td>260</td>
<td>0.010</td>
<td>0.002</td>
</tr>
<tr>
<td>275</td>
<td>0.015</td>
<td>0.004</td>
</tr>
<tr>
<td>300</td>
<td>0.010</td>
<td>0.006</td>
</tr>
<tr>
<td>325</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>350</td>
<td>0.020</td>
<td>-0.050</td>
</tr>
<tr>
<td>375</td>
<td>0.005</td>
<td>-0.100</td>
</tr>
<tr>
<td>400</td>
<td>0.080</td>
<td>-0.200</td>
</tr>
<tr>
<td>425</td>
<td>0.004</td>
<td>-0.100</td>
</tr>
<tr>
<td>450</td>
<td>0.010</td>
<td>0.005</td>
</tr>
<tr>
<td>475</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>500</td>
<td>0.001</td>
<td>0.005</td>
</tr>
<tr>
<td>520</td>
<td>-0.100</td>
<td>0.500</td>
</tr>
</tbody>
</table>
Table 5
Statistics on Asset Weights

<table>
<thead>
<tr>
<th>Period</th>
<th>Asset Class</th>
<th>Minimum</th>
<th>Median</th>
<th>Mean</th>
<th>Maximum</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Given $\eta = 0.75$ and $k = 0.0001$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Years</td>
<td>stock</td>
<td>-0.7057</td>
<td>-0.0421</td>
<td>-0.1203</td>
<td>0.0061</td>
<td>0.1672</td>
</tr>
<tr>
<td>10 Years</td>
<td>stock</td>
<td>-0.7108</td>
<td>-0.0034</td>
<td>-0.0625</td>
<td>0.1083</td>
<td>0.1636</td>
</tr>
<tr>
<td>5 Years</td>
<td>long term bond</td>
<td>-0.0728</td>
<td>0.1792</td>
<td>0.2853</td>
<td>1.5888</td>
<td>0.3434</td>
</tr>
<tr>
<td>10 Years</td>
<td>long term bond</td>
<td>-0.2357</td>
<td>-0.0154</td>
<td>0.1272</td>
<td>1.6067</td>
<td>0.3655</td>
</tr>
<tr>
<td>5 Years</td>
<td>cash</td>
<td>0.1132</td>
<td>0.8697</td>
<td>0.8350</td>
<td>1.0668</td>
<td>0.1809</td>
</tr>
<tr>
<td>10 Years</td>
<td>cash</td>
<td>0.1031</td>
<td>1.0175</td>
<td>0.9354</td>
<td>1.1399</td>
<td>0.2076</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Given $\eta = 1.00$ and $k = 0.0001$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Years</td>
<td>stock</td>
<td>-3.9395</td>
<td>-0.1611</td>
<td>-0.6277</td>
<td>0.0024</td>
<td>0.9193</td>
</tr>
<tr>
<td>10 Years</td>
<td>stock</td>
<td>-4.0801</td>
<td>-0.0076</td>
<td>-0.3197</td>
<td>0.0950</td>
<td>0.7961</td>
</tr>
<tr>
<td>5 Years</td>
<td>long term bond</td>
<td>-0.0489</td>
<td>0.7039</td>
<td>1.4995</td>
<td>8.8692</td>
<td>1.8794</td>
</tr>
<tr>
<td>10 Years</td>
<td>long term bond</td>
<td>-0.2082</td>
<td>0.0147</td>
<td>0.7282</td>
<td>8.8981</td>
<td>1.6783</td>
</tr>
<tr>
<td>5 Years</td>
<td>cash</td>
<td>-3.9653</td>
<td>0.4364</td>
<td>0.1281</td>
<td>1.0465</td>
<td>0.9769</td>
</tr>
<tr>
<td>10 Years</td>
<td>cash</td>
<td>-3.9161</td>
<td>0.9929</td>
<td>0.5915</td>
<td>1.1188</td>
<td>0.8932</td>
</tr>
</tbody>
</table>
The funding ratio volatility and the trading activity volatility increase with the difference between the current funding ratio and the target funding ratio. Notice that the volatility of trading volatilities, fund levels, and funding ratios when $\eta = 1$ are greater than those $\eta = 0.75$. This is reasonable because the larger the difference is, the more the assets have to be increased and thus the more aggressive the trading must be. Furthermore, we observe that shorter evaluation periods result in higher volatilities. A possible explanation is that shorter evaluation periods make the optimal trading strategies more sensitive to financial markets because the plan manager has a shorter time to achieve the goal.

5 Summary and Closing Comments

Stochastic control is potentially a helpful tool for managing pension plans. It represents a significant improvement over the one-period approach traditionally used by plan managers because it can explicitly consider the inter-period dynamics and aim at long-term rather than short-term optimality. Furthermore, dynamic control models can simultaneously handle plan funding policies and investment decisions.

Our model is the most comprehensive one so far as it combines the merits from different models. The Haberman and Sung (1994) approach is used to develop our objective function, i.e., we consider the contribution risk (the stability of contributions) and the solvency risk (the security of funds). The Brennan, Schwartz, and Lagnado (1997) model is used to enlarge the set of investable assets so that it contains both risk-free and risky assets. For liabilities, we employ the stochastic simulations in Chang (1999, 2002) that explicitly characterize the plan members. These allow us to derive a system of differential equations, for which the solution represents optimal funding policies and asset allocations. We then apply the theoretical model to an actual Taiwanese pension plan and numerically obtain optimal solutions.

There are three areas for further research:

- One may add short sale constraints into our model because our optimal strategies usually involve certain amount of short sales. Most pension funds, however, are not allowed to engage in short sales because of the associated downside risk;

- One may want to include transaction costs. Note that high transaction costs may reduce the relative advantage of active trading.
to passive trading, which might result in different optimal trading strategies; and finally,

- Include optimal hedging policies.

References


### Table A1
Simplified Male Service Table
Survival Probabilities

<table>
<thead>
<tr>
<th>x</th>
<th>( p_x^{(T)} )</th>
<th>x</th>
<th>( p_x^{(T)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>0.860048</td>
<td>38</td>
<td>0.937401</td>
</tr>
<tr>
<td>16</td>
<td>0.859854</td>
<td>39</td>
<td>0.937279</td>
</tr>
<tr>
<td>17</td>
<td>0.859326</td>
<td>40</td>
<td>0.983539</td>
</tr>
<tr>
<td>18</td>
<td>0.859181</td>
<td>41</td>
<td>0.983369</td>
</tr>
<tr>
<td>19</td>
<td>0.859278</td>
<td>42</td>
<td>0.983176</td>
</tr>
<tr>
<td>20</td>
<td>0.801437</td>
<td>43</td>
<td>0.982957</td>
</tr>
<tr>
<td>21</td>
<td>0.801529</td>
<td>44</td>
<td>0.982763</td>
</tr>
<tr>
<td>22</td>
<td>0.801620</td>
<td>45</td>
<td>0.982559</td>
</tr>
<tr>
<td>23</td>
<td>0.801712</td>
<td>46</td>
<td>0.982352</td>
</tr>
<tr>
<td>24</td>
<td>0.801708</td>
<td>47</td>
<td>0.982149</td>
</tr>
<tr>
<td>25</td>
<td>0.881447</td>
<td>48</td>
<td>0.981952</td>
</tr>
<tr>
<td>26</td>
<td>0.881441</td>
<td>49</td>
<td>0.981673</td>
</tr>
<tr>
<td>27</td>
<td>0.881436</td>
<td>50</td>
<td>0.990293</td>
</tr>
<tr>
<td>28</td>
<td>0.881438</td>
<td>51</td>
<td>0.989965</td>
</tr>
<tr>
<td>29</td>
<td>0.881445</td>
<td>52</td>
<td>0.989596</td>
</tr>
<tr>
<td>30</td>
<td>0.937983</td>
<td>53</td>
<td>0.989173</td>
</tr>
<tr>
<td>31</td>
<td>0.937946</td>
<td>54</td>
<td>0.988972</td>
</tr>
<tr>
<td>32</td>
<td>0.937987</td>
<td>55</td>
<td>0.988738</td>
</tr>
<tr>
<td>33</td>
<td>0.937804</td>
<td>56</td>
<td>0.988465</td>
</tr>
<tr>
<td>34</td>
<td>0.937741</td>
<td>57</td>
<td>0.988150</td>
</tr>
<tr>
<td>35</td>
<td>0.937668</td>
<td>58</td>
<td>0.987794</td>
</tr>
<tr>
<td>36</td>
<td>0.937585</td>
<td>59</td>
<td>0.987288</td>
</tr>
<tr>
<td>37</td>
<td>0.937496</td>
<td>60</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: \( p_x^{(T)} \) = Probability a male plan member age \( x \) remains in the plan at age \( x + 1 \).
Table A2

**Simplified Female Service Table**

<table>
<thead>
<tr>
<th>$x$</th>
<th>$p_x^{(τ)}$</th>
<th>$x$</th>
<th>$p_x^{(τ)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>0.860180</td>
<td>38</td>
<td>0.938062</td>
</tr>
<tr>
<td>16</td>
<td>0.860134</td>
<td>39</td>
<td>0.938021</td>
</tr>
<tr>
<td>17</td>
<td>0.860081</td>
<td>40</td>
<td>0.984412</td>
</tr>
<tr>
<td>18</td>
<td>0.860059</td>
<td>41</td>
<td>0.984340</td>
</tr>
<tr>
<td>19</td>
<td>0.860054</td>
<td>42</td>
<td>0.984258</td>
</tr>
<tr>
<td>20</td>
<td>0.802068</td>
<td>43</td>
<td>0.984166</td>
</tr>
<tr>
<td>21</td>
<td>0.802065</td>
<td>44</td>
<td>0.984053</td>
</tr>
<tr>
<td>22</td>
<td>0.802064</td>
<td>45</td>
<td>0.983935</td>
</tr>
<tr>
<td>23</td>
<td>0.802063</td>
<td>46</td>
<td>0.983814</td>
</tr>
<tr>
<td>24</td>
<td>0.802061</td>
<td>47</td>
<td>0.983691</td>
</tr>
<tr>
<td>25</td>
<td>0.881830</td>
<td>48</td>
<td>0.983571</td>
</tr>
<tr>
<td>26</td>
<td>0.881818</td>
<td>49</td>
<td>0.983423</td>
</tr>
<tr>
<td>27</td>
<td>0.881800</td>
<td>50</td>
<td>0.992203</td>
</tr>
<tr>
<td>28</td>
<td>0.881781</td>
<td>51</td>
<td>0.992035</td>
</tr>
<tr>
<td>29</td>
<td>0.881759</td>
<td>52</td>
<td>0.991851</td>
</tr>
<tr>
<td>30</td>
<td>0.938305</td>
<td>53</td>
<td>0.991652</td>
</tr>
<tr>
<td>31</td>
<td>0.938279</td>
<td>54</td>
<td>0.991392</td>
</tr>
<tr>
<td>32</td>
<td>0.938258</td>
<td>55</td>
<td>0.991124</td>
</tr>
</tbody>
</table>

*Note: $p_x^{(τ)}$ = Probability a female plan member age $x$ remains in the plan at age $x + 1$.*

Table A3

**Basic Statistics on New Entrants**

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>Number of New Entrants</th>
<th>Average Annual Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 19</td>
<td>82</td>
<td>23,356</td>
</tr>
<tr>
<td>20 24</td>
<td>163</td>
<td>27,660</td>
</tr>
<tr>
<td>25 29</td>
<td>273</td>
<td>38,404</td>
</tr>
<tr>
<td>30 34</td>
<td>88</td>
<td>38,718</td>
</tr>
<tr>
<td>35 39</td>
<td>17</td>
<td>46,297</td>
</tr>
<tr>
<td>40 44</td>
<td>7</td>
<td>43,305</td>
</tr>
<tr>
<td>45 49</td>
<td>4</td>
<td>36,053</td>
</tr>
</tbody>
</table>
Model Risks and Surplus Management Under a Stochastic Interest Rate Process

Jennifer L. Wang* and Rachel J. Huang†

Abstract‡

This paper uses simulations to explore the effects of incorrectly identifying the underlying interest rate process on assets, liabilities, and surplus levels. We show that mismodeling the interest rate (called model risk) could not only lead to a misstatement of the company's surplus, but could also cause a mismatch between the company's assets and liabilities. Our simulations demonstrate that three aspects of interest rates affect model risk: (i) volatility, (ii) level of long-term interest rate, and (iii) the speed at which the drift rate adjusts. We conclude that asset-liability managers should not ignore the impact of the model risks, regardless of the length of their planning horizon.

Key words and phrases: asset and liability management, immunization strategy, interest rate risk, model risk

*Jennifer L. Wang, Ph.D., is associate professor of risk management and insurance and deputy director of the Insurance Research and Education Center at the College of Commerce, National Chengchi University in Taiwan. She received a Ph.D in risk management and insurance from Temple University. Her research interests are in pensions, annuities, insurance finance, and insurance accounting. She has published many papers in various international journals including Journal of Risk and Insurance, Journal of Insurance Issues, and Benefits Quarterly.

Dr. Wang’s address is: Department of Risk Management and Insurance, National Chengchi University, 64, Sec. 2, Chihnan Rd., Taipei, 116, TAIWAN. Internet address: jenwang@nccu.edu.tw

†Rachel J. Huang is an instructor of finance at Ming Chuan University and a Ph.D. student in finance at National Taiwan University.

Ms. Huang’s address is: Department of Finance, National Taiwan University, 50, Lane 144, Sec. 4, Keelung Rd., Taipei, 106, TAIWAN. Internet address: rachelhuang@mba.ntu.edu.tw

‡The authors gratefully acknowledge the helpful comments of two anonymous referees and the editor of this journal and Dr. Larry Y. R. Tzeng and seminar participants at the 2001 Risk Management Seminar at National Chengchi University. We deeply appreciate Zn-Hong Chen's research assistance and the financial support from National Science Council of Taiwan.
1 Introduction

A serious problem insurance companies face is the problem of interest rate fluctuations on their assets and liabilities, the so-called C-3 risk by actuaries. Simply put, if assets have a longer maturity date than liabilities, a rise in interest rates will lead to a decrease in the net value of the insurer, while a fall in interest rates will lead to an increase in the net value of the insurer.

To deal with this problem, Redington (1952) introduced a so-called immunization strategy of setting the duration of assets equal to the asset/liability ratio times the duration of liabilities. Redington's approach is now a standard technique used by many authors, including Grove (1974), Bierwag (1987) and Reitano (1992), for immunizing the surplus of an insurance company against interest rate risk. See Panjer (1998, Chapter 3) for a detailed review of the actuarial approach to immunization.

Bellhouse and Panjer (1981), Beekman and Fuelling (1990), Frees (1990), Norberg (1995), and Lai and Frees (1995) have explored the impact of stochastic interest rates on the reserves of life insurance. On the other hand, Briys and Varenne (1997) and Tzeng, Wang, and Soo (2000) have extended the traditional duration approach to address the case where interest rates follow a stochastic process. Tzeng, Wang, and Soo (2000) show that, with certain adjustments, the classical immunization strategy still can be used for surplus management.

Although this line of research has provided some insightful strategies for asset-liability management of insurance companies, most papers focus on the change in interest rates and overlook the cost of mismodeling the interest rate process itself.¹

Though many models of the stochastic behavior of interest rates have been proposed, two models are most popular: Vasicek (1977) and Cox, Ingersoll, and Ross (1985). The Vasicek model assumes the interest rate process is a mean-reverting process with constant volatility. The Cox, Ingersoll, and Ross model assumes the interest rate process is a mean-reverting process but with volatility that is proportional to the level of the interest rate. Other models have been proposed by Dothan and Feldman (1986); Ho and Lee (1986); Chan et al., (1992); and Heath, Jarrow, and Morton (1992).

¹In practice, surplus managers are interested mostly in comparing how surplus levels change as strategies change. Although incorporating a stochastic interest model may not influence the decision in choosing an investment strategy, it certainly generates more accurate asset allocation in terms of immunization of surplus.
This paper considers a hypothetical insurance company and uses simulations and the Vasicek (1977) and Cox, Ingersoll, and Ross (1985) models\(^2\) to measure the cost of misidentifying the interest rate model (i.e., the model risk) in two ways: (i) miscalculating the company's value, and (ii) mismatching the company's assets and liabilities. The paper is organized as follows: Section 2 describes some of the properties of the two interest rate models. Section 3 describes the relevant aspects of the hypothetical insurance company. Section 4 contains the results of the simulations.

### 2 The Vasicek and Cox, Ingersoll, and Ross Interest Rate Models

Although many alternative processes\(^3\) have been suggested for modeling interest-rate behaviors, only a few of them have a closed-form solution for the price of a zero-coupon bond. Among these, Vasicek (1977) and Cox, Ingersoll, and Ross (1985) are most commonly used. Vasicek (1977) models the interest rate process, \(r_t\), as

\[
dr_t = a_V (b_V - r_t) dt + \sigma_V dz,
\]

where \(a_V\), \(b_V\), and \(\sigma_V\) are constants and \(dz\) follows a standard Brownian motion. The term \(a_V (b_V - r_t)\) is called the drift rate, and \(\sigma_V\) is the standard deviation of the interest rate process.

Vasicek (1977) solves equation (1) and shows that the current price of a one-dollar zero-coupon bond maturing in \(t\) periods, \(P(t)\),

\[
P(t) = P_V(t) = \alpha_V(t) \exp(-\beta_V(t)t),
\]

\(^2\)Although these two models are most commonly used interest rate models, they suffer certain limitations. Sometimes, the surplus manager would like to replicate the diverse nature of the yield curve. Neither the Vasicek nor the Cox, Ingersoll, and Ross model allows the yield curve to change from a positively sloped yield curve to a negatively sloped yield curve.

\(^3\)Interest rate models such as those of Vasicek (1977); Cox, Ingersoll, and Ross (1985); Dothan and Feldman (1986); Ho and Lee (1986); Chan et al., (1992); and Heath, Jarrow, and Morton (1992) can be chosen by insurance companies for their own management purposes.
where \( r \) is the current level of interest rates,

\[
\beta_V(t) = \frac{1 - \exp(-a_V t)}{a_V}, \quad \text{and} \quad \alpha_V(t) = \exp \left( \frac{(\beta_V(t) - t)(a_V^2 b_V - 0.5 \sigma_V^2) - \sigma_V^2 \beta^2(t)}{a_V^2} \right).
\]

Under the Cox, Ingersoll, and Ross (1985), the interest rate process, \( r_t \), is modeled as

\[
dr_t = a_I (b_I - r_t) dt + \sigma_I \sqrt{r_t} dz
\]

where \( a_I, b_I, \) and \( \sigma_I \) are constants and \( dz \) follows a standard Brownian motion. Here again, the drift rate is \( a_I (b_I - r_t) \). The standard deviation, however, is now \( \sigma_I \sqrt{r_t} \). Cox, Ingersoll, and Ross (1985) solve equation (5) and show that

\[
P(t) = P_I(t) = \alpha_I(t) \exp(-\beta_I(t)r),
\]

where \( r \) is the current level of interest rates,

\[
y_I^2 = a_I^2 + 2 \sigma_I^2,
\]

\[
\alpha_I(t) = \left( \frac{2 y_I e^{t(a_I + y_I)/2}}{(y_I + a_I)(e^{ty_I} - 1) + 2 y_I} \right)^{2a_I b_I / \sigma_I^2}, \quad \text{and}
\]

\[
\beta_I(t) = \frac{2 (e^{ty_I} - 1)}{(y_I + a_I)(e^{ty_I} - 1) + 2 y_I}.
\]

It is important to recognize that, though they have the same functional form for the drift rate, the Vasicek (1977) and Cox, Ingersoll, and Ross (1985) models have different assumptions for interest-rate variations. Vasicek (1977) assumes a constant variation in the interest rate in each period, while Cox, Ingersoll, and Ross (1985) assume that the variation in the interest rate in a period is proportional to the square root of the interest rate in the period.

3 Model of the Hypothetical Insurance Company

Suppose a hypothetical insurance company has a current balance sheet (at the start of period 1) as shown in Table 1.
Table 1

Balance Sheet of Hypothetical Insurance Company

<table>
<thead>
<tr>
<th>Assets</th>
<th>Liabilities</th>
<th>Surplus</th>
</tr>
</thead>
<tbody>
<tr>
<td>$9,045,110</td>
<td>$8,545,110</td>
<td>$500,000</td>
</tr>
</tbody>
</table>

Let $R(t)$ and $C(t)$ denote the cash inflows and cash outflows, respectively, of the hypothetical insurance company $t$ periods in the future. Following the approach proposed by Tzeng, Wang, and Soo (2000), the assets and liabilities of an insurance company, $A$ and $L$, satisfy the following equations:

\[ A = \sum_{t=1}^{n} R(t)P^A(t), \quad \text{and} \]
\[ L = \sum_{t=1}^{n} C(t)P^L(t), \]

where $P^A(t)$ and $P^L(t)$ are the current price of a one-dollar zero-coupon bond maturing in $t$ periods based on the interest rate process followed by the assets and the liabilities, respectively. The surplus of insurance company, $S$, is then equal to

\[ S = A - L = \sum_{t=1}^{n} R(t)P^A(t) - \sum_{t=1}^{n} C(t)P^L(t). \]

For simplicity, we further assume the company is a run-off case,\(^4\) and the liabilities\(^5\) are to be paid out over fifteen years, as shown in Table 2. This means that the present value, using discount rate $P^L(t)$, of cash outflows would be equal to the total liability. On the other hand, the

\(^4\)A run-off case means that the company would not consider or implement any new business line over fifteen years.

\(^5\)In practice, an insurance company’s liability schedule is often hard to predict. Becker (1988) discusses the difficulty of correctly measuring the value of the liability of an insurance company. Recent research findings on the effective duration of insurance liabilities—see, for example, Babbel, Merrill, and Planning (1997) and Briys and Varenne (1997)—can help to make more accurate predictions. We have made the liability schedule independent of the interest rate in order to concentrate on the analysis of model risk. In practice, however, interest rate changes do have a significant impact on lapse rates, policy loans, and surrenders, as documented in Briys and Varenne (1997) and hence on the duration of liabilities.
present value of cash inflows that the company pursues should satisfy the balance sheet condition, i.e.,

$$\sum_{t=1}^{15} [\alpha_V(t) \exp(-\beta_V(t)r)]R(t) = 9,045,110. \quad (13)$$

### Table 2

| Liabilities (Cash Outflows) of Hypothetical Insurance Company |
|---|---|---|---|
| $C(t)$ | $C(t)$ | $C(t)$ |
| $\$591,500$ | $\$824,600$ | $\$1,087,400$ |
| $\$633,700$ | $\$871,300$ | $\$1,133,500$ |
| $\$677,400$ | $\$932,700$ | $\$1,187,300$ |
| $\$723,500$ | $\$984,200$ | $\$1,212,600$ |
| $\$775,800$ | $\$1,036,500$ | $\$1,253,800$ |

Let $r_t^A$ and $r_t^L$ denote the rate of return on assets and liabilities, respectively. Assume the insurance policies are interest-rate sensitive, and the company always maintains its interest rate for valuing liabilities as a fixed proportion of its rate for valuing assets. This means that

$$r_t^L = kr_t^A$$

where $k$ is a positive constant. If the interest rate of assets follow Vasicek's (1977) model, i.e., $r_t^A = r_t$, then the interest rate for valuing liabilities would satisfy $r_t^L = kr_t$, i.e.,

$$dr_t^L = \alpha_V(kb_V - r_t^L) dt + \sqrt{k} \sigma_V dz. \quad (14)$$

This means that the long run level and the volatility of the liability rate of return are proportional to those of the asset rate of return. The adjustment speed for the liability rate of return to its long-term level is the same as that for the asset rate of return.

On the other hand, if the asset rate of return follows Cox, Ingersoll, and Ross's (1985) model, we have

$$dr_t^L = a_I(kb_I - r_t^L) dt + \sqrt{k} \sigma_I \sqrt{r_t^L} dz. \quad (15)$$

Here, the long-term level of the liability interest rate is still $k$ times that of the asset return, as in Vasicek's model. In the Cox, Ingersoll, and
Ross (1985) model, however, the standard deviation of the liability rate of return is $\sqrt{k\sigma_1^2 r_t^L}$.

Assume that the current interest rate of asset is $r = 6\%$. The parameters of the Vasicek (1977) and Cox, Ingersoll, and Ross (1985) models are obtained from Chan et al., (1992), who estimate them from U.S. Treasury yield data from June 1964 to December 1989.\(^6\) Thus, we can generate the parameters as follows: $a_V = 0.1779$, $b_V = 0.0866$, $\sigma_V = 0.02$; and $a_I = 0.2339$, $b_I = 0.0808$, $\sigma_I = 0.0854$.\(^7\) We assume $k = 80\%$, so that the adjustment speed, long-term level, and standard error of $r_t^L$ are 0.1779, 0.0693, and 0.0160, respectively.

4 The Immunization Equations

Let us suppose that the hypothetical company manages surplus by assuming that the interest-rate process follows Vasicek's (1977) model with the parameters given at the end of Section 3. We further assume, however, that interest rates actually follow the Cox, Ingersoll, and Ross (1985) model with the parameters given at the end of Section 3. The deviation of expected surplus from the actual surplus is referred to as mismodeling cost.

Tzeng, Wang, and Soo (2000) show that, if a closed-form solution of $P^A(t)$ and $P^L(t)$ exists, an immunization strategy can be generated by

$$\frac{dS}{dr} = 0,$$

where $r$ is the spot rate. For this hypothetical company, the above immunization strategy can be expressed as

$$\frac{dS}{dr} = \sum_{t=1}^{n} R(t) \frac{dP^A(t)}{dr} - \sum_{t=1}^{n} C(t) \frac{dP^L(t)}{dr} = 0. \quad (16)$$

---

\(^6\)The proxy of the short-term interest rate in their model is the Treasury yield, which is generated from Fama (1984) and maintained by the Center for Research in Security Prices (CRSP). The one-month yield is the average of the bid-and-ask price for Treasury bills and is normalized as a standard month with 30.4 days. It should be recognized that, besides the prices of short-term bonds, the prices of long-term bonds and the price of interest options could provide additional information for interest rate volatility especially when more sophisticated models [such as Heath, Jarrow and Morton (1992)] are adopted.

\(^7\)In Chan et al.'s (1992) Table III, the expectation of the short-term interest rate under Vasicek's setting is $E[r_{t+1} - r_t] = 0.0154 - 0.1779r_t$. Therefore, we have $a_V = 0.1779$, $b_V = 0.0866$. By the same token, under the Cox, Ingersoll, and Ross model, $E[r_{t+1} - r_t] = 0.0189 - 0.2339r_t$. Thus, we have $a_I = 0.2339$, and $b_I = 0.0189 / 0.2339 = 0.0808$. 

Substituting the cash outflows and parameters chosen for the Vasicek (1977) model, the above equation is equivalent to

\[ \sum_{t=1}^{15} \alpha_V(t) \beta_V(t) \exp(-\beta_V(t) r) R(t) = 26,049,488. \quad (17) \]

From equations (2) to (9), it is obvious that the immunization strategies under Vasicek (1977) and Cox, Ingersoll, and Ross (1985) can be substantially different. Moreover, given the same set of cash in-flows and out-flows, the value of a company's surplus depends on the interest rate model used. Thus, the model risks associated with surplus management actually stems from two sources: misevaluation and mismatch. Misevaluation of the company's surplus refers to incorrectly calculating the surplus due to mismodeling the interest rates, i.e., incorrectly identifying the underlying interest rate model. Mismatch refers to the lack of immunization of a company's assets and liabilities due to mismodeling interest rates.

In practice, insurance companies must satisfy certain statutory regulations such as minimum solvency margins and restrictions against borrowing. If there is a minimum solvency margin of \( M(t) \) in period \( t \) and the insurance company can reinvest its net cash flows in each period in the same investment portfolio, then the solvency constraints for the insurance company can be expressed as

\[ \sum_{t=1}^{j} (R(t) - C(t)) \frac{P^A(t)}{P^V(j)} \geq M(j), \quad j = 1, \ldots, 15, \quad (18) \]

and \( R(t) \geq 0 \) for \( t = 1, \ldots, 15 \). There may exist multiple solutions that satisfy equations (13), (17), and (18). To keep all the comparisons on an equal basis, we choose a maximum-convexity strategy\(^8\) as the optimal strategy for the insurance company. If we assume that the solvency margin, \( M(t) \), is $10,000, the company's optimal immunization strategy can be modeled as

\(^8\)Douglas (1990) and Christensen and Sorensen (1994) suggested that if asset-liability managers expect the volatility of interest rates to be greater than what appears in the term-structure, then the company's optimal objective would be to maximize its convexity of the surplus subject to the zero surplus duration and its budget constraints. Gagnon and Johnson (1994) and Barber and Copper (1997), however, have demonstrated that matching the convexities of asset and liability does not always improve the immunization results.
max  \frac{d^2S}{dr^2} = \sum_{t=1}^{15} [\alpha(t) \beta_\nu(t) \exp(-\beta(t)r)]R(t) \\
\text{such that}

\sum_{t=1}^{15} [\alpha(t) \exp(-\beta(t)r)]R(t) = 9,045,110,

\sum_{t=1}^{15} [\alpha(t) \beta_\nu(t) \exp(-\beta(t)r)]R(t) = 26,049,488,

\sum_{t=1}^{15} (R(t) - C(t)) \frac{p_\nu(t)}{p_\nu(j)} \geq 10,000, \quad j = 1, \ldots, 15, \quad \text{and} \quad R(t) \geq 0, \quad t = 1, \ldots, 15.

Notice that when the company's surplus (S), liability schedule (C(t)), and the parameters of the stochastic interest rate processes are given, equations (13), (17), and (18) are all linear functions with respect to R(t). Therefore, equation (19) can be solved by linear programming, and the optimal allocation of cash flows is shown in Table 3.

Table 3
Optimal Income Stream (Cash Inflows)
Of Hypothetical Insurance Company

<table>
<thead>
<tr>
<th>t</th>
<th>R(t)</th>
<th>t</th>
<th>R(t)</th>
<th>t</th>
<th>R(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$5,035,935</td>
<td>6</td>
<td>$0</td>
<td>11</td>
<td>$1,086,624</td>
</tr>
<tr>
<td>2</td>
<td>$0</td>
<td>7</td>
<td>$0</td>
<td>12</td>
<td>$1,132,644</td>
</tr>
<tr>
<td>3</td>
<td>$0</td>
<td>8</td>
<td>$331,756</td>
<td>13</td>
<td>$1,186,887</td>
</tr>
<tr>
<td>4</td>
<td>$0</td>
<td>9</td>
<td>$1,035,246</td>
<td>14</td>
<td>$1,211,622</td>
</tr>
<tr>
<td>5</td>
<td>$0</td>
<td>10</td>
<td>$5,437,539</td>
<td>15</td>
<td>$5,437,539</td>
</tr>
</tbody>
</table>
5 The Results of the Simulation

The simulation is divided into two parts: First, we compare the differences between $\alpha_V$ and $\alpha_I$, and between $\beta_V$ and $\beta_I$. Then we evaluate the cost of mismodeling.

5.1 Differences in Vasicek (1977) and Cox, Ingersoll, and Ross (1985)

As mentioned earlier, the model risks actually result from the differences in the $\alpha$ and $\beta$ terms in the two models. Therefore, it is important that we examine these differences under different parameters values for $a$, $b$, $\sigma$, and $t$ as it will help to identify the severity of the model risks.

Tables A1, A2, A3, and A4 in the appendix display $\alpha_V - \alpha_I$, $(\alpha_V - \alpha_I)/\alpha_V$, $\beta_V - \beta_I$, and $(\beta_V - \beta_I)/\beta_V$, respectively, for $b = 3$, various time periods, and various levels of $a$ and $\sigma$.

Table A1 shows that $|\alpha_V - \alpha_I|$ increases as $\sigma$ increases, but decreases as $a$ increases. In addition, it is important to recognize that $|\alpha_V - \alpha_I|$ approaches zero as $\sigma$ approaches zero because the Vasicek (1979) and Cox, Ingersoll, and Ross (1985) models collapse into the same model when the variance of the interest rate process approaches zero. Table A2 shows that the relative difference, $|(\alpha_V - \alpha_I)/\alpha_V|$, increases as $\sigma$ increases. For large $a$, $|(\alpha_V - \alpha_I)/\alpha_V|$ is an increasing convex function with respect to $t$. For small $a$, however, there is no clear impact pattern on $(\alpha_V - \alpha_I)/\alpha_V$. Table A3 shows $|\beta_V - \beta_I|$ decreases as $a$ increases, but increases as $\sigma$ or $t$ increases. As the same pattern observed in $|\alpha_V - \alpha_I|$, we find $|\beta_V - \beta_I|$ also will approach zero when $\sigma$ is sufficiently small. In Table A4 we find $|(\beta_V - \beta_I)/\beta_V|$ also decreases as $a$ increases, but increases as $\sigma$ or $t$ increases.

Further results obtained by varying $b$, but not reported in these tables, show that:

- $|\alpha_V - \alpha_I|$ decreases as $b$ increases;

- For large $b$, $|(\alpha_V - \alpha_I)/\alpha_V|$ is an increasing convex function with respect to $t$. For small $b$, however, it shows no clear impact pattern; and,

- As expected, $b$ has no impact on $|\beta_V - \beta_I|$ because in both the Vasicek (1977) and Cox, Ingersoll, and Ross (1985) models, $\beta$ is independent of the level of the long-term interest rate $b$.

\footnote{Although the results of this paper depend on model forms in the simulation, they still serve as a case for demonstrating the severity of model risks.}
Based on the results of the simulation, we can conclude that low long-term interest rate levels, high interest rate volatility, or low drift rate momentum increases model risk. We do not have a conclusive finding with respect to an increase in the time horizon. Thus, we would caution financial planners at insurance companies to not ignore the possible effects of model risks, regardless of the length of their planning horizon.

5.2 Costs of Mismodeling

The costs of model risks are measured in two ways: miscalculation of a company's value and mismatch of a company's assets and liabilities. Given the cash outflows and inflows in Tables 2 and 3, we then calculate the values of a company's assets, liabilities, and surplus under the Vasicek (1977) and Cox, Ingersoll, and Ross (1985) models. Table 4 shows the estimated values of a company's assets, liabilities, and surplus for each model. Notice that miscalculation of the surplus value is roughly 5 percent, which is a substantial amount.

<table>
<thead>
<tr>
<th>Table 4</th>
<th>The Cost of Miscalculating the Company's Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assets</td>
</tr>
<tr>
<td>Expected:</td>
<td>$9,045,000</td>
</tr>
<tr>
<td>Actual:</td>
<td>$9,138,000</td>
</tr>
<tr>
<td>Cost:</td>
<td>$93,000</td>
</tr>
<tr>
<td>% Change:</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

Notes: Expected refers to the Vasicek (1977) model; Actual refers to the Cox, Ingersoll, and Ross (1985) model; Cost = Actual – Expected; % Change = Cost / Expected; Numbers are rounded to the nearest $1,000.

To measure the mismatch cost caused by the model risks, we further assume that the current interest rate immediately shifts from \( r = 6\% \) to \( r = i\% \), where \( i = 2, 3, \ldots, 10 \).\(^{10}\) The estimated surplus values under the shift in the interest rate is assumed to be non-stochastic, although our simulation can be applied to both stochastic and non-stochastic changes in interest rates. In practice, company managers may be more concerned with the non-stochastic changes in interest rates in the short run, although they may recognize the underlying stochastic structure of interest rates in the long run. In addition, if the interest rate is allowed to vary within two standard deviations, then a maximum 4 percent shock may be acceptable.
Vasicek (1977) and under Cox, Ingersoll, and Ross (1985) are shown in Table 5.

### Table 5

<table>
<thead>
<tr>
<th>r</th>
<th>Vasicek (1977)</th>
<th>IMMUZ Effect</th>
<th>CIR (1985)</th>
<th>Differences (4)-(2)</th>
<th>% Cost (4)–(5) (5)</th>
<th>% Cost (4)–$25,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>2%</td>
<td>$500,000</td>
<td>100%</td>
<td>$503,000</td>
<td>0.6%</td>
<td>-4.3%</td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td>$500,000</td>
<td>100%</td>
<td>$509,000</td>
<td>1.7%</td>
<td>-3.2%</td>
<td></td>
</tr>
<tr>
<td>4%</td>
<td>$500,000</td>
<td>100%</td>
<td>$513,000</td>
<td>2.8%</td>
<td>-2.3%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>$500,000</td>
<td>100%</td>
<td>$519,000</td>
<td>3.9%</td>
<td>-1.2%</td>
<td></td>
</tr>
<tr>
<td>6%</td>
<td>$500,000</td>
<td>100%</td>
<td>$525,000</td>
<td>5.0%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>7%</td>
<td>$500,000</td>
<td>100%</td>
<td>$531,000</td>
<td>6.2%</td>
<td>+1.0%</td>
<td></td>
</tr>
<tr>
<td>8%</td>
<td>$500,000</td>
<td>100%</td>
<td>$537,000</td>
<td>7.5%</td>
<td>+2.2%</td>
<td></td>
</tr>
<tr>
<td>9%</td>
<td>$500,000</td>
<td>100%</td>
<td>$544,000</td>
<td>8.7%</td>
<td>+3.5%</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>$500,000</td>
<td>100%</td>
<td>$550,000</td>
<td>9.9%</td>
<td>+4.7%</td>
<td></td>
</tr>
</tbody>
</table>

Notes: IMMUZ = Immunization; CIR = Cox, Ingersoll, and Ross; and % Cost = Percentage cost of mismodeling. Numbers are rounded to the nearest $1,000.

Columns (2) and (4) in Table 5 show the estimated surplus values at different interest rates under the processes of Vasicek (1977) and Cox, Ingersoll, and Ross (1985), respectively. Column (3) demonstrates the immunization effect of surplus management if the immunization strategy is derived from Vasicek (1977) and the underlined interest rate follows Vasicek (1977). On the other hand, Columns (5) and (6) demonstrate the percentage difference in the surplus value if the immunization strategy is derived from Vasicek (1977), whereas the underlined interest rate follows Cox, Ingersoll, and Ross (1985).

The results of Table 5 show that mismodeling causes a mismatch of a company’s assets and liabilities and exposes the company’s surplus to interest-rate risk. Although the cost of mismodeling is not as high as mismeasurement in our simulation, a one-percent change in the interest rate could still influence a company’s surplus value by more than one percent. Moreover, all other risks, such as equity, operational, liquidity, etc., remain unchanged in the simulation. The simulation shows that it could cost the company ±5 percent of its surplus purely because of mismodeling.
6 Summary and Conclusions

In practice, asset-liability managers often rely on sophisticated models to develop risk management strategies. The over-reliance on such models may cause unpredictable crises when the real world does not behave according to the models. This paper investigates the impact of interest rate model risks on an insurance company’s surplus using two popular interest-rate models: Vasicek (1977) and Cox, Ingersoll, and Ross (1985).

We find that differences in parameters between Vasicek (1977) and Cox, Ingersoll, and Ross (1985) are higher when the long-term interest-rate level is low, the volatility of the interest rate is high, and the momentum of the drift rate is low. In other words, a low level of the long-term interest rate, high volatility of the interest rate, and low momentum of the drift rate increase the model risks. We do not have a conclusive finding with respect to an increase in the time horizon. Thus, managers in insurance companies should not ignore the possible impact of the model risks whether they are engaged in short-term or long-term financial planning. We further show that the cost of failing to recognize model risks can be extremely high. Because of mismodeling, misevaluation could cause about a 5 percent shock on a company’s surplus. A mismatch of a company’s assets and liabilities also could cause at least a one-percent fluctuation for a one percentage change in the interest rate.

In this paper we focus on estimating the cost of model risk for a yearly adjustment surplus management strategy; thus, the liability schedules of an insurance company are assumed to be independent of interest rate, and the shock of interest rate is a one-time shock. In the real world, however, many factors—such as surrender rate, lapse rate, and policy loan as suggested in Briys and Varenne (1997)—could make a liability schedule sensitive to the path of the interest rate. A dynamic immunization relaxing the above two assumptions could provide further understanding for asset-liability management in future studies.

References


| $t$ | $a = 0.1$ | | | | $a = 0.2$ | | | | $a = 0.3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 0.1$ | $\sigma = 0.2$ | $\sigma = 0.3$ | | $\sigma = 0.1$ | $\sigma = 0.2$ | $\sigma = 0.3$ | | $\sigma = 0.1$ | $\sigma = 0.2$ | $\sigma = 0.3$ | |
| 1 | 1.2E-03 | 1.1E-02 | 3.2E-02 | | 9.3E-04 | 8.4E-03 | 2.4E-02 | | 6.9E-04 | 6.3E-03 | 1.8E-02 | |
| 2 | 5.6E-03 | 5.4E-02 | 1.7E-01 | | 2.5E-03 | 2.4E-02 | 7.5E-02 | | 1.1E-03 | 1.1E-02 | 3.5E-02 | |
| 3 | 8.5E-03 | 9.3E-02 | 3.8E-01 | | 1.8E-03 | 2.1E-02 | 8.5E-02 | | 4.3E-04 | 4.9E-03 | 2.1E-02 | |
| 4 | 7.3E-03 | 1.0E-01 | 7.1E-01 | | 6.7E-04 | 9.9E-03 | 6.5E-02 | | 6.2E-05 | 1.1E-03 | 7.9E-03 | |
| 5 | 4.2E-03 | 9.2E-02 | 1.5E+00 | | 1.5E-04 | 3.4E-03 | 4.2E-02 | | 1.1E-06 | 1.4E-04 | 2.3E-03 | |
| 6 | 1.9E-03 | 7.3E-02 | 4.1E+00 | | 2.0E-05 | 9.1E-04 | 2.6E-02 | | -1.1E-06 | 7.0E-06 | 5.8E-04 | |
| 7 | 6.5E-04 | 5.5E-02 | 1.7E+01 | | 1.6E-06 | 2.1E-04 | 1.7E-02 | | -2.5E-07 | -1.3E-06 | 1.4E-04 | |
| 8 | 1.9E-04 | 4.3E-02 | 1.1E+02 | | -2.5E-08 | 4.4E-05 | 1.1E-02 | | -3.4E-08 | -4.6E-07 | 3.0E-05 | |
| 9 | 4.6E-05 | 3.5E-02 | 1.2E+03 | | -3.1E-08 | 8.4E-06 | 7.8E-03 | | -3.7E-09 | -9.1E-08 | 6.5E-06 | |
| 10 | 9.7E-06 | 3.1E-02 | 2.1E+04 | | -6.2E-09 | 1.5E-06 | 5.9E-03 | | -3.4E-10 | -1.4E-08 | 1.4E-06 | |
| 11 | 1.8E-06 | 2.9E-02 | 6.0E+05 | | -8.6E-10 | 2.6E-07 | 4.7E-03 | | -2.8E-11 | -2.0E-09 | 2.9E-07 | |
| 12 | 3.1E-07 | 3.1E-02 | 2.7E+07 | | -1.0E-10 | 4.3E-08 | 4.0E-03 | | -2.1E-12 | -2.7E-10 | 6.1E-08 | |
| 13 | 4.6E-08 | 3.6E-02 | 1.8E+09 | | -1.0E-11 | 7.0E-09 | 3.5E-03 | | -1.5E-13 | -3.4E-11 | 1.3E-08 | |
| 14 | 6.4E-09 | 4.6E-02 | 1.9E+11 | | -9.7E-13 | 1.1E-09 | 3.3E-03 | | -1.1E-14 | -4.1E-12 | 2.6E-09 | |
| 15 | 8.0E-10 | 6.5E-02 | 2.8E+13 | | -8.5E-14 | 1.7E-10 | 3.1E-03 | | -7.2E-16 | -4.8E-13 | 5.3E-10 |
Table A2

<table>
<thead>
<tr>
<th>$t$</th>
<th>$\sigma = 0.1$</th>
<th>$\sigma = 0.2$</th>
<th>$\sigma = 0.3$</th>
<th>$\sigma = 0.1$</th>
<th>$\sigma = 0.2$</th>
<th>$\sigma = 0.3$</th>
<th>$\sigma = 0.1$</th>
<th>$\sigma = 0.2$</th>
<th>$\sigma = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.4E-03</td>
<td>1.3E-02</td>
<td>3.5E-02</td>
<td>1.2E-03</td>
<td>1.1E-02</td>
<td>3.0E-02</td>
<td>1.0E-03</td>
<td>9.4E-03</td>
<td>2.6E-02</td>
</tr>
<tr>
<td>2</td>
<td>9.8E-03</td>
<td>8.5E-02</td>
<td>2.2E-01</td>
<td>7.0E-03</td>
<td>6.3E-02</td>
<td>1.7E-01</td>
<td>4.9E-03</td>
<td>4.5E-02</td>
<td>1.2E-01</td>
</tr>
<tr>
<td>3</td>
<td>2.8E-02</td>
<td>2.3E-01</td>
<td>5.3E-01</td>
<td>1.7E-02</td>
<td>1.5E-01</td>
<td>3.8E-01</td>
<td>8.9E-03</td>
<td>8.5E-02</td>
<td>2.5E-01</td>
</tr>
<tr>
<td>4</td>
<td>5.5E-02</td>
<td>4.2E-01</td>
<td>8.0E-01</td>
<td>2.7E-02</td>
<td>2.4E-01</td>
<td>5.9E-01</td>
<td>8.8E-03</td>
<td>1.1E-01</td>
<td>3.6E-01</td>
</tr>
<tr>
<td>5</td>
<td>9.0E-02</td>
<td>6.0E-01</td>
<td>9.4E-01</td>
<td>3.3E-02</td>
<td>3.3E-01</td>
<td>7.6E-01</td>
<td>1.4E-03</td>
<td>9.4E-02</td>
<td>4.5E-01</td>
</tr>
<tr>
<td>6</td>
<td>1.3E-01</td>
<td>7.6E-01</td>
<td>9.9E-01</td>
<td>3.2E-02</td>
<td>3.9E-01</td>
<td>8.7E-01</td>
<td>-1.6E-02</td>
<td>3.9E-02</td>
<td>5.2E-01</td>
</tr>
<tr>
<td>7</td>
<td>1.7E-01</td>
<td>8.7E-01</td>
<td>1.0E+00</td>
<td>2.0E-02</td>
<td>4.5E-01</td>
<td>9.4E-01</td>
<td>-4.4E-02</td>
<td>-6.5E-02</td>
<td>5.7E-01</td>
</tr>
<tr>
<td>8</td>
<td>2.0E-01</td>
<td>9.4E-01</td>
<td>1.0E+00</td>
<td>-3.1E-03</td>
<td>4.8E-01</td>
<td>9.7E-01</td>
<td>-8.4E-02</td>
<td>-2.3E-01</td>
<td>6.1E-01</td>
</tr>
<tr>
<td>9</td>
<td>2.3E-01</td>
<td>9.7E-01</td>
<td>1.0E+00</td>
<td>-4.1E-02</td>
<td>5.1E-01</td>
<td>9.9E-01</td>
<td>-1.4E-01</td>
<td>-4.7E-01</td>
<td>6.5E-01</td>
</tr>
<tr>
<td>10</td>
<td>2.6E-01</td>
<td>9.9E-01</td>
<td>1.0E+00</td>
<td>-9.5E-02</td>
<td>5.2E-01</td>
<td>1.0E+00</td>
<td>-2.0E-01</td>
<td>-8.1E-01</td>
<td>6.8E-01</td>
</tr>
<tr>
<td>11</td>
<td>2.8E-01</td>
<td>1.0E+00</td>
<td>1.0E+00</td>
<td>-1.7E-01</td>
<td>5.2E-01</td>
<td>1.0E+00</td>
<td>-2.8E-01</td>
<td>-1.3E+00</td>
<td>7.1E-01</td>
</tr>
<tr>
<td>12</td>
<td>2.9E-01</td>
<td>1.0E+00</td>
<td>1.0E+00</td>
<td>-2.6E-01</td>
<td>5.2E-01</td>
<td>1.0E+00</td>
<td>-3.7E-01</td>
<td>-1.9E+00</td>
<td>7.3E-01</td>
</tr>
<tr>
<td>13</td>
<td>2.9E-01</td>
<td>1.0E+00</td>
<td>1.0E+00</td>
<td>-3.8E-01</td>
<td>5.1E-01</td>
<td>1.0E+00</td>
<td>-4.7E-01</td>
<td>-2.8E+00</td>
<td>7.5E-01</td>
</tr>
<tr>
<td>14</td>
<td>2.8E-01</td>
<td>1.0E+00</td>
<td>1.0E+00</td>
<td>-5.3E-01</td>
<td>5.0E-01</td>
<td>1.0E+00</td>
<td>-6.0E-01</td>
<td>-3.9E+00</td>
<td>7.7E-01</td>
</tr>
<tr>
<td>15</td>
<td>2.6E-01</td>
<td>1.0E+00</td>
<td>1.0E+00</td>
<td>-7.1E-01</td>
<td>4.8E-01</td>
<td>1.0E+00</td>
<td>-7.3E-01</td>
<td>-5.5E+00</td>
<td>7.9E-01</td>
</tr>
</tbody>
</table>
Table A3

The Differences $\beta_V - \beta_I$ for $b = 3$

<table>
<thead>
<tr>
<th>$t$</th>
<th>$a = 0.1$</th>
<th>$a = 0.2$</th>
<th>$a = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma = 0.1$</td>
<td>$\sigma = 0.2$</td>
<td>$\sigma = 0.3$</td>
</tr>
<tr>
<td>1</td>
<td>1.5E-03</td>
<td>1.3E-02</td>
<td>3.6E-02</td>
</tr>
<tr>
<td>2</td>
<td>1.1E-02</td>
<td>9.2E-02</td>
<td>2.3E-01</td>
</tr>
<tr>
<td>3</td>
<td>3.3E-02</td>
<td>2.6E-01</td>
<td>6.0E-01</td>
</tr>
<tr>
<td>4</td>
<td>7.0E-02</td>
<td>5.2E-01</td>
<td>1.1E+00</td>
</tr>
<tr>
<td>5</td>
<td>1.2E-01</td>
<td>8.4E-01</td>
<td>1.6E+00</td>
</tr>
<tr>
<td>6</td>
<td>1.9E-01</td>
<td>1.2E+00</td>
<td>2.1E+00</td>
</tr>
<tr>
<td>7</td>
<td>2.7E-01</td>
<td>1.6E+00</td>
<td>2.6E+00</td>
</tr>
<tr>
<td>8</td>
<td>3.6E-01</td>
<td>2.0E+00</td>
<td>3.1E+00</td>
</tr>
<tr>
<td>9</td>
<td>4.6E-01</td>
<td>2.3E+00</td>
<td>3.5E+00</td>
</tr>
<tr>
<td>10</td>
<td>5.7E-01</td>
<td>2.7E+00</td>
<td>3.9E+00</td>
</tr>
<tr>
<td>11</td>
<td>6.8E-01</td>
<td>3.0E+00</td>
<td>4.2E+00</td>
</tr>
<tr>
<td>12</td>
<td>7.9E-01</td>
<td>3.3E+00</td>
<td>4.5E+00</td>
</tr>
<tr>
<td>13</td>
<td>9.0E-01</td>
<td>3.6E+00</td>
<td>4.8E+00</td>
</tr>
<tr>
<td>14</td>
<td>1.0E+00</td>
<td>3.8E+00</td>
<td>5.1E+00</td>
</tr>
<tr>
<td>15</td>
<td>1.1E+00</td>
<td>4.0E+00</td>
<td>5.3E+00</td>
</tr>
</tbody>
</table>
Table A4
The Relative Differences $(\beta_V - \beta_I) / \beta_V$ for $b = 3$

<table>
<thead>
<tr>
<th>$t$</th>
<th>$\sigma = 0.1$</th>
<th>$\sigma = 0.2$</th>
<th>$\sigma = 0.3$</th>
<th>$\sigma = 0.1$</th>
<th>$\sigma = 0.2$</th>
<th>$\sigma = 0.3$</th>
<th>$\sigma = 0.1$</th>
<th>$\sigma = 0.2$</th>
<th>$\sigma = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$1.6E-03$</td>
<td>$1.4E-02$</td>
<td>$3.8E-02$</td>
<td>$1.5E-03$</td>
<td>$1.3E-02$</td>
<td>$3.6E-02$</td>
<td>$1.4E-03$</td>
<td>$1.3E-02$</td>
<td>$3.4E-02$</td>
</tr>
<tr>
<td>2</td>
<td>$6.0E-03$</td>
<td>$5.1E-02$</td>
<td>$1.3E-01$</td>
<td>$5.4E-03$</td>
<td>$4.6E-02$</td>
<td>$1.2E-01$</td>
<td>$4.9E-03$</td>
<td>$4.2E-02$</td>
<td>$1.1E-01$</td>
</tr>
<tr>
<td>3</td>
<td>$1.3E-02$</td>
<td>$1.0E-01$</td>
<td>$2.3E-01$</td>
<td>$1.1E-02$</td>
<td>$8.9E-02$</td>
<td>$2.1E-01$</td>
<td>$9.5E-03$</td>
<td>$7.8E-02$</td>
<td>$1.8E-01$</td>
</tr>
<tr>
<td>4</td>
<td>$2.1E-02$</td>
<td>$1.6E-01$</td>
<td>$3.3E-01$</td>
<td>$1.8E-02$</td>
<td>$1.3E-01$</td>
<td>$2.9E-01$</td>
<td>$1.5E-02$</td>
<td>$1.1E-01$</td>
<td>$2.5E-01$</td>
</tr>
<tr>
<td>5</td>
<td>$3.1E-02$</td>
<td>$2.1E-01$</td>
<td>$4.1E-01$</td>
<td>$2.5E-02$</td>
<td>$1.8E-01$</td>
<td>$3.5E-01$</td>
<td>$2.0E-02$</td>
<td>$1.4E-01$</td>
<td>$3.0E-01$</td>
</tr>
<tr>
<td>6</td>
<td>$4.2E-02$</td>
<td>$2.7E-01$</td>
<td>$4.7E-01$</td>
<td>$3.2E-02$</td>
<td>$2.1E-01$</td>
<td>$4.0E-01$</td>
<td>$2.4E-02$</td>
<td>$1.7E-01$</td>
<td>$3.4E-01$</td>
</tr>
<tr>
<td>7</td>
<td>$5.4E-02$</td>
<td>$3.1E-01$</td>
<td>$5.2E-01$</td>
<td>$3.9E-02$</td>
<td>$2.5E-01$</td>
<td>$4.4E-01$</td>
<td>$2.9E-02$</td>
<td>$1.9E-01$</td>
<td>$3.6E-01$</td>
</tr>
<tr>
<td>8</td>
<td>$6.6E-02$</td>
<td>$3.6E-01$</td>
<td>$5.6E-01$</td>
<td>$4.6E-02$</td>
<td>$2.8E-01$</td>
<td>$4.7E-01$</td>
<td>$3.2E-02$</td>
<td>$2.1E-01$</td>
<td>$3.8E-01$</td>
</tr>
<tr>
<td>9</td>
<td>$7.8E-02$</td>
<td>$3.9E-01$</td>
<td>$5.9E-01$</td>
<td>$5.2E-02$</td>
<td>$3.0E-01$</td>
<td>$4.9E-01$</td>
<td>$3.6E-02$</td>
<td>$2.2E-01$</td>
<td>$4.0E-01$</td>
</tr>
<tr>
<td>10</td>
<td>$9.0E-02$</td>
<td>$4.2E-01$</td>
<td>$6.1E-01$</td>
<td>$5.8E-02$</td>
<td>$3.2E-01$</td>
<td>$5.1E-01$</td>
<td>$3.8E-02$</td>
<td>$2.4E-01$</td>
<td>$4.1E-01$</td>
</tr>
<tr>
<td>11</td>
<td>$1.0E-01$</td>
<td>$4.5E-01$</td>
<td>$6.3E-01$</td>
<td>$6.4E-02$</td>
<td>$3.3E-01$</td>
<td>$5.2E-01$</td>
<td>$4.1E-02$</td>
<td>$2.4E-01$</td>
<td>$4.2E-01$</td>
</tr>
<tr>
<td>12</td>
<td>$1.1E-01$</td>
<td>$4.7E-01$</td>
<td>$6.5E-01$</td>
<td>$6.9E-02$</td>
<td>$3.5E-01$</td>
<td>$5.3E-01$</td>
<td>$4.3E-02$</td>
<td>$2.5E-01$</td>
<td>$4.2E-01$</td>
</tr>
<tr>
<td>13</td>
<td>$1.2E-01$</td>
<td>$4.9E-01$</td>
<td>$6.6E-01$</td>
<td>$7.3E-02$</td>
<td>$3.6E-01$</td>
<td>$5.4E-01$</td>
<td>$4.4E-02$</td>
<td>$2.5E-01$</td>
<td>$4.3E-01$</td>
</tr>
<tr>
<td>14</td>
<td>$1.3E-01$</td>
<td>$5.1E-01$</td>
<td>$6.7E-01$</td>
<td>$7.7E-02$</td>
<td>$3.6E-01$</td>
<td>$5.4E-01$</td>
<td>$4.5E-02$</td>
<td>$2.6E-01$</td>
<td>$4.3E-01$</td>
</tr>
<tr>
<td>15</td>
<td>$1.4E-01$</td>
<td>$5.2E-01$</td>
<td>$6.8E-01$</td>
<td>$8.0E-02$</td>
<td>$3.7E-01$</td>
<td>$5.5E-01$</td>
<td>$4.6E-02$</td>
<td>$2.6E-01$</td>
<td>$4.3E-01$</td>
</tr>
</tbody>
</table>
Some Comments on the Pricing of an Exotic Excess of Loss Treaty

Jean-François Walhin*

Abstract

This paper uses a multivariate analog of Panjer's algorithm to develop a method for pricing a complex excess of loss treaty. The treaty is such that some layers inure to the benefit of other layers. The structure of this treaty is discussed. Numerical examples are provided.

Key words and phrases: multivariate Panjer's algorithm, paid reinstatements, inuring layers, order of claims

1 Excess of Loss Treaties

1.1 The Basics

The classic collective risk model assumes that an insurer has a portfolio of similar policies that experiences \( N \) claims in a year—any other period, such as a quarter or month, will do. The sizes of the claims are \( X_1, X_2, \ldots, X_N \) and are independent and identically distributed with common distribution function \( F(x) \). The aggregate losses faced by the insurer for a year, \( S \), is then given by

\[
S = X_1 + \cdots + X_N.
\]

One way to manage these losses is through reinsurance. Excess of loss reinsurance is a means to share risks between the insurer and the reinsurer. The insurer always remains liable for the part of the claim below
a given attachment point (deductible) $P$. The reinsurer, on the other hand, pays the excess of each loss above $P$ and up to a limit $P + L$, thus offering some capacity between $P$ and the limit $P + L$. The quantity $L$ is often called the amount of capacity offered. So, for each claim $X_i$, the liability of the excess of loss reinsurer is $R_i$, where

$$R_i = \min(L, \max(0, X_i - P)).$$

In practice, reinsurers use the term $L \times X$ to refer to the contract defined in equation (1). The aggregate claims of the reinsurer is

$$S_R = R_1 + \cdots + R_N.$$  

Sometimes a line of business is protected by several reinsurance treaties such that for each $X_i$, the $j^{th}$ reinsurance treaty pays

$$R_i^{(j)} = \min(L_j, \max(0, X_i - P_j)),$n$$

where $P_{j+1} = P_j + L_j$ for $j = 1, 2, 3, \ldots$. In this case the $j^{th}$ reinsurance treaty is called the $j^{th}$ layer.

When the reinsurer offers the capacity $L k + 1$ times per year, we say there are $k$ reinstatements, $k = 0, 1, \ldots$. The annual liability of the reinsurer is $\min\{(k + 1)L, S_R\}$. Reinstatements may be paid or free. If they are free, then the insurer can use the $k$ reinstatements without payment of a reinstatement premium. If they are paid, however, the insurer has to pay the reinsurer a reinstatement premium each time the insurer uses the whole layer or part of the layer. Reinstatement premiums are usually defined as a percentage of the initial reinsurance premium. Thus, for example, the reinsurer would state: $L \times X$ with three reinstatements payable at 100%. This means that the offered capacity may be used up to four times, and each time it is used the insurer has to pay a reinstatement premium, which is 100% of the original premium prorated for the used capacity.

In order to reduce the aggregate claims paid by the reinsurer (and hence the reinsurance premium charged) the reinsurance treaty may include an annual aggregate deductible, AAD. In the general case where there are $k$ reinstatements and an annual aggregate deductible of $AAD$, the annual liability of the reinsurer is:

$$S^{(AL)} = \min\{(k + 1)L, \max(0, S_R - AAD)\}.$$  

The following practical example is used to illustrate the ideas and terminology mentioned above. Consider the following treaty:

**Treaty 1. An excess of loss treaty with two layers,**
Walhin: Pricing of an Exotic Excess of Loss Treaty

- **Layer 1:** 100 xs 100 with two reinstatements payable at 150% after an annual aggregate deductible of 50. The reinsurance premium is 25.

- **Layer 2:** 300 xs 200 with one reinstatement payable at 100%. The reinsurance premium is 10.

Suppose further that the insurer experiences the following claims under the treaty: 120, 250, 150, 130. The claims are paid as follows:

- **Layer 1:** Within layer 1, the claims for the reinsurer are: 20, 100, 50, 30. The aggregate claims is 200. As there is an annual aggregate deductible, the individual claims within the layer are aggregated (to give 200), and the insurer pays 50 of the 200. The remaining 150 is to be paid by the reinsurer. As 150 = 100 + 50, layer 1 is used completely once with 50 remaining. Fortunately there are two reinstatements, so the remaining 50 is paid by the reinsurer. As the reinstatements are not free, the reinsurer will ask for two reinstatement premiums: the first for the full layer used, i.e., for the full reinstatement premium, which is 150% × 25 = 37.5; and the second for the partial (50/100) layer used, i.e., for 150% × \( \frac{50}{100} \times 25 = 18.75 \).

- **Layer 2:** Within layer 2 the attachment point is 200 per claim, so the claims faced by the reinsurer are 0, 50, 0, 0. The aggregate claims is 50, which will be paid by the reinsurer. As a reinstatement is payable, however, there is a compulsory reinstatement premium: 100% × \( \frac{50}{300} \times 10 = 1.666 \).

These results are summarized below.

<table>
<thead>
<tr>
<th>Reinsurance Premium</th>
<th>Layer 1: 25 + 37.5 + 18.75 = 81.25</th>
<th>Layer 2: 10 + 1.666 = 11.666</th>
</tr>
</thead>
<tbody>
<tr>
<td>Losses Paid</td>
<td>By insurer: 100 + 100 + 100 + 100 + 50 = 450.</td>
<td>By reinsurer for layer 1: 200 - 50 = 150.</td>
</tr>
<tr>
<td></td>
<td>By reinsurer for layer 2: 50.</td>
<td>1.2 The Notion of Inuring</td>
</tr>
</tbody>
</table>

Recently, I have been given the opportunity to price the following excess of loss treaty:
Treaty 2.

- **Layer 1**: $7.5 \times 2.5$ with three reinstatements payable at 100% after an annual aggregate deductible of 10.
- **Layer 2**: $15 \times 2.5$ with three reinstatements payable at 100% after an annual aggregate deductible of 5.
- **Layer 3**: $22.5 \times 2.5$ with two reinstatements payable at 100%.

Notice that in Treaty 1, layer 2 is such that the upper limit of layer 1 is the attachment point of layer 2. Thus, we must apply layer 1 first, i.e., layer 1 has priority over layer 2. In Treaty 2, however, each layer has the same priority, 2.5, which is why I call this an exotic excess of loss treaty. A rule has to be given to assign a priority to each layer, i.e., to know which layer has to pay first, second, and third. The rule is

- Layer 1 inures to the benefit of layer 2; and
- Layers 1 and 2 inure to the benefit of layer 3.

This means that an excess claim must be paid by layer 3 unless layer 1 or layer 2 is able to pay for it, and layer 1 must pay before layer 2.

There are two ways to interpret the term inure:

(i) Each claim, from ground up (i.e., from the first dollar) is reduced by application of the lower layer; and

(ii) The part of each claim within the layer is reduced by application of the lower layer.

Let us analyze these interpretations with a numerical example. Assume the following claims hit Treaty 2: 20, 5, and 25; and that these claim amounts are from ground up. Under interpretation (i), the results are given in Table 1. Now let us change the order of claims hitting Treaty 2 to 5, 25, and 20. The results are given in Table 2. We observe that, though the sum of total payments within the layers has not changed (42.5), the distribution of these payments has changed, making it questionable whether interpretation (i) makes for good actuarial practice.

It is instructive to analyze what happens in case of a loss larger than the 25, i.e., larger than the largest limit (the one of layer 3). The results are given in Table 3. Observe that due to the reductions of the losses, a loss larger than 25 is still paid entirely (except the deductible) by the reinsurance. The total payments now becomes 52.5.
Next we analyze interpretation (ii) in Table 4. Now let us change the order of claims; the results are shown in Table 5. Notice that when the order of claims changes, the total payments within the layers remain the same. It is easier to analyze the treaty on an aggregate basis as is shown in Table 6. Finally, we analyze the case of a loss larger than the 25, i.e. larger than the largest limit (the one of layer 3), in Table 7. Here, even with larger claims, the sum of total payments remains the same.

Given the results and our observations, we use interpretation (ii) as our definition of inuring.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretation (i)</td>
</tr>
<tr>
<td>In layer 1</td>
</tr>
<tr>
<td>Paid by layer 1</td>
</tr>
<tr>
<td>Reduced claims</td>
</tr>
<tr>
<td>In layer 2</td>
</tr>
<tr>
<td>Paid by layer 2</td>
</tr>
<tr>
<td>Reduced claim</td>
</tr>
<tr>
<td>In layer 3</td>
</tr>
<tr>
<td>Paid by layer 3</td>
</tr>
<tr>
<td>Total payment</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>In layer 1</td>
</tr>
<tr>
<td>Paid by layer 1</td>
</tr>
<tr>
<td>Reduced claims</td>
</tr>
<tr>
<td>In layer 2</td>
</tr>
<tr>
<td>Paid by layer 2</td>
</tr>
<tr>
<td>Reduced claim</td>
</tr>
<tr>
<td>In layer 3</td>
</tr>
<tr>
<td>Paid by layer 3</td>
</tr>
<tr>
<td>Total payment</td>
</tr>
</tbody>
</table>

Notes: *Due to the annual aggregate deductible of layer 1; and
**Due to the annual aggregate deductible of layer 2.

2 The General Mathematical Model

2.1 The Aggregate Loss Model

Our analysis will be conducted within the collective risk model. This model essentially states:

- the number of claims, $N$, is a random variable $N$;
- the claim amounts $X_1, X_2, \ldots, X_N$ are independent realizations of a random variable $X$; and
- $X$ and $N$ are independent.
Table 2
Interpretation (i)
With Changed Order of Claims

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>25</td>
<td>20</td>
<td>Total</td>
</tr>
<tr>
<td>In layer 1</td>
<td>2.5</td>
<td>7.5</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>Paid by layer 1</td>
<td>0</td>
<td>0</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>Reduced claims</td>
<td>5</td>
<td>25</td>
<td>12.5</td>
<td></td>
</tr>
<tr>
<td>In layer 2</td>
<td>2.5</td>
<td>15</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Paid by layer 2</td>
<td>0</td>
<td>12.5</td>
<td>10</td>
<td>22.5</td>
</tr>
<tr>
<td>Reduced claim</td>
<td>5</td>
<td>12.5</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>In layer 3</td>
<td>2.5</td>
<td>10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Paid by layer 3</td>
<td>2.5</td>
<td>10</td>
<td>0</td>
<td>12.5</td>
</tr>
<tr>
<td>Total payment</td>
<td></td>
<td></td>
<td></td>
<td>42.5</td>
</tr>
</tbody>
</table>

Table 3
Interpretation (i)
With One Claim Larger than 25

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>35</td>
<td>20</td>
<td>Total</td>
</tr>
<tr>
<td>In layer 1</td>
<td>2.5</td>
<td>7.5</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>Paid by layer 1</td>
<td>0</td>
<td>0</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>Reduced claims</td>
<td>5</td>
<td>35</td>
<td>12.5</td>
<td></td>
</tr>
<tr>
<td>In layer 2</td>
<td>2.5</td>
<td>15</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Paid by layer 2</td>
<td>0</td>
<td>12.5</td>
<td>10</td>
<td>22.5</td>
</tr>
<tr>
<td>Reduced claim</td>
<td>5</td>
<td>22.5</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>In layer 3</td>
<td>2.5</td>
<td>20</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Paid by layer 3</td>
<td>2.5</td>
<td>20</td>
<td>0</td>
<td>22.5</td>
</tr>
<tr>
<td>Total payment</td>
<td></td>
<td></td>
<td></td>
<td>52.5</td>
</tr>
</tbody>
</table>
Walhin: Pricing of an Exotic Excess of Loss Treaty

Table 4
Interpretation (ii)

<table>
<thead>
<tr>
<th>Layer</th>
<th>20</th>
<th>5</th>
<th>25</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 1</td>
<td>7.5</td>
<td>2.5</td>
<td>7.5</td>
<td>17.5</td>
</tr>
<tr>
<td>Paid by layer 1</td>
<td>0</td>
<td>0</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>Layer 2</td>
<td>15</td>
<td>2.5</td>
<td>15</td>
<td>32.5</td>
</tr>
<tr>
<td>Paid by layer 2</td>
<td>10</td>
<td>2.5</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Layer 3</td>
<td>17.5</td>
<td>2.5</td>
<td>22.5</td>
<td>42.5</td>
</tr>
<tr>
<td>Paid by layer 3</td>
<td>17.5</td>
<td>2.5</td>
<td>22.5</td>
<td></td>
</tr>
<tr>
<td>Real payment by layer 1</td>
<td>0</td>
<td>0</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>Real payment by layer 2</td>
<td>10</td>
<td>2.5</td>
<td>7.5</td>
<td>20</td>
</tr>
<tr>
<td>Real payment by layer 3</td>
<td>7.5</td>
<td>0</td>
<td>7.5</td>
<td>15</td>
</tr>
<tr>
<td>Total payment</td>
<td>42.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Real payment by layer 2 = Paid by layer 2 – Real payment by layer 1; and Real payment by layer 3 = Paid by layer 3 – Real payment by layer 2 – Real payment by layer 1.

Table 5
Interpretation (ii)
With Changed Order of Claims

<table>
<thead>
<tr>
<th>Layer</th>
<th>5</th>
<th>25</th>
<th>20</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 1</td>
<td>2.5</td>
<td>7.5</td>
<td>7.5</td>
<td>17.5</td>
</tr>
<tr>
<td>Paid by layer 1</td>
<td>0</td>
<td>0</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>Layer 2</td>
<td>2.5</td>
<td>15</td>
<td>15</td>
<td>32.5</td>
</tr>
<tr>
<td>Paid by layer 2</td>
<td>0</td>
<td>12.5</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Layer 3</td>
<td>2.5</td>
<td>22.5</td>
<td>17.5</td>
<td>42.5</td>
</tr>
<tr>
<td>Paid by layer 3</td>
<td>2.5</td>
<td>22.5</td>
<td>17.5</td>
<td></td>
</tr>
<tr>
<td>Real payment by layer 1</td>
<td>0</td>
<td>0</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>Real payment by layer 2</td>
<td>0</td>
<td>12.5</td>
<td>7.5</td>
<td>20</td>
</tr>
<tr>
<td>Real payment by layer 3</td>
<td>2.5</td>
<td>10</td>
<td>2.5</td>
<td>15</td>
</tr>
<tr>
<td>Total payment</td>
<td>2.5</td>
<td>22.5</td>
<td>17.5</td>
<td>42.5</td>
</tr>
</tbody>
</table>
Table 6
Interpretation (ii)
With Changed Order of Claims
Analyzed on an Aggregate Basis

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>25</th>
<th>20</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>In layer 1</td>
<td>2.5</td>
<td>7.5</td>
<td>7.5</td>
<td>17.5</td>
</tr>
<tr>
<td>In layer 2</td>
<td>2.5</td>
<td>15</td>
<td>15</td>
<td>32.5</td>
</tr>
<tr>
<td>In layer 3</td>
<td>2.5</td>
<td>22.5</td>
<td>17.5</td>
<td>42.5</td>
</tr>
</tbody>
</table>

Payments in layer 1: \(17.5 - 10 = 7.5\).
Payments in layer 2: \(32.5 - 7.5 - 5 = 20\).
Payments in layer 3: \(42.5 - 20 - 7.5 = 15\).

Table 7
Interpretation (ii)
With One Claim Larger than 25

<table>
<thead>
<tr>
<th></th>
<th>20</th>
<th>35</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>In layer 1</td>
<td>7.5</td>
<td>7.5</td>
<td>2.5</td>
<td>17.5</td>
</tr>
<tr>
<td>In layer 2</td>
<td>15</td>
<td>15</td>
<td>2.5</td>
<td>32.5</td>
</tr>
<tr>
<td>In layer 3</td>
<td>17.5</td>
<td>22.5</td>
<td>2.5</td>
<td>42.5</td>
</tr>
</tbody>
</table>
For more on collective risk models, see, for example, Klugman et al., (1998, Chapter 4).

In our model, we assume the number of claims, \( N \), is a Poisson random variable with mean \( \lambda \) and that the claims follow a limited Pareto distribution with distribution function \( F_X(x) \) given by

\[
F_X(x) = \begin{cases} 
0 & x \leq A, \\
(A^{-\alpha} - x^{-\alpha})/(A^{-\alpha} - B^{-\alpha}) & A < x \leq B, \\
1 & x > B
\end{cases}
\]

where \( A, B, \) and \( \alpha \) are non-negative constants. A limited Pareto distribution is used because it is known from the treaty that there are no losses above a certain threshold, and we are interested only in losses above a certain attachment point. Nevertheless, our approach can be used with any distribution.

A useful tool for determining probabilities in the collective risk model is Panjer's algorithm (Panjer, 1981). This algorithm requires that the distribution of \( X \) be of lattice type; therefore, the limited Pareto distribution is made discrete using, for example, the local moment matching method with one moment. [See, for example, Gerber (1982).] This method ensures that the sum of the masses is 1 and that the first moment of the distribution is conserved.

For a given span \( h = (B - A)/m \), it is not difficult to show that the probabilities of the lattice version of \( X \) are given by \( \tilde{f}_X \):

\[
\begin{align*}
\tilde{f}_X(A) &= 1 - \frac{(A + h)^{1-\alpha} - A^{1-\alpha} - (1 - \alpha)hB^{-\alpha}}{h(1 - \alpha)(A^{-\alpha} - B^{-\alpha})}, \\
\tilde{f}_X(A + jh) &= \frac{2(A + jh)^{1-\alpha} - (A + (j - 1)h)^{1-\alpha} - (A + (j + 1)h)^{1-\alpha}}{h(1 - \alpha)(A^{-\alpha} - B^{-\alpha})}, \quad j = 1, \ldots, m - 1, \\
\tilde{f}_X(B) &= 1 - \tilde{f}_X(A) - \tilde{f}_X(A + h) - \cdots - \tilde{f}_X(B - h).
\end{align*}
\]

2.2 An Exotic Excess of Loss Model

Let \( D \) be the common priority of all layers, \( L_j \) be the limit of layer \( j \), \( AAD_j \) be the annual aggregate deductible of layer \( j \), and \( AAL_j \) be the annual aggregate limit of layer \( j \). In a classical excess of loss treaty, one would naturally define the part of the loss \( X_i \) hitting the various layers as: \( R_i^{(1)} \) for the first layer, \( R_i^{(2)} \) for the second layer, and \( R_i^{(3)} \) for
the third layer. We have

\[ R_i^{(1)} = \min(L_1, \max(0, X_i - D)), \]
\[ R_i^{(2)} = \min(L_2, \max(0, X_i - D)), \]
\[ R_i^{(3)} = \min(L_3, \max(0, X_i - D)). \]

In our exotic excess of loss treaty the aggregate claims for each layer is given by

\[ S_1 = \min(AAL_1, \max(0, \sum_{i=1}^{N} R_i^{(1)} - AAD_1)), \]
\[ S_2 = \min(AAL_2, \max(0, \sum_{i=1}^{N} R_i^{(2)} - S_1 - AAD_2)), \]
\[ S_3 = \min(AAL_3, \max(0, \sum_{i=1}^{N} R_i^{(3)} - S_2 - S_1 - AAD_3)). \]

The term \(-S_1\) in equation (6) indicates that layer 1 inures to the benefit of layer 2. Similarly, the term \(-S_2 - S_1\) in equation (7) indicates that layers 1 and 2 inure to the benefit of layer 3.

In order to price Treaty 2, we now need the distributions of \(S_1, S_2,\) and \(S_3\). Now the distribution of \(S_1\) is easy to obtain by applying Panjer’s algorithm to the case where \(N\) is Poisson. The problem is more complicated, however, for \(S_2\) and \(S_3\). Indeed, \(S_1\) and \(R_1^{(2)} + \cdots + R_N^{(2)}\) are correlated, and \(S_1, S_2,\) and \(R_1^{(3)} + \cdots + R_N^{(3)}\) are also correlated. Thus, the joint distribution of \(\sum R_i^{(1)}, \sum R_i^{(2)},\) and \(\sum R_i^{(3)}\) is needed. Fortunately, a multivariate analog of Panjer’s algorithm exists; see Walhin and Paris (2000) or Sundt (1999).

Let

\[ f^{(S)}(s_1, s_2, s_3) = \mathbb{P}[S_1 = s_1, S_2 = s_2, S_3 = s_3] \]
\[ f^{(R)}(r_1, r_2, r_3) = \mathbb{P}[R^{(1)} = r_1, R^{(2)} = r_2, R^{(3)} = r_3] \]
\[ f^{(0)}_0 = \mathbb{P}[R^{(1)} = 0, R^{(2)} = 0, R^{(3)} = 0] \]
The multivariate analog of the Panjer's algorithm is as follows:

\[
f^{(S)}(0, 0, 0) = \Psi_N(f^{(R)})
\]

\[
f^{(S)}(s_1, s_2, s_3) = \frac{1}{(1 - af^{(R)}_0)} \sum_{x_1, x_2, x_3} \left( f^{(S)}(s_1 - x_1, s_2 - x_2, s_3 - x_3) \times f^{(R)}(x_1, x_2, x_3) \times \left[ a + b \frac{x_1}{s_1} \right] \right), \quad s_1 \geq 1
\]

\[
f^{(R)}(x_1, x_2, x_3) \times \left[ a + b \frac{x_2}{s_2} \right], \quad s_2 \geq 1
\]

\[
f^{(S)}(s_1, s_2, s_3) = \frac{1}{(1 - af^{(R)}_0)} \sum_{x_1, x_2, x_3} \left( f^{(S)}(s_1 - x_1, s_2 - x_2, s_n - x_n) \times f^{(R)}(x_1, x_2, x_3) \times \left[ a + b \frac{x_3}{s_3} \right] \right), \quad s_3 \geq 1
\]

where \( \Psi_N(u) = \mathbb{E}[u^N] \) denotes the probability generating function of \( N \), the probabilities \( p_n = \mathbb{P}[N = n] \) satisfy

\[p_n = (a + \frac{b}{n}) p_{n-1}\]

for \( n = 1, 2, \ldots \), and

\[
\sum_{x_1, x_2, x_3} g(x_1, x_2, x_3) = \sum_{x_1=0}^{\min(s_1, m_1)} \sum_{x_2=0}^{\min(s_2, m_2)} \sum_{x_3=0}^{\min(s_3, m_3)} g(x_1, x_2, x_3) - g(0, 0, 0),
\]

where

\[m_1 = \max(x \mid f^{(R)}(x, x_2, x_3) > 0),\]

\[m_2 = \max(x \mid f^{(R)}(x_1, x, x_3) > 0),\]

and

\[m_3 = \max(x \mid f^{(R)}(x_1, x_2, x) > 0).\]

Though the execution of this algorithm can be time-consuming, we can take advantage of a particular feature of the vector \((R^{(1)}, R^{(2)}, R^{(3)})\):

\[R^{(1)}_i \leq R^{(2)}_i \leq R^{(3)}_i \quad \text{for } i = 1, 2, \ldots, \]
This implies that the sums in the algorithm may be rewritten as:

\[
\sum_{x_1, x_2, x_3} g(x_1, x_2, x_3) = \sum_{x_3=0} \sum_{x_2=0} \sum_{x_1=0} g(x_1, x_2, x_3) - g_0
\]

where \(g_0 = g(0, 0, 0)\). As a corollary, we have \(S_1 \leq S_2 \leq S_3\). Therefore, the algorithm needs only to be evaluated for values of \((S_1 = s_1, S_2 = s_2, S_3 = s_3)\) such that \(s_1 \leq s_2 \leq s_3\). Moreover, only a few of the values of the random vector \((R^{(1)}, R^{(2)}, R^{(3)})\) have a positive probability. It may be more efficient to rewrite the algorithm as:

\[
f^{(S)}(0, 0, 0) = \Psi_N(f^{(R)}(0, 0, 0))
\]

\[
f^{(S)}(s_1, s_2, s_3) = \frac{1}{(1 - af^{(R)}(0, 0, 0))} \sum_{j=0}^{t} \left[ a + b \frac{Z(j, 1)}{s_1} \right] x
\]

\[
f^{(S)}(s_1, s_2, s_3) = \frac{1}{(1 - af^{(R)}(0, 0, 0))} \sum_{j=0}^{t} \left[ a + b \frac{Z(j, 2)}{s_2} \right] x
\]

\[
f^{(S)}(s_1, s_2, s_3) = \frac{1}{(1 - af^{(R)}(0, 0, 0))} \sum_{j=0}^{t} \left[ a + b \frac{Z(j, 3)}{s_3} \right] x
\]

where \(Z\) denotes a matrix with \(t\) rows and four columns. The number of rows in \(Z\) represents the number of points of \((R^{(1)}, R^{(2)}, R^{(3)})\) with positive mass [(excluding the possible point \((0, 0, 0)\)]. Column \(j\) represents the value of the \(R^j\) for \(j = 1, 2, 3\). The fourth column gives the probability associated with the realization \((R^{(1)} = Z(j, 1), R^{(2)} = Z(j, 2), R^{(3)} = Z(j, 3))\).
3 Numerical Results for Treaty 2

Let us develop actual prices for the exotic treaty given in Treaty 2. Two cases are considered: free reinstatements and paid reinstatements. In both cases we use a span $h = 2.5$ to discretize the distribution. The following parameters values are used in this section:

$$
A = 2.5, \quad B = 25, \quad \alpha = 0.85, \quad \lambda = 10.61
$$

$$
D = 2.5, \quad L_1 = 7.5, \quad L_2 = 15, \quad L_3 = 22.5
$$

$$
AAD_1 = 10, \quad AAD_2 = 5, \quad AAD_3 = 0,
$$

$$
AAL_1 = 4 \times 7.5 = 30, \quad AAL_2 = 4 \times 15 = 60, \quad \text{and} \quad AAL_3 = 3 \times 22.5 = 67.5.
$$

3.1 Free Reinstatements

First we find the pure premiums for the three layers using equations (5), (6), and (7):

$$
\mathbb{E}[S_1] = 21.20, \quad \mathbb{E}[S_2] = 17.31, \quad \mathbb{E}[S_3] = 7.18.
$$

Summing these premiums we obtain 45.68, which is the premium for an unlimited cover:

$$
\mathbb{E}[N] \times \mathbb{E}[\max(0, X - D)] = 10.61 \times 4.3056 = 45.68.
$$

This shows that, using the data given, this arrangement is about the same as an unlimited cover. The total liability offered by the reinsurance is $30 + 60 + 67.5 = 157.5$.

We can simplify the structure (for pricing purposes) and assume, for example, the following cover (which offers essentially the same capacity):

- 7.5 xs 2.5 with 9 reinstatements
- 7.5 xs 10 with 7 reinstatements
- 7.5 xs 17.5 with 5 reinstatements.

This cover offers the same aggregate liability: $67.5 + 52.5 + 37.5 = 157.5$, but the prices obtained by the classical univariate Panjer algorithm are

$$
\mathbb{E}[S_1] = 34.47, \quad \mathbb{E}[S_2] = 9.06, \quad \mathbb{E}[S_3] = 2.06,
$$
for a total of 45.59, which is a little smaller than the premium for an
unlimited cover (i.e., for cover without an annual aggregate limit). This
suggests that this alternative has a cover that is a little smaller than the
one given by the exotic excess of loss treaty.

If we change the covers to

- 7.5 xs 2.5 with 12 reinstatements;
- 7.5 xs 10 with 6 reinstatements; and
- 7.5 xs 17.5 with 3 reinstatements,

we obtain the following prices:

\[
\begin{align*}
E[S_1] &= 34.50, \\
E[S_2] &= 9.11, \\
E[S_3] &= 2.06.
\end{align*}
\]

The sum of these prices is 45.67, which is almost an unlimited cover.
Thus, it is not difficult to offer an almost unlimited cover to the insurer
without the exotic cover.

If these premiums are then applied to the exotic excess of loss treaty,
we obtain the following expected gains for the reinsurer:

\[
\begin{align*}
E[G_1] &= 13.3, \\
E[G_2] &= -8.2, \\
E[G_3] &= -5.12,
\end{align*}
\]

i.e., a total loss of 0.02. This shows that for free reinstatements, the
reinsurer who participates on all the layers (with the same share) can
almost replicate any treaty. We will see in the next section that this
situation dramatically changes when there are paid reinstatements.

### 3.2 Paid Reinstatements

Recall that the original exotic excess of loss treaty (Treaty 2) is with
paid reinstatements. Using the joint distribution of \( S_1, S_2, \) and \( S_3 \)
given in equations (5), (6) and (7), then from Sundt (1991) it is easy to calculate
the pure premium, \( P_t: \)

\[
P_t = \frac{E[S_i]}{1 + \sum_{j=1}^{k} \frac{c_j}{L_i} E[\max(0, S_i - (j - 1)L_i)]}
\]

(14)

where \( k \) is the number of reinstatements and \( c_j \) is the price of the \( j^{th} \)
reinstatement. With all reinstatements at 100% we obtain the following
prices:

\[
P_1 = 6.22, \quad P_2 = 8.15, \quad P_3 = 5.44.
\]
The premiums obtained within the classical excess of loss treaty

- 7.5 xs 2.5 with 12 reinstatements;
- 7.5 xs 10 with 6 reinstatements; and
- 7.5 xs 17.5 with 3 reinstatements

are

\[ P_1 = 6.16, \quad P_2 = 4.11, \quad P_3 = 1.61. \]

If these premiums are then applied to the exotic excess of loss treaty, we obtain the following expected gains for the reinsurer:

\[ \mathbb{E}[G_1] = -0.22, \quad \mathbb{E}[G_2] = -8.61, \quad \mathbb{E}[G_3] = -5.06, \]

i.e., a total loss of 13.89. This loss shows the importance of using the correct model to price each layer.

### 3.3 Annual Aggregate Deductibles

Finally, we analyze the effect of the annual aggregate deductibles within the exotic excess of loss treaty. First, let us assume that there are no annual aggregate deductibles. Table 8 shows the premiums already derived for various levels of annual aggregate deductibles (AAD).

<table>
<thead>
<tr>
<th>Table 8 Premiums for Various Layers With and Without Reinstatements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free Reinstatements</td>
</tr>
<tr>
<td>AAD₁ = AAD₂ = AAD₃ = 0</td>
</tr>
<tr>
<td>L₁</td>
</tr>
<tr>
<td>L₂</td>
</tr>
<tr>
<td>L₃</td>
</tr>
<tr>
<td>AAD₁ = 10, AAD₂ = 5, AAD₃ = 0</td>
</tr>
<tr>
<td>L₁</td>
</tr>
<tr>
<td>L₂</td>
</tr>
<tr>
<td>L₃</td>
</tr>
</tbody>
</table>

Table 9 shows the premiums with AAD₁ = 20 and AAD₂ = 10. The effect of the annual aggregate deductibles on the first two layers is nil.
Indeed, the sum of the premiums with free reinstatements is constant (45.68). The only effect of these annual aggregate deductibles is to distribute the claims differently between the layers. With very large annual aggregate deductibles, \( \text{AAD}_1 = 60 \), \( \text{AAD}_2 = 90 \), and \( \text{AAD}_3 = 0 \), there is an effect with respect to the total liability of the reinsurer. So far the effect of the annual aggregate deductibles is due to the annual aggregate limit of the third layer. If this layer has unlimited reinstatements, then we would not observe the effect of large deductibles on layers 1 and 2. The final part of Table 9 shows the case with an annual aggregate deductible on layer 3. In this case the sum of the premiums with free reinstatements is 38.67, showing the effect of the annual aggregate deductible on the third layer. Clearly only an annual aggregate deductible on the third layer has the effect of a classical annual aggregate deductible.

| Table 9 |
|---|---|
| **Premiums for Various Layers** |
| **With and Without Reinstatements** |

<table>
<thead>
<tr>
<th>Reinstatements</th>
<th>Free</th>
<th>@ 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AAD_1 = 20, AAD_2 = 10, AAD_3 = 0</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>14.12</td>
<td>5.26</td>
</tr>
<tr>
<td>L2</td>
<td>19.50</td>
<td>8.54</td>
</tr>
<tr>
<td>L3</td>
<td>12.07</td>
<td>7.86</td>
</tr>
<tr>
<td><strong>AAD_1 = 60, AAD_2 = 90, AAD_3 = 0</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>L2</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>L3</td>
<td>43.41</td>
<td>16.47</td>
</tr>
<tr>
<td><strong>AAD_1 = 10, AAD_2 = 5, AAD_3 = 15</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>21.13</td>
<td>6.22</td>
</tr>
<tr>
<td>L2</td>
<td>17.37</td>
<td>8.15</td>
</tr>
<tr>
<td>L3</td>
<td>0.17</td>
<td>0.17</td>
</tr>
</tbody>
</table>
4 Closing Comments

We have shown how to price an exotic excess of loss treaty using a multivariate analog of Panjer's algorithm. The exotic structure presents mathematical difficulties that can be avoided by using a certain definition of inuring that ensures the order of the claims has no effect on the pricing. Numerical examples show that it is important to correctly identify the actuarial model in order to obtain accurate pricing. We also show that other classical reinsurance structures may also provide a similar level of cover.

Our calculations were based on a step size of $h = 2.5$, which may appear too large. A short sensitivity analysis, however, shows that halving the step size to $h = 1.25$ does not significantly affect the premiums for the original reinsurance program.

References


Modeling Size-of-Loss Distributions for Exact Data in WinBUGS

David P.M. Scollnik*

Abstract†

This paper discusses how the statistical software WinBUGS can be used to implement a Bayesian analysis of several popular severity models applied to exact size-of-loss data. The particular models targeted are the gamma, inverse gamma, loggamma, lognormal, (two-parameter) Pareto, inverse (two-parameter) Pareto, Weibull, and inverse Weibull distributions. It is possible to implement additional size-of-loss models (including those for truncated data) using methods analogous to those described herein.

Key words and phrases: Bayesian, severity, Markov chain Monte Carlo, simulation

1 Introduction

1.1 Why WinBUGS?

BUGS (Bayesian inference using Gibbs sampling) is a specialized suite of statistical software packages for implementing Markov chain Monte Carlo (MCMC)-based analysis of full probability models in which all unknowns are treated as random variables. The BUGS programming language allows the user to make a straightforward specification of the full probability model under consideration. The Windows version of

*David P.M. Scollnik, A.S.A., Ph.D., is an associate professor of actuarial science and statistics at the University of Calgary. He holds a Ph.D. in statistics from the University of Toronto.

Dr. Scollnik's address is: Department of Mathematics and Statistics, University of Calgary, 2500 University Drive N.W., Calgary, Alberta, Canada, T2N 1N4. Internet address: scollnik@math.ucalgary.ca and <http://www.math.ucalgary.ca/~scollnik>

†An earlier unpublished version of this paper was supported by a grant from the Actuarial Education and Research Fund (AERF). This version was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).
BUGS is known as WinBUGS, and is available from the BUGS Project website at: <http://www.mrc-bsu.cam.ac.uk/bugs>.

Scollnik (2001) describes how a number of different actuarial models can be implemented and analyzed in accordance with the Bayesian paradigm using the MCMC simulation method via BUGS. The MCMC method can be used to generate a dependent sequence of random draws from a Markov chain with a stationary distribution equal to the distribution associated with some probabilistic model of interest, even if the distribution is multi-dimensional with a very complicated form. A wide variety of simulation-based inferences for the model then can be developed on the basis of these dependent simulated values.

Due to its astonishing flexibility and to its ability to simplify the analysis of even extremely complicated multi-dimensional random models, the MCMC method has become increasingly popular over the last dozen or so years, as is evident in the statistical and related literature. See Scollnik (2001) for a detailed description of the MCMC method, list of references, and summary of recent actuarial applications.

1.2 Objectives

The main purpose of this paper is to show actuaries how WinBUGS can be used to implement a Bayesian analysis of several popular severity models when the data consist of the exact size of losses, i.e., before items such as deductibles and policy limits are applied. Scollnik (2001) considers only the case of grouped size-of-loss data, i.e., where losses are grouped according to size.

The particular models (distributions) studied in this paper are the gamma, inverse gamma, loggamma, lognormal, (two-parameter) Pareto, inverse (two-parameter) Pareto, Weibull, and inverse Weibull distributions. Each of these models is applied to the size-of-loss data in Table 1, after which we discuss how Bayesian posterior prediction and model checking and selection can be performed. Several authors have demonstrated that Bayesian predictions are an improvement over traditional classical statistical predictions based on conditioned maximum likelihood estimates. Bayesian predictions incorporate parameter uncertainty and prior information, which are, in effect, ignored by classical statistical predictions. See, for example, Dickson, Tedesco, and Zehnwirth (1998), Cairns (2000), and Scollnik (2002) for a discussion of this point along with some numerical comparisons.

While this paper introduces relevant WinBUGS programming tips and implementation details, Scollnik (2001) should be referenced for more detailed information about MCMC-based simulations in general,
and for specific details regarding the actual operation of the WinBUGS software in particular. The reader is assumed to have a basic working knowledge of WinBUGS. In addition, we assume the reader is familiar with the general nature of Bayesian inference. A quick overview of the Bayesian approach is as follows: Suppose the data consist of \( n \) independent observations \( x_i, i = 1, 2, \ldots, n \), from a common density function \( g(x_i | \alpha, \beta) \) where \( \alpha \) and \( \beta \) are random parameters (possibly vector valued) with joint prior density \( \pi(\alpha, \beta) \). From Bayes theorem and the conditional independence of losses, the posterior distribution is

\[
\pi(\alpha, \beta | x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n} g(x_i | \alpha, \beta) \pi(\alpha, \beta)}{\int \prod_{i=1}^{n} g(x_i | \alpha, \beta) \pi(\alpha, \beta) \, d\alpha \, d\beta} \propto \pi(\alpha, \beta) \prod_{i=1}^{n} g(x_i | \alpha, \beta).
\]

In the Bayesian context, inferences concerning the unknown model parameters are constructed from the posterior distribution. The posterior distribution describes all that is known about the unknown model parameters in light of the observed data and prior information. The posterior knowledge can be summarized using summary statistics such as posterior means, quantiles, and variances or summarized graphically by posterior density plots and the like. Instead of deriving the form of the posterior distribution and the value of its desired summary statistics analytically, it is common and often easier to simulate random draws from the posterior distribution and then use this posterior sample to fashion the posterior inferences (e.g., via empirical posterior summary statistics and density plots).

MCMC is one method of simulating random draws from a Markov chain with a stationary distribution equal to the posterior distribution. WinBUGS is a useful and easy-to-use software package that can be used to implement these simulations. WinBUGS does not require that the user analytically derive the posterior distribution first. Rather, the user need only specify the conditional model generating the data and the prior distribution for any unknown parameters. WinBUGS uses this information to construct, and simulate random draws from, a Markov chain with a stationary distribution equal to the correct posterior distribution for the model and data under consideration.

One advantage of a simulation-based Bayesian analysis is that the simulation results can be reused. For instance, random draws from the posterior distribution of \((\alpha, \beta)\) can be used to estimate the posterior mean of \( \alpha \) or of \( \beta \). The same random draws, however, also can be
used to make inferences about any function of \( \alpha \) and \( \beta \), say \( h(\alpha, \beta) \), by simply applying \( h(\cdot, \cdot) \) to each random draw of \((\alpha, \beta)\) and then summarizing the results. So if \( m \) draws of \((\alpha_j, \beta_j)\), \( j = 1, 2, \ldots, m \), are made from the distribution \( \pi(\alpha, \beta|x_1, \ldots, x_n) \), then one can use \( h(\alpha_j, \beta_j) \), \( j = 1, 2, \ldots, m \) to fashion inferences about the posterior distribution of \( h(\alpha, \beta) \). The function of interest may even be the likelihood function, i.e.,

\[
h(\alpha, \beta) = l(\alpha, \beta|x_1, \ldots, x_n) = \prod_{i=1}^{n} g(x_i|\alpha, \beta)
\]
as in the example above, or the log-likelihood function.

Good discussions of Bayesian inference are provided in Klugman (1992) and Klugman et al., (1998, Section 2.8). Makov (2001) gives an overview of principal applications of Bayesian methods in actuarial science, while Scollnik (2001) includes many additional references to recent papers in actuarial science with a Bayesian perspective. Gelman, Carlin, Stern, and Rubin (1995) is an excellent non-actuarial text on Bayesian data analysis that also discusses many simulation methods, including MCMC, for use in Bayesian analyses.

### Table 1

<table>
<thead>
<tr>
<th>Twenty Exact Size of Losses</th>
<th>59</th>
<th>71</th>
<th>127</th>
<th>217</th>
</tr>
</thead>
<tbody>
<tr>
<td>223</td>
<td>524</td>
<td>537</td>
<td>1,089</td>
<td></td>
</tr>
<tr>
<td>1,127</td>
<td>1,181</td>
<td>1,189</td>
<td>1,516</td>
<td></td>
</tr>
<tr>
<td>1,681</td>
<td>1,708</td>
<td>1,784</td>
<td>3,639</td>
<td></td>
</tr>
<tr>
<td>5,386</td>
<td>6,100</td>
<td>9,945</td>
<td>15,295</td>
<td></td>
</tr>
</tbody>
</table>

2 Size-of-Loss Model Specification in WinBUGS

Though WinBUGS can be used to analyze complex stochastic models, it explicitly supports only a few continuous distributions (size-of-loss models). These include the beta, chi-squared, double exponential, exponential, gamma, normal, \( t \), (single-parameter) Pareto, uniform, and Weibull distributions. Before using any one of these distributions, however, the practitioner should note the parameterization of distributions given in the WinBUGS User Manual in order to avoid any possibility of confusion.
Though several of our models are not explicitly supported by WinBUGS, some of them can be constructed from those available in WinBUGS using mixtures of distributions and/or by applying simple transformations, such as the inverse or logarithmic transform, to the data. The remainder can be implemented using the general purpose 'ones' or 'zeroes' tricks described below in our discussion of the Pareto models.

Dempster (1974; reprinted in 1997) suggested that one might examine the posterior distribution of the loglikelihood to assist with model selection. To this end the node NLL in our WinBUGS program, which represents the negative log of the likelihood function for the observed exact size of loss values, will be used. The value of NLL depends on the unobserved model parameters, and the different values it takes as the model is updated in WinBUGS can be monitored like those of any other node. Our strategy is simple: monitor the value of NLL and choose the model with the smallest posterior mean for NLL. See Spiegelhalter, Best, Carlin, and van der Linde (2001) for modifications of this approach that are particularly useful when the models under consideration differ in complexity (the number of free parameters).

In this section we will review the definitions for our targeted models, and discuss how they may be coded in WinBUGS. It should be understood that the code can be ported over to 'classic' BUGS with little effort.

2.1 The Gamma, Inverse Gamma, and Loggamma Models

Let \( x \) denote an observed exact size-of-loss value. In WinBUGS, the declaration

\[
x \sim \text{dgamma}(\alpha, \beta)
\]
corresponds to a definition of the gamma model with probability density function

\[
f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,
\]

with \( \alpha > 0 \) and \( \beta > 0 \).

Consider what happens if we assign a gamma distribution, as in equation (1), to the transformed variable \( y = \log(x) \) instead. The resulting pdf of \( x \) is

\[
g(x|\alpha, \beta) = \frac{\beta^\alpha(\log(x))^{\alpha-1}}{\Gamma(\alpha) x^{\beta+1}}, \quad x > 1,
\]

with \( \alpha, \beta > 0 \), which is the definition of the density function for the loggamma model. In WinBUGS, the lines of code
describe the loggamma model defined in equation (2). Specifically, the first line states the relationship between $x$ and $y$ and the second line assigns the relevant density type to $y$. The order of the two lines is actually immaterial to WinBUGS.

As before, let $x$ denote the exact size of loss. This time assign a gamma distribution, as in equation (1), to the inverse transformed variable $y = 1/x$. The resulting pdf of $x$ is

$$h(x|\alpha, \beta) = \frac{\beta^\alpha \exp(-\beta/x)}{\Gamma(\alpha) x^{\alpha+1}}, \quad x > 0,$$

with $\alpha, \beta > 0$, which is the definition of the density function for the inverse gamma model. In WinBUGS, the lines of code

```r
y <- log(x)
y ~ dgamma(alpha, beta)
```

describe the inverse gamma model.

All of the models described above will need to be completed by adding prior density specifications for the model parameters. Suppose that we are interested in modeling an inverse gamma model to the twenty exact size of loss observations appearing in Table 1. Then our specification of a complete inverse gamma model in WinBUGS might proceed as shown below.

**CODE FOR THE INVERSE GAMMA MODEL**

```r
model;
{
  # Compute negative loglikelihood (NLL) in terms of x.
  NLL <- - sum( loglik[] )
  for( i in 1 : N ) {
    loglik[i] <- alpha * log( beta ) - loggam( alpha ) -
    beta / x[i] - ( alpha + 1 ) * log( x[i] )
  }

  # Define exact size-of-loss random variables.
  for( i in 1 : N ) {
    y[i] <- 1 / x[i]
    y[i] ~ dgamma( alpha, beta )
  }
}
```
# Define 'naive' prior densities for founder nodes.
alpha ~ dgamma( 0.001, 0.001 )
beta ~ dgamma( 0.001, 0.001 )

# More informative priors, each with mean = mle and
# sd = 5 x mle. See discussion below for more details.
#
# alpha ~ dgamma( aparm1, aparm2 )
# beta ~ dgamma( bparm1, bparm2 )
#
# amle <- 0.5661338 ; aparm1 <- 0.04
# aparm2 <- aparm1 / amle
# bmle <- 193.6986 ; bparm1 <- 0.04
# bparm2 <- bparm1 / bmle

DATA
list( N = 20,
      x = c( 59, 71, 127, 217, 223, 524, 537, 1089, 1127,
            1181, 1189, 1516, 1681, 1708, 1784, 3639,
            5386, 6100, 9945, 15295 ) )

INITS
list( alpha = 2, beta = 2 )

The prior density specifications assigned to the random parameters 
\( \alpha \) and \( \beta \) in the sample code above are independently gamma random
variables with common mean and variance of 1 and 1000, respectively.
This is a naive assumption. While the selection of gamma distributions
is reasonable enough for parameters that are non-negative valued (like
\( \alpha \) and \( \beta \)), it is difficult to believe that an experienced actuary cannot give
a more informed specification of the prior mean and variance. When
all else fails, it may be reasonable—or at least, be not too objectionable
from a pragmatic point of view—to assign each variable a gamma
distribution a priori with mean and standard deviation equal to its maximum
likelihood estimate (mle) and, say, five times its mle, respectively.
Such a distribution is approximately centered in the appropriate region
yet is still widely spread. Code implementing this mle strategy is provided above for illustration’s sake. The mle values themselves were determined outside of WinBUGS using standard techniques. When the data are used to estimate prior parameters in this way, the analysis is sometimes called empirical Bayes (Gelman, et al., 1995, page 123).

Illustrative WinBUGS code for the various models described above appear in the file exact.odc available on this author's website at: <http://www.math.ucalgary.ca/~scollnik/abcd/>. The same is true for each of the models described in the following sections.

2.2 The Lognormal Model

In WinBUGS, the declaration

\[
x \sim \text{dnorm}(\mu, \tau)
\]

corresponds to a definition of the normal model with density function

\[
f(x|\mu, \tau) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(x - \mu)^2\right), \quad -\infty < x < \infty, \quad (4)
\]

with \(-\infty < \mu < \infty\) and \(\tau > 0\). In this parameterization, \(\tau\) is called the precision or inverse variance parameter.

Consider what happens if we assign a normal distribution as in equation (4) to the transformed variable \(y = \log(x)\). The resulting pdf of \(x\) is

\[
g(x|\mu, \tau) = \frac{\sqrt{\tau}}{x\sqrt{2\pi}} \exp\left(-\frac{\tau}{2} [\log(x) - \mu]^2\right), \quad (5)
\]

for \(x > 0\), with \(-\infty < \mu < \infty\) and \(\tau > 0\). This is the definition of the density function for the lognormal model. In WinBUGS, the lines of code

\[
y <- \log(x) \\
y \sim \text{dnorm}(\mu, \tau)
\]

describe the lognormal model defined above. Another way in which to define the same lognormal model is with the declaration

\[
x \sim \text{dlnorm}(\mu, \tau)
\]

This appears to work for all recent versions of WinBUGS, even though the dlnorm density is undocumented in the User Manual for some recent versions of WinBUGS.
As usual, we still need to complete the model with a prior density specification and also define the data and initial values. Our complete model specification might proceed as shown below:

**CODE FOR THE LOGNORMAL MODEL**

```r
model;
{
    # Compute negative loglikelihood (NLL) in terms of x.
    NLL <- - sum( loglik[] )
}

for( i in 1 : N ) {
    loglik[i] <- - log( sqrt( 2 * Pi / tau ) ) -
    log( x[i] ) - pow( log( x[i] ) -
    mu, 2 ) * tau / 2
}

Pi <- 3.14159265

# Define the exact size of loss random variables.
for( i in 1 : N ) {
    y[i] <- log( x[i] )
    y[i] ~ dnorm( mu, tau )
}

# Define 'naive' prior densities for the founder nodes.
mu ~ dnorm( 0, 0.001 )
tau ~ dgamma( 0.001, 0.001 )

# More informative priors, each with mean = mle and
# sd = 5 x mle. See discussion below for more details.
#
# mu ~ dnorm( mparm1, mparm2 )
# tau ~ dgamma( tparm1, tparm2 )
#
# mmle <- 6.936106 ; mparm1 <- mmle
# mparm2 <- 1 / pow( 5 * mmle, 2 )
# tmle <- 0.432222 ; tparm1 <- 0.04
# tparm2 <- tparm1 / tmle
```
DATA

```
list( N = 20,
     x = c( 59, 71, 127, 217, 223, 524, 537, 1089, 1127,
           1181, 1189, 1516, 1681, 1708, 1784, 3639,
           5386, 6100, 9945, 15295 ) )
```

INITS

```
list( mu = 2, tau = 2 )
```

Again, a definition of the NLL is included in the code and its values can be monitored and used to assist with model selection. Note that we adopted a prior normal distribution for the parameter $\mu$ (i.e., instead of a gamma distribution), as the support for this parameter is the entire real number line. Included for illustration's sake, is a more informative prior density specification for each model parameter, as before, centered at that parameter's mle and with standard deviation equal to five times the mle.

2.3 The Weibull and Inverse Weibull Models

In WinBUGS, the declaration

```
x ~ dweib( tau, lambda )
```

corresponds to a definition of the Weibull model with density function

$$f(x|\tau, \lambda) = \tau \lambda x^{\tau-1} \exp(-\lambda x^{\tau}), \quad x > 0,$$

with $\tau > 0$ and $\lambda > 0$.

Consider what happens if we assign a Weibull distribution as in (6) to the transformed variable $y = 1/x$. The density of $x$ is

$$h(x|\alpha, \beta) = \frac{\tau \lambda \exp(-\lambda/x^{\tau})}{x^{\tau+1}}, \quad x > 0,$$

with $\tau > 0$ and $\lambda > 0$. This is the definition of the density function for the inverse Weibull model. In WinBUGS, the lines of code

```
y <- 1 / x
y ~ dweib( tau, lambda )
```
describe the inverse Weibull model.

As usual, we still need to complete either model with a prior density specification and also define the data and initial values. We omit a presentation of either complete model specification as they are both similar to those presented earlier in this section. As mentioned earlier, the code is available on this author's website.

2.4 The Pareto and Inverse Pareto Models

The discussion of the Pareto and inverse Pareto models has been left for last, as the tricks used to implement these models have more general application and deserve to be emphasized. In recent versions of WinBUGS, specifically (beta) Version 1.2 (May, 1999) or later, a version of the Pareto model is available with the declaration

\[ x \sim \text{dpar}(\alpha, \theta) \]

This declaration, however, corresponds to the single-parameter Pareto model with density function

\[ f(x|\alpha, \theta) = \frac{\alpha \theta^\alpha}{x^{\alpha+1}}, x > \theta, \]

with \( \alpha > 0 \) and \( \theta > 0 \). This form of the Pareto distribution may be appropriate in certain instances, for example when modeling losses above a given deductible. This distribution is used in the analysis of the motor example in Section 4 of Scollnik (2000).

As the data in Table 1 have no deductible associated with them, a more sensible version of the Pareto distribution for this context would be the two-parameter model with density function

\[ f(x|\alpha, \theta) = \frac{\alpha \theta^\alpha}{(x + \theta)^{\alpha+1}}, \quad x > 0, \]  

with \( \alpha > 0 \) and \( \theta > 0 \). A related distribution is the inverse Pareto model which arises in the expected manner by assigning a Pareto distribution as in equation (8) to the transformed variable \( y = 1/x \). The density of \( x \) is

\[ h(x|\alpha, \theta) = \frac{\alpha \theta^\alpha x^{\alpha-1}}{(1 + x\theta)^{\alpha+1}}, \quad x > 0, \]  

with \( \alpha > 0 \) and \( \theta > 0 \). Although neither of these distributions is explicitly supported in WinBUGS, we are aware of two ways in which to implement them.
The first is based on a trick that was originally found on the FAQ (frequently asked questions) page of the BUGS website at <http://www.mrc-bsu.cam.ac.uk/bugs>. This 'ones' trick now appears in Section 3.2 of the WinBUGS User Manual and its discussion there reads as follows:

Suppose your data is $y$ (of length $n$) and you want to fit the model $p(y) = f(y, t)$ where $t$ are the unknown parameters and $f$ is the formula of the density that is not currently handled by BUGS.

The trick is to create a new vector 'ones', that comprises just 1's and is of length $n$ (note the use of the data transformation ability described in Section 3.7). Then use the BUGS code:

```r
for(i in 1 : n) {
    ones[i] <- 1
    ones[i] ~ dbern( p[i] )
    p[i] <- f(y[i], t) / K
}
```

where $K$ is a sufficiently large constant to ensure that all sampled values of $p[i]$ are less than one. This should provide a likelihood term proportional to $f(y, t)$.

To illustrate, in the case of a random sample from the two-parameter Pareto model, with density function equation (8), we would assign

$$p[i] <- \alpha \times \frac{\text{pow}(\theta, \alpha)}{\text{pow}(x[i] + \theta, \alpha + 1)}$$

When using the inverse Pareto model, with density function equation (9), we would use the lines of code

```r
y[i] <- 1 / x[i]
p[i] <- \alpha \times \frac{\text{pow}(\theta, \alpha)}{\text{pow}(y[i] + \theta, \alpha + 1) / \text{pow}(x[i], 2)}
```

It should be apparent to the reader that this 'ones' trick can be used to construct the likelihood function for a sample drawn from any continuous distribution, including truncated models, provided that the relevant density function may be expressed using the operators $+,-,\times,/,$ and the standard mathematical functions (e.g., exp, log, abs, and sqrt)
listed in Table I of the WinBUGS User Manual. Incidentally, as the likelihood function for the observed data is the product of the \( p[i] \) terms, it will be an easy matter to calculate the node NLL, as it is simply equal to the negative logarithm of this product.

A variation of the 'ones' trick was first suggested to us through a public communication by Serguei N. Smirnov on an email discussion list devoted to BUGS at: <http://www.jiscmail.ac.uk/lists/bugs.html>. Smirnov's idea was to modify the 'ones' trick by using an exponential distribution in place of the Bernoulli as follows

```plaintext
for(i in 1 : n) {
  zeroes[i] <- 0
  zeroes[i] ~ dexp( p[i] )
  p[i] <- f(y[i],t)
}
```

The advantage to Smirnov's method is that a large constant \( K \) need no longer be specified.

The second method with which to implement the two-parameter Pareto and inverse Pareto models relies on the observation that a two-parameter Pareto random variable can be defined as a mixture of two gamma random variables. (See, for example, Hogg and Klugman, 1984, page 54.) Specifically, if the distribution of \( x \) given \( T \) is \( \Gamma(1,T) \) and the distribution of \( T \) given \( (X \text{ and } \epsilon) \) is \( \Gamma(\alpha,\theta) \), then the distribution of \( x \) given \( (x \text{ and } \epsilon) \) has the density function equation (8). To code this relationship in WinBUGS, we would use the lines of code

```plaintext
x[i] ~ dgamma( 1, tau[i] )
tau[i] ~ dgamma( alpha, theta )
```

It is important to note that each observation requires its own mixing parameter \( \tau[i] \); see Section 2.7.3.4 of Klugman et al., (1998), for a further discussion of this point and of mixture modeling in general. The lines

```plaintext
y[i] <- 1 / x[i]
y[i] ~ dgamma( 1, tau[i] )
tau[i] ~ dgamma( alpha, theta )
```

serve to define the inverse Pareto model.

Other distributions with interpretations as mixture models may be implemented in an analogous manner. Although hard and fast advice
is difficult to give, our experience suggests that the 'ones' and 'zeroes' tricks lead to complete model specifications which update more quickly and also take fewer updates to converge in WinBUGS. The mixture modeling approach is still valuable, though, as it may be used to generate posterior predictive draws from the two-parameter Pareto models described above. This is discussed below.

No matter which of the the methods we adopt, we still need to complete the model with a prior density specification and also define the data and initial values in the usual way. Illustrative code for a complete model specification appears in the aforementioned exact.odc file at the author's website.

3 Posterior Predictive Draws and Model Checks

The preceding discussion described how a variety of size of loss models can be implemented in WinBUGS. By examining the values of the NLL node associated with each model, selection between competing models is facilitated. Models with low values of the NLL are generally preferred, ceteris paribus. Although examination of the values taken by the NLL node will provide some guidance as to how well a particular model fits a given data set, it does not tell the complete story. Model checking is also important and is discussed in Gelman et al., (1995, especially Chapters 6 and 18). One method presented by these authors, that of posterior predictive checks, involves drawing simulated values from the posterior predictive distribution of replicated data and comparing these samples to the observed data (Gelman, et al., 1995, pages 162-174). Systematic differences between the simulations and observed data indicate potential failings of the model.

The method of posterior predictive checks is fairly simple to implement using WinBUGS. The first step is to generate a replicated sample from the same model (i.e., from the same distribution and with the same model parameter values) that is assumed to have generated the observations at hand. The replicated sample is of the same size as the original and would use the identical covariate values if the model happened to contain explanatory variables. Often, this replicated sample is easily obtained by essentially duplicating the code used to model the original observations. For example, suppose we were assuming a loggamma model as in equation (2) for the data in Table 1 and so had specified

\[ y[i] \leftarrow \log( x[i] ) \]
Scollnik: Modeling Size-of-Loss Distributions

```r
y[i] ~ dgamma( alpha, beta )
```

Then the replicated data would be defined analogously with the lines

```r
x.rep[i] <- exp( y.rep[i] )
y.rep[i] ~ dgamma( alpha, beta )
```

Note the transformations are now coded from `y.rep[i]` to `x.rep[i]` as the former is logically defined in advance of the latter. The same idea works for all of the distributions previously discussed except the Pareto and inverse Pareto. As these two are not explicitly supported in WinBUGS, we utilize the mixture model interpretation of the Pareto distribution in order to generate the predictive draws. In the case of the Pareto model with density function equation (8), this is accomplished with the lines of code

```r
x.rep[i] ~ dgamma( 1, tau.rep[i] )
```

```r
tau.rep[i] ~ dgamma( alpha, theta )
```

whereas the code segment

```r
x.rep[i] <- 1 / y.rep[i]
y.rep[i] ~ dgamma( 1, tau.rep[i] )
tau.rep[i] ~ dgamma( alpha, theta )
```

would be appropriate if we were assuming the inverse Pareto model with density function equation (9).

The next step is to compare the simulated values from the posterior predictive distribution to the observed data. This may be accomplished using graphical summaries or through the use of test quantities. Here, we will briefly describe the latter approach and direct the reader to Figures 6.3-6.5, 13.2, and 16.2-16.3 in Gelman et al., (1995) for examples of the former. In any case, the reader is once again referred to Gelman et al., (1995, pages 162-174) for a more extensive discussion of posterior predictive model checking.

Let \( x = (x_1, x_2, \ldots, x_n) \) be the observed data, let \( \theta \) be the vector of unknown model parameters, and let \( x^{rep} \) be the replicated data as defined above that might have been observed if a new sample of observations were sampled from the same distribution and with the same model parameter values used to generate \( x \). A test quantity, also called a discrepancy measure, \( T(x, \theta) \) is a scalar summary of the parameters and data that is used as a standard when comparing the observed data to the replicate simulations. The possibilities include, but are certainly not limited to,
\[ T(x, \theta) = \min(x_i), \]
\[ T(x, \theta) = \sum x_i, \quad \text{and} \]
\[ T(x, \theta) = \bar{x} - E(X_i|\theta). \]

Test quantities are suggested by the problem context, and some examples are considered below. Any given discrepancy measure can also be calculated using the posterior simulations of \((x^{\text{rep}}, \theta)\) in order to obtain values we denote \(T(x^{\text{rep}}, \theta)\).

The Bayesian posterior predictive \(p\)-value is defined as the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity and given the assumed model. Mathematically, we write

\[
\text{Bayes } p\text{-value} = \mathbb{P} \left[ T(x^{\text{rep}}, \theta) \geq T(x, \theta) \mid x \right],
\]

with the probability understood to be taken over the joint posterior distribution of \((x^{\text{rep}}, \theta)\); that is, over the joint conditional distribution of \((x^{\text{rep}}, \theta)\) given the observed data. WinBUGS will automatically generate random draws from this posterior distribution, provided that the replicated data were defined in WinBUGS as described earlier in this section.

When the tail-area probability equation (10) is close to 0 or 1 for some meaningful test quantity, the assumed model is suspect. In this case, the definition of the discrepancy measure might suggest how the model can be improved. For example, suppose \(T(x, \theta) = \max(x_i)\) and the Bayes \(p\)-value is approximately 0.84. This says that nearly 17 times out of 20 the assumed model will generate a predictive sample containing a maximum value greater than that observed in the original sample. The practitioner will have to decide whether or not this is a crucial model failing, given the problem context. It needn’t be, say, if the practitioner’s real interest is in developing inferences with respect to the distribution of total future claims and the test quantity \(T(x, \theta) = \sum x_i\) happens to yield a Bayes \(p\)-value close to 0.50. But if it is judged to be a crucial failing, the practitioner might try a model with a thinner tail. As an alternative course of action, the practitioner may keep the assumed model for the original sample but impose a reasonable a priori upper bound on each predictive draw. See Gelman et al., (1995, pages 463–468) for an example of this sort.

When inference is proceeding on the basis of a MCMC simulation, as with WinBUGS, it is easy to estimate the Bayes \(p\)-value by monitoring
the values taken on by an indicator variable assigned equal to 1 when \( T(x_{rep}, \theta) \geq T(x, \theta) \), and 0 otherwise. The average of these values is an estimate of equation (10). In the particular case of the exact size of losses in Table 1, it may make sense to monitor the minimum, maximum, and total losses in each of the replicated data sets. The WinBUGS code following below could be used to implement the appropriate posterior predictive checks. The approximate Bayes p-values are equal to the estimated posterior means of the nodes \( p_{repmin} \), \( p_{repmax} \), and \( p_{repsum} \).

ILLUSTRATIVE CODE FOR POSTERIOR PREDICTIVE CHECKS

```r
# Use the step function to define indicator variables
# with which to estimate the Bayes p-values. The
# step function is equal to 0 (1) when its argument
# is less than (greater than or equal to) zero.
# So, for example, \( p_{repmin} \leftarrow \text{step}( x_{repmin} - x_{min} \right)
# is assigned the value of 1 if \( x_{repmin} \geq x_{min} \).

p_{repmin} \leftarrow \text{step}( x_{repmin} - x_{min} \)
p_{repmax} \leftarrow \text{step}( x_{repmax} - x_{max} \)
p_{repsum} \leftarrow \text{step}( x_{repsum} - x_{sum} \)

# Calculate min, max, and total of observed data.

x_{min} \leftarrow \text{ranked}( x[], 1 \)
x_{max} \leftarrow \text{ranked}( x[], N \)
x_{sum} \leftarrow \text{sum}( x[] \)

# Calculate min, max, and total of replicated data.

x_{repmin} \leftarrow \text{ranked}( x_{rep[]}, 1 \)
x_{repmax} \leftarrow \text{ranked}( x_{rep[]}, N \)
x_{repsum} \leftarrow \text{sum}( x_{rep[]} \)
```

4 Fitting the Models to the Data in Table 1

Finally, we are ready to apply the models and methods discussed in this section to the exact size of loss data in Table 1. In each case we have assumed independent prior distributions for all model parameters. Positive model parameters were assigned prior gamma distribu-
tions and the lognormal model's real parameter $\mu$ was assigned a normal prior distribution. Each model parameter had its prior distribution assigned a mean and standard deviation equal to its mle and five times its mle, respectively. These distributions are clearly informative, but we would argue only very weakly so. In practice, the actuarial practitioner often will be able to ascertain more informative prior distributions than these from past experience.

Each model compiled readily in WinBUGS and updated fairly quickly. The loggamma model was typical of the majority and took three seconds to burn-in for 5000 updates and 25 seconds to run for an additional 20,000 iterations on a dual 200 MHz Pentium Pro PC. The Pareto and inverse Pareto models were slowest, and each took about twice as long to run as the others. Summary statistics for the eight models appear in Table 2. The estimates from WinBUGS are based on the final 20,000 of the 25,000 iterations performed for each model. On the basis of the summary statistics for the NLL node, the lognormal and Pareto models rank as our first and second choices. The posterior predictive checks we monitored give us no reason to suspect either model.

Note that WinBUGS will always output estimated posterior means and SDs for the nodes $x.\text{repmi}$, $x.\text{repmax}$, and $x.\text{repsum}$ using sample moment calculations applied to the 20,000 simulated values of each, even though the corresponding theoretical posterior predictive moments may not exist under the assumed model. In these cases, the posterior mean and SD estimates should be ignored. If it is believed a priori that certain predictive moments do exist, then the model parameters should be constrained appropriately.
### Table 2

Estimated Posterior Summary Statistics

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Mean</th>
<th>SD</th>
<th>2.5%</th>
<th>Median</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gamma</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLL</td>
<td>177.3</td>
<td>1.03</td>
<td>176.3</td>
<td>177.0</td>
<td>180.1</td>
</tr>
<tr>
<td>alpha</td>
<td>0.6241</td>
<td>0.1699</td>
<td>0.341</td>
<td>0.6075</td>
<td>0.9989</td>
</tr>
<tr>
<td>beta</td>
<td>2.35E-4</td>
<td>9.23E-5</td>
<td>8.72E-5</td>
<td>2.23E-4</td>
<td>4.45E-4</td>
</tr>
<tr>
<td>p.repmin</td>
<td>0.3686</td>
<td>0.4824</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>p.repmax</td>
<td>0.1905</td>
<td>0.3927</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>p.repsum</td>
<td>0.4953</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>x.repmin</td>
<td>82.57</td>
<td>131.9</td>
<td>0.02948</td>
<td>31.4</td>
<td>466.0</td>
</tr>
<tr>
<td>x.repmax</td>
<td>11160.0</td>
<td>7002.0</td>
<td>3519.0</td>
<td>9392.0</td>
<td>2.9E+4</td>
</tr>
<tr>
<td>x.repsum</td>
<td>58390.0</td>
<td>26510.0</td>
<td>22660.0</td>
<td>53140.0</td>
<td>123200.0</td>
</tr>
<tr>
<td><strong>Inverse Gamma</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLL</td>
<td>179.0</td>
<td>1.045</td>
<td>178.0</td>
<td>178.7</td>
<td>181.8</td>
</tr>
<tr>
<td>alpha</td>
<td>0.5504</td>
<td>0.1476</td>
<td>0.3059</td>
<td>0.536</td>
<td>0.8788</td>
</tr>
<tr>
<td>beta</td>
<td>188.0</td>
<td>75.82</td>
<td>67.81</td>
<td>178.5</td>
<td>359.8</td>
</tr>
<tr>
<td>p.repmin</td>
<td>0.7805</td>
<td>0.4139</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>p.repmax</td>
<td>0.7222</td>
<td>0.4479</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>p.repsum</td>
<td>0.7859</td>
<td>0.4102</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>x.repmin</td>
<td>106.0</td>
<td>63.92</td>
<td>27.7</td>
<td>92.03</td>
<td>263.1</td>
</tr>
<tr>
<td>x.repmax</td>
<td>9.8E+11</td>
<td>9.9E+13</td>
<td>2482.0</td>
<td>50780.0</td>
<td>1.117E+8</td>
</tr>
<tr>
<td>x.repsum</td>
<td>6.4E+13</td>
<td>6.9E+15</td>
<td>16160.0</td>
<td>214500.0</td>
<td>6.082E+8</td>
</tr>
</tbody>
</table>

† Note the discussion in the main text concerning the existence of the theoretical posterior predictive moments for the nodes x.repmin, x.repmax, and x.repsum.
Table 2 (Continued)
Estimated Posterior Summary Statistics

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Mean</th>
<th>SD</th>
<th>2.5%</th>
<th>Median</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Loggamma</strong>†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLL</td>
<td>177.0</td>
<td>0.9604</td>
<td>176.0</td>
<td>176.7</td>
<td>179.6</td>
</tr>
<tr>
<td>alpha</td>
<td>18.52</td>
<td>5.826</td>
<td>9.211</td>
<td>17.72</td>
<td>31.67</td>
</tr>
<tr>
<td>beta</td>
<td>2.669</td>
<td>0.8512</td>
<td>1.311</td>
<td>2.553</td>
<td>4.575</td>
</tr>
<tr>
<td>p.repmin</td>
<td>0.7228</td>
<td>0.4476</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>p.repmax</td>
<td>0.5332</td>
<td>0.4989</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>p.repsum</td>
<td>0.6401</td>
<td>0.48</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>x.repmin</td>
<td>117.2</td>
<td>91.01</td>
<td>18.08</td>
<td>93.77</td>
<td>352.6</td>
</tr>
<tr>
<td>x.repmax</td>
<td>175700.0</td>
<td>4.432E+6</td>
<td>2672.0</td>
<td>17010.0</td>
<td>590800.0</td>
</tr>
<tr>
<td>x.repsum</td>
<td>608500.0</td>
<td>2.988E+7</td>
<td>17600.0</td>
<td>73250.0</td>
<td>1.432E+6</td>
</tr>
<tr>
<td><strong>Lognormal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLL</td>
<td>176.5</td>
<td>1.017</td>
<td>175.5</td>
<td>176.2</td>
<td>179.2</td>
</tr>
<tr>
<td>mu</td>
<td>6.933</td>
<td>0.365</td>
<td>6.22</td>
<td>6.934</td>
<td>7.66</td>
</tr>
<tr>
<td>tau</td>
<td>0.4105</td>
<td>0.1335</td>
<td>0.1969</td>
<td>0.3956</td>
<td>0.712</td>
</tr>
<tr>
<td>p.repmin</td>
<td>0.6247</td>
<td>0.4842</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>p.repmax</td>
<td>0.4294</td>
<td>0.495</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>p.repsum</td>
<td>0.5647</td>
<td>0.4958</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>x.repmin</td>
<td>109.6</td>
<td>102.0</td>
<td>6599</td>
<td>81.54</td>
<td>380.4</td>
</tr>
<tr>
<td>x.repmax</td>
<td>32870.0</td>
<td>161600.0</td>
<td>2764.0</td>
<td>12900.0</td>
<td>159100.0</td>
</tr>
<tr>
<td>x.repsum</td>
<td>1.02E+5</td>
<td>310700.0</td>
<td>18770.0</td>
<td>59560.0</td>
<td>397900.0</td>
</tr>
</tbody>
</table>

† Note the discussion in the main text concerning the existence of the theoretical posterior predictive moments for the nodes x.repmin, x.repmax, and x.repsum.
### Table 2 (Continued)

**Estimated Posterior Summary Statistics**

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>Estimates from WinBUGS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td><strong>Pareto</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NLL</td>
<td>176.7</td>
</tr>
<tr>
<td></td>
<td>alpha</td>
<td>3.484</td>
</tr>
<tr>
<td></td>
<td>theta</td>
<td>6827.0</td>
</tr>
<tr>
<td></td>
<td>p.repmin</td>
<td>0.6236</td>
</tr>
<tr>
<td></td>
<td>p.repmax</td>
<td>0.3059</td>
</tr>
<tr>
<td></td>
<td>p.repsum</td>
<td>0.5065</td>
</tr>
<tr>
<td></td>
<td>x.repmin</td>
<td>126.8</td>
</tr>
<tr>
<td></td>
<td>x.repmax</td>
<td>3.585E+7</td>
</tr>
<tr>
<td></td>
<td>x.repsum</td>
<td>4.725E+7</td>
</tr>
<tr>
<td><strong>Inverse Pareto</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NLL</td>
<td>177.1</td>
</tr>
<tr>
<td></td>
<td>alpha</td>
<td>1.536</td>
</tr>
<tr>
<td></td>
<td>theta</td>
<td>0.0002069</td>
</tr>
<tr>
<td></td>
<td>p.repmin</td>
<td>0.593</td>
</tr>
<tr>
<td></td>
<td>p.repmax</td>
<td>0.5238</td>
</tr>
<tr>
<td></td>
<td>p.repsum</td>
<td>0.6594</td>
</tr>
<tr>
<td></td>
<td>x.repmin</td>
<td>106.7</td>
</tr>
<tr>
<td></td>
<td>x.repmax</td>
<td>96650.0</td>
</tr>
<tr>
<td></td>
<td>x.repsum</td>
<td>219100.0</td>
</tr>
</tbody>
</table>

† Note the discussion in the main text concerning the existence of the theoretical posterior predictive moments for the nodes x.repmin, x.repmax, and x.repsum.
### Table 2 (Continued)  
Estimated Posterior Summary Statistics

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Model Estimates from WinBUGS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td><strong>Weibull</strong></td>
<td></td>
</tr>
<tr>
<td>NLL</td>
<td>176.8</td>
</tr>
<tr>
<td>alpha</td>
<td>0.7236</td>
</tr>
<tr>
<td>beta</td>
<td>0.006146</td>
</tr>
<tr>
<td>p.repmin</td>
<td>0.4091</td>
</tr>
<tr>
<td>p.repmax</td>
<td>0.2338</td>
</tr>
<tr>
<td>p.repsum</td>
<td>0.4978</td>
</tr>
<tr>
<td>x.repmin</td>
<td>84.74</td>
</tr>
<tr>
<td>x.repmax</td>
<td>12590.0</td>
</tr>
<tr>
<td>x.repsum</td>
<td>60370.0</td>
</tr>
<tr>
<td><strong>Inverse</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Weibull</strong></td>
<td></td>
</tr>
<tr>
<td>NLL</td>
<td>178.1</td>
</tr>
<tr>
<td>tau</td>
<td>0.6671</td>
</tr>
<tr>
<td>lambda</td>
<td>71.95</td>
</tr>
<tr>
<td>p.repmin</td>
<td>0.745</td>
</tr>
<tr>
<td>p.repmax</td>
<td>0.6937</td>
</tr>
<tr>
<td>p.repsum</td>
<td>0.7718</td>
</tr>
<tr>
<td>x.repmin</td>
<td>109.5</td>
</tr>
<tr>
<td>x.repmax</td>
<td>1.796E+8</td>
</tr>
<tr>
<td>x.repsum</td>
<td>8.665E+9</td>
</tr>
</tbody>
</table>

† Note the discussion in the main text concerning the existence of the theoretical posterior predictive moments for the nodes x.repmin, x.repmax, and x.repsum.
In the case of the Pareto model, for example, the restrictions $\alpha > 1$ and $\alpha > 2$ would need to be imposed in order to ensure the existence of a finite posterior predictive mean and variance, respectively (Klugman et al., 1998, page 575). The posterior probability attached to these restrictions can be checked by monitoring the frequency with which they arise in the MCMC simulation-based analysis of the unconstrained model. This procedure is illustrated in the analysis of the motor example in Section 4 of Scollnik (2000), and in the analysis of the grouped example ("Modeling Grouped Size of Loss Data in WinBUGS") in Section 7 of Scollnik (2001).

5 Implementing Predictive Inference

Suppose that $f(x|\psi)$ is the loss model responsible for generating the original observed losses, and that $g(y|\psi)$ is the loss model responsible for generating the losses that will be observed in the next period. Given the model parameters, $\psi$, we assume that the past and future losses are mutually independent. The predictive density $h(y|x)$ associated with a future loss is defined as the theoretical average of $g(y|\psi)$ taken with respect to the posterior distribution of the model parameters. That is,

$$h(y|x) = \int g(y|\psi)p(\psi|x)d\psi.$$ (11)

In Section 1.2, we discussed how to simulate a dependent sequence of random draws from a posterior distribution of model parameters, like $p(\psi|x)$, using WinBUGS. Let us assume that WinBUGS has been used in this manner to generate a sequence of such draws, which we will denote as $\psi^{(t)}$, for $t = m, \ldots, n$ ($m = 5,001$ and $n = 25,000$, in the example above). Provided that the model parameter vector $\psi$ was monitored in WinBUGS over these $n - m + 1$ iterations, we can click the Coda button on the Sample Monitor Tool dialog box to dump an ASCII (text) representation of its simulated values. These can be read into a spreadsheet or mathematical/statistical package and then used to estimate equation (11) on the basis of the ergodic sample average

$$h(y|x) \approx \frac{1}{n - m + 1} \sum_{t=m}^{n} g(y|\psi^{(t)}).$$ (12)

This is easily evaluated for any value(s) of $y$ desired. Note that the conditional model $g(y|\psi)$ needn’t be identical to the model $f(x|\psi)$ responsible for generating the original observed losses. In particular,
it may be modified in accordance with the effect(s) of inflation and/or policy limit modifications. For instance, if the original model was

\[ f(x|\alpha, \theta) \sim \text{Pareto}(\alpha, \theta) \]

and inflation through the next period was 100r percent, then the conditional loss model at the end of this period would be

\[ g(y|\alpha, \theta) \sim \text{Pareto}(\alpha, [1 + r]\theta), \]

as noted in Table 5.1 of Hogg and Klugman (1984, page 180). To simulate a variable representing a predictive draw from a loss model, use lines of code patterned after Section 3. For the Pareto loss model with inflation, for example, we would code

\[
y \sim \text{dgamma}(1, \text{tau.y}) \\
\text{tau.y} \sim \text{dgamma}(\alpha, \text{theta.y}) \\
\text{theta.y} \leftarrow (1 + r) \times \text{theta}
\]

The value of \( r \) would be set as a constant, loaded as part of the data list.

Scollnik (2002) provides a detailed illustration of predictive inference constructed via WinBUGS in the context of two possible regression models for a set of bivariate claims data (of the actual loss and allocated loss adjustment expense variety) and develops predictive forecasts of the total loss distributions under these two models for two different coverages. The reader is directed to this example for further insight into the predictive modeling process.

6 Concluding Remarks

This paper discusses how a number of different actuarial models for exact size-of-loss data can be implemented and analyzed in accordance with the Bayesian paradigm using WinBUGS. It does not, however, discuss how the models themselves are developed and selected for consideration, nor does it discuss how the likelihood function is specified when the sample data are incomplete—for instance, when there are left-truncated (due to deductibles) and right-censored losses (i.e., capped by policy limits).\(^1\) Provided that the resulting likelihood function can be defined using the mathematical operators available in WinBUGS, the

\(^1\)These issues are discussed in Klugman et al., (1998), and Guiahi (2001).
size of loss model always can be coded in WinBUGS by using the ‘ones’ or ‘zeroes’ tricks described in Section 2.4 above.

Another topic not discussed is the data preparation steps that may be required prior to model fitting. In practice, the data must be examined and corrected for data entry and reporting errors. Some of the data may belong to more current periods and some to older periods, so that some trending may be required to bring the data to current levels. In some contexts, it is also possible that some losses have not completely settled so that some adjustments to ultimate values also may be required. Some of these issues are discussed in McClenahan (1996) and Brown and Gottlieb (2001). It is also possible for some or all of these steps to be included as part of the complete probabilistic model. For instance, a random component representing missing data (e.g., reported but not settled claim amounts) could be included in the model and then the complete model be analyzed using a Bayesian method. See Ntzoufras and Dellaportas (2002) for a discussion and analysis of four such models.

References


Improving Mortality: A Rule of Thumb and Regulatory Tool

John H. Pollard*

Abstract†

We develop a simple exact formula for determining cohort life expectancies under constant continuous uniform improvement in mortality using only a cross-sectional (period) Gompertz life table for the lives concerned and a simple approximation applicable to all life tables. The present values of annuities for such lives can be determined simply and accurately across the whole age span.

Key words and phrases: Gompertz, life annuities, cohort mortality, cross-sectional mortality, period life table

1 Introduction

The latter part of the twentieth century has seen mortality rates in many developed countries improving steadily over prolonged periods.

*John H. Pollard completed his B.Sc. degree at the University of Sydney with first class honors in 1963 and his Ph.D. at the University of Cambridge in 1967. After a post-doctoral year at the University of Chicago, he returned to Australia, becoming Professor and Chair of Actuarial Studies at Macquarie University in 1977.

Professor Pollard is the author of seven books on mortality, statistics, life insurance, and non-life insurance. His research publications range from mortality and population mathematics to stochastic interest models and reserving in non-life insurance. A past president of the Institute of Actuaries of Australia and of the Statistical Society of Australia, he has advised a number of major insurance companies, the World Health Organization, and the Australian Government and was a director of Swiss Re Australia from 1983 to 2001. He chaired the Mortality Committee of the International Union for the Scientific Study of Population from 1979 to 1983.

Dr. Pollard's address is: Division of Economic and Financial Studies, Macquarie University, NSW 2109, AUSTRALIA. Internet address: jpollard@efs.mq.edu.au

†The author would like to thank the anonymous referees for their suggestions that have improved the overall layout of this paper.
Similar improvements have been observed with life insurance data. It is interesting to inquire about the effects such steady improvements have on expectation of life and make comparisons with the life expectancies reported in commonly prepared cross-sectional life tables. For insurance companies offering life annuities, allowance must be made for future mortality improvements to avoid the adverse financial consequences such improvements would otherwise have on company funds.

In a changing mortality environment, the value of an immediate annuity, for example, must depend inter alia on the age of the proposer and the year the annuity commences. Different values are therefore required according to each of these two variables, and a separate underlying life table must be computed for each combination. A simple change in the annual improvement rate or use of an updated base cross-sectional table necessitates a complete recalculation of all the underlying life tables and values. For a regulator, checking values used can be tedious.

Under the Gompertz law of mortality, a simple exact formula can be obtained for a cohort expectation of life at any age under constant uniform mortality improvement. An approximate rule derived from this formula, which is surprisingly accurate across a wide range of life tables, may be expressed as follows.

\[
\text{cohort } \hat{e}_x = \frac{9}{9 - 100r} \times \text{base } \hat{e}_{x+150r} \tag{1}
\]

where the annual rate of mortality improvement is 100r% and 0 \leq 100r \leq 3. A precise definition of "constant uniform mortality improvement" is given later.

Accurate approximations for the associated continuous life annuities can be obtained using

\[
\text{cohort } \bar{a}_x = \text{base } \bar{a}_x \times \left( \frac{\bar{d}_n}{\bar{d}_v} \right) \tag{2}
\]

where \(n = \text{cohort } \hat{e}_x\) and \(v = \text{base } \hat{e}_x\).

The accuracy of the above formulae is such that errors introduced in using them are relatively minor compared with the uncertainty in selecting an assumed annual rate of mortality improvement. The formulae have the important advantage that a special generation life table does not need to be prepared for each and every age in a particular base year.
2 Expectation of Life Under Gompertz

The force of mortality under the well-known Gompertz law of mortality may be expressed in the following form:

\[ \mu_x = k \exp[k(x - m)] \]  

where \( m \) is the mode of the curve of deaths and \( k \) reflects the rate at which mortality increases with age. (See, for example, Benjamin and Pollard, 1993, page 298; Pollard, 1991.) The complete expectation of life at age \( x \) is

\[ \hat{e}_x = \frac{1}{k} \exp\left(\frac{\mu_x}{k}\right) E_1(\frac{\mu_x}{k}) \]  

where \( E_1(\cdot) \), an exponential integral, is defined by

\[ E_1(t) = \int_1^{\infty} \frac{e^{-tu}}{u} du. \]

A proof of this result is given in the appendix, where a power series expansion for \( E_1(\cdot) \) is provided.

For the purposes of this paper, the important thing to note in respect of equation (4) is that the complete expectation of life is simply equal to a function of the single variable \( \mu_x/k \) divided by \( k \) and may be written

\[ \hat{e}_x = \frac{f(\mu_x/k)}{k}. \]

3 Improving Mortality

If mortality is improving at a constant instantaneous rate of \( r \) per annum at all ages and at all future times, a life age \( x \) in the base year and subject to a force of mortality \( \mu_x \) at that time, will experience a force of mortality \( t \) years later of \( \mu_{x+t} \exp(-rt) \). If the base cross-sectional table follows the Gompertz law of equation (3), \( \mu_{x+t} = \mu_x \exp(kt) \), then the force of mortality \( t \) years after the base year for the life under consideration is

\[ \mu_{x+t} e^{-rt} = \mu_x e^{(k-r)t}. \]

In other words, the cohort of lives age \( x \) in the base year will experience Gompertz mortality over their subsequent lifetimes with mortality
increasing with age at a rate of \( k - r \) compared with \( k \) in the base cross-sectional table. The cohort expectation of life is therefore

\[
\hat{r}e_x = \frac{f\left(\frac{\mu_x}{k-r}\right)}{k-r}.
\]

If we select an age \( x + n \) in the cross-sectional life table such that

\[
\frac{\mu_{x+n}}{k} = \frac{\mu_x}{k-r},
\]

then

\[
(k - r)\hat{r}e_x = f\left(\frac{\mu_x}{(k-r)}\right) = f\left(\frac{\mu_{x+n}}{k}\right) = k\hat{r}e_{x+n}.
\]

To obtain the improving mortality expectation of life, \( r\hat{r}e_x \), all we have to do is multiply the cross-sectional expectation of life at age \( x + n \) by \( k/(k - r) \). Because in the base Gompertz life table \( \mu_{x+n} = \mu_x \exp(kn) \), the \( n \) required to satisfy equation (5) is

\[
n = \frac{1}{k} \ln \left[ \frac{k}{(k-r)} \right].
\]

All the above formulae are exact. We now note that for most cross-sectional life tables, a Gompertz approximation to the adult age mortality will indicate a value of \( k \) in the neighborhood of 0.09, in which case

\[
n = \frac{1}{0.09} \ln \left( \frac{0.09}{0.09-r} \right) \approx 130r
\]

which is exact when \( r = 0 \). For \( r = 0.01, 0.02, \) and 0.03 and \( k = 0.09 \), equation (7) provides values of \( n \) of 1.31, 2.79, and 4.51, respectively. As a rule of thumb, therefore, we simply add 1.5 to the age for each percentage point in the annual mortality improvement rate (and propor
ta between). The cross-sectional expectation of life at the resultant age is then multiplied by \( 9/(9 - 100r) \).
4 The Accuracy of the Rule of Thumb

The expectations of life at ages 30-35 in English Life Table 15 (Males) are shown in Table 1. As an example, let us imagine that this is the base cross-sectional life table and we require the expectation of life for a male age 30 at the time of the base table under a regime of continually improving mortality at a rate of 2% per annum.

<table>
<thead>
<tr>
<th>Age $x$</th>
<th>30</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_x$</td>
<td>44.88</td>
<td>43.92</td>
<td>42.96</td>
<td>42.01</td>
<td>41.05</td>
<td>40.09</td>
</tr>
</tbody>
</table>

According to the rule of thumb of the previous section, we simply take the expectation of life at age 33 according to the base table and multiply by $9/7$. The approximate expectation of life is $42.01 \times 9/7 = 54.01$. Exact calculation using a specially prepared cohort life table produces a value of 54.10. Given the large difference between the cross-sectional $e_x(44.88)$ and the cohort $e_x(54.10)$, the rule of thumb produces a remarkably accurate approximation (only -0.2% error).

Further comparisons are presented in Tables 2, 3, and 4 using English Life Table 15 (Male), English Life Table 15 (Female), and the base cross-sectional table underlying the $a(90)$ (Male) Life Table, none of which has a strict Gompertz shape. Interestingly, the rule of thumb works well, even at the juvenile ages, except when the annual mortality improvement rate is high. Although juvenile mortality does not follow the Gompertz pattern, it is now so low in developed-country populations that further mortality improvements affecting the complete expectation of life are largely concentrated in the later adult ages.

Continuous long-term mortality improvements at rates approaching 2% would be exceptional; the fact that the rule of thumb becomes progressively less accurate for values of $r$ above 0.02 is therefore of little concern. Actuaries will usually be interested in monetary functions associated with the improving mortality, particularly annuities. The most accurate simple way of evaluating the latter approximately is to apply equation (2); that is, to multiply the life annuity in the base cross-sectional table by the ratio of the continuous annuity certain with a term equal to the approximated expectation of life under the improving mortality regime to the continuous annuity certain with its term equal to the base table expectation of life. The examples in Tables 2,
3, and 4 reveal that the approximations are remarkably accurate across the entire age range.

5 Concluding Remarks

The rule of thumb we have described for evaluating expectation of life and annuity values under continuously improving mortality regimes provides surprisingly accurate approximations for these life table and monetary functions. Attempts at gaining greater accuracy by using values of \( k \) derived from the actual base life table itself (e.g., setting \( k \) equal to the force of mortality at the mode of the curve of deaths and using a value for the age adjustment calculated using the exact equation (5)) fail to produce worthwhile improvements in the accuracy of calculated life expectancies and, given the uncertainty in the long-term mortality improvement rate, the additional work involved in making the more refined calculations is not justified.

Where the base life table relates to a mortality experience \( s \) years earlier rather than the date when the life of interest was age \( x \), and where there has been continuous mortality improvements at rate \( p \) over the intervening period, the Gompertz law with \( k = 0.09 \) indicates that the life should be treated as being of age \( x - ps/0.09 \). This well-known adjustment should work well at all except the juvenile ages.

Regulators concerned with the solvency of life insurance companies need simple effective rules to apply to companies and their products. The rule we propose has potential applications for these purposes.

References


Appendix

As is well known, the probability of survival from age $x$ to age $x + t$ is

$$t p_x = \exp \left( - \int_0^t \mu_{x+s} \, ds \right),$$

and the complete expectation of life at age $x$ is

$$\hat{e}_x = \int_0^\infty t p_x \, dt.$$

Using the Gompertz force of mortality given in equation (3) yields

$$t p_x = \exp \left[ - \left( \frac{\mu_x}{k} \right) (e^{kt} - 1) \right]. \quad (8)$$

Substituting $z = e^{kt}$ and integrating with respect to $z$ yields

$$\hat{e}_x = \frac{e^{\mu_x/k}}{k} \int_1^\infty \frac{1}{z} e^{-z \mu_x/k} \, dz = \frac{e^{\mu_x/k}}{k} E_1 (\mu_x / k) \quad (9)$$

where $E_1 (t)$ is the exponential integral defined by

$$E_1 (t) = \int_1^\infty \frac{e^{-tu}}{u} \, du = -\gamma - \ln t - \sum_{n=1}^{\infty} \frac{(-t)^n}{n \, n!}$$

and $\gamma = 0.5772157...$ is Euler’s constant. See, for example, CRC Standard Mathematical Tables page 315 or Abramowitz and Stegun (1965, Chapter 5) for more on this integral. Abramowitz and Stegun also provide accurate approximations to $E_1 (t)$. 
Table 2
Effect of Continually Improving Mortality
On English Life Table Number 15 (Male)
Life Expectancies and Life Annuities

<table>
<thead>
<tr>
<th>Age</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELT15 Table</td>
<td>73.41</td>
<td>69.13</td>
<td>64.20</td>
<td>54.45</td>
<td>44.88</td>
<td>17.85</td>
<td>3.51</td>
</tr>
<tr>
<td>Improvement 1% pa</td>
<td>81.69</td>
<td>76.11</td>
<td>70.55</td>
<td>59.65</td>
<td>48.87</td>
<td>18.82</td>
<td>3.58</td>
</tr>
<tr>
<td>Approximation</td>
<td>81.04</td>
<td>76.11</td>
<td>70.57</td>
<td>59.64</td>
<td>48.95</td>
<td>19.03</td>
<td>3.59</td>
</tr>
<tr>
<td>Exact Value</td>
<td>0.8</td>
<td>0.0</td>
<td>-0.0</td>
<td>0.0</td>
<td>-0.2</td>
<td>-1.1</td>
<td>-0.4</td>
</tr>
<tr>
<td>Improvement 2% pa</td>
<td>91.40</td>
<td>85.08</td>
<td>78.72</td>
<td>66.33</td>
<td>54.01</td>
<td>20.11</td>
<td>3.69</td>
</tr>
<tr>
<td>Approximation</td>
<td>90.24</td>
<td>84.78</td>
<td>78.55</td>
<td>66.21</td>
<td>54.10</td>
<td>20.46</td>
<td>3.69</td>
</tr>
<tr>
<td>Exact Value</td>
<td>1.2</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
<td>-0.2</td>
<td>-1.7</td>
<td>0.0</td>
</tr>
<tr>
<td>Improvement 3% pa</td>
<td>104.43</td>
<td>97.03</td>
<td>89.63</td>
<td>75.24</td>
<td>60.85</td>
<td>21.91</td>
<td>3.92</td>
</tr>
<tr>
<td>Approximation</td>
<td>98.58</td>
<td>92.96</td>
<td>86.47</td>
<td>73.30</td>
<td>60.03</td>
<td>22.19</td>
<td>3.80</td>
</tr>
<tr>
<td>Exact Value</td>
<td>5.9</td>
<td>4.4</td>
<td>3.7</td>
<td>2.6</td>
<td>1.4</td>
<td>-1.3</td>
<td>3.1</td>
</tr>
</tbody>
</table>

Continuous Life Annuity at 5% pa Interest

<table>
<thead>
<tr>
<th>Age</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELT15 Table</td>
<td>19.59</td>
<td>19.54</td>
<td>19.30</td>
<td>18.65</td>
<td>17.69</td>
<td>11.08</td>
<td>3.06</td>
</tr>
<tr>
<td>Improvement 1% pa</td>
<td>19.78</td>
<td>19.74</td>
<td>19.53</td>
<td>18.97</td>
<td>18.08</td>
<td>11.45</td>
<td>3.12</td>
</tr>
<tr>
<td>Approximation</td>
<td>19.77</td>
<td>19.75</td>
<td>19.54</td>
<td>18.95</td>
<td>18.07</td>
<td>11.46</td>
<td>3.12</td>
</tr>
<tr>
<td>Exact Value</td>
<td>0.0</td>
<td>-0.0</td>
<td>-0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>-0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Improvement 2% pa</td>
<td>19.92</td>
<td>19.92</td>
<td>19.75</td>
<td>19.27</td>
<td>18.49</td>
<td>11.91</td>
<td>3.20</td>
</tr>
<tr>
<td>Approximation</td>
<td>19.92</td>
<td>19.94</td>
<td>19.76</td>
<td>19.25</td>
<td>18.44</td>
<td>11.87</td>
<td>3.19</td>
</tr>
<tr>
<td>Exact Value</td>
<td>0.0</td>
<td>-0.1</td>
<td>-0.0</td>
<td>0.1</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Improvement 3% pa</td>
<td>20.03</td>
<td>20.06</td>
<td>19.93</td>
<td>19.55</td>
<td>18.90</td>
<td>12.51</td>
<td>3.38</td>
</tr>
<tr>
<td>Approximation</td>
<td>20.04</td>
<td>20.08</td>
<td>19.94</td>
<td>19.51</td>
<td>18.80</td>
<td>12.31</td>
<td>3.26</td>
</tr>
<tr>
<td>Exact Value</td>
<td>-0.0</td>
<td>-0.1</td>
<td>-0.0</td>
<td>0.2</td>
<td>0.5</td>
<td>1.6</td>
<td>3.7</td>
</tr>
</tbody>
</table>
### Table 3
Effect of Continually Improving Mortality
On English Life Table Number 15 (Female)
Life Expectancies and Life Annuities

<table>
<thead>
<tr>
<th>Age</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expectation of Life</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELT15 Table</td>
<td>78.96</td>
<td>74.56</td>
<td>69.61</td>
<td>59.75</td>
<td>49.94</td>
<td>22.08</td>
<td>4.35</td>
</tr>
<tr>
<td>Improvement 1% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>87.73</td>
<td>82.21</td>
<td>76.64</td>
<td>65.56</td>
<td>54.53</td>
<td>23.46</td>
<td>4.42</td>
</tr>
<tr>
<td>Exact Value</td>
<td>86.82</td>
<td>81.81</td>
<td>76.29</td>
<td>65.30</td>
<td>54.38</td>
<td>23.58</td>
<td>4.48</td>
</tr>
<tr>
<td>Error (%)</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>-0.5</td>
<td>-1.3</td>
</tr>
<tr>
<td>Improvement 2% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>98.39</td>
<td>92.05</td>
<td>85.68</td>
<td>73.04</td>
<td>60.44</td>
<td>25.27</td>
<td>4.57</td>
</tr>
<tr>
<td>Exact Value</td>
<td>96.07</td>
<td>90.54</td>
<td>84.42</td>
<td>72.15</td>
<td>59.91</td>
<td>25.40</td>
<td>4.62</td>
</tr>
<tr>
<td>Error (%)</td>
<td>2.4</td>
<td>1.7</td>
<td>1.5</td>
<td>1.2</td>
<td>0.9</td>
<td>-0.5</td>
<td>-1.1</td>
</tr>
<tr>
<td>Improvement 3% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>112.57</td>
<td>105.15</td>
<td>97.73</td>
<td>83.00</td>
<td>68.32</td>
<td>27.74</td>
<td>4.83</td>
</tr>
<tr>
<td>Exact Value</td>
<td>103.82</td>
<td>98.28</td>
<td>92.02</td>
<td>79.20</td>
<td>66.03</td>
<td>27.59</td>
<td>4.78</td>
</tr>
<tr>
<td>Error (%)</td>
<td>8.4</td>
<td>7.0</td>
<td>6.2</td>
<td>4.8</td>
<td>3.5</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>Continuous Life Annuity at 5% pa Interest</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELT15 Table</td>
<td>19.81</td>
<td>19.78</td>
<td>19.60</td>
<td>19.08</td>
<td>18.28</td>
<td>12.69</td>
<td>3.70</td>
</tr>
<tr>
<td>Improvement 1% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>19.96</td>
<td>19.95</td>
<td>19.80</td>
<td>19.35</td>
<td>18.63</td>
<td>13.12</td>
<td>3.75</td>
</tr>
<tr>
<td>Exact Value</td>
<td>19.95</td>
<td>19.95</td>
<td>19.80</td>
<td>19.35</td>
<td>18.61</td>
<td>13.09</td>
<td>3.79</td>
</tr>
<tr>
<td>Error (%)</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>-1.1</td>
</tr>
<tr>
<td>Improvement 2% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>20.07</td>
<td>20.09</td>
<td>19.97</td>
<td>19.60</td>
<td>18.98</td>
<td>13.63</td>
<td>3.87</td>
</tr>
<tr>
<td>Exact Value</td>
<td>20.08</td>
<td>20.10</td>
<td>19.97</td>
<td>19.59</td>
<td>18.94</td>
<td>13.53</td>
<td>3.88</td>
</tr>
<tr>
<td>Error (%)</td>
<td>-0.0</td>
<td>-0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.7</td>
<td>-0.3</td>
</tr>
<tr>
<td>Improvement 3% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>20.16</td>
<td>20.19</td>
<td>20.11</td>
<td>19.82</td>
<td>19.32</td>
<td>14.27</td>
<td>4.06</td>
</tr>
<tr>
<td>Error (%)</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.4</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Table 4
Effect of Continually Improving Mortality
On a(90) Base Life Expectancies and Life Annuities

<table>
<thead>
<tr>
<th>Age</th>
<th>20</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Table</td>
<td>56.03</td>
<td>46.34</td>
<td>19.11</td>
<td>4.04</td>
</tr>
<tr>
<td>Improvement 1% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>61.35</td>
<td>50.50</td>
<td>20.26</td>
<td>4.14</td>
</tr>
<tr>
<td>Exact Value</td>
<td>61.59</td>
<td>50.75</td>
<td>20.46</td>
<td>4.16</td>
</tr>
<tr>
<td>Error (%)</td>
<td>-0.4</td>
<td>-0.5</td>
<td>-1.0</td>
<td>-0.5</td>
</tr>
<tr>
<td>Improvement 2% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>68.26</td>
<td>55.84</td>
<td>21.77</td>
<td>4.30</td>
</tr>
<tr>
<td>Exact Value</td>
<td>68.41</td>
<td>56.18</td>
<td>22.08</td>
<td>4.28</td>
</tr>
<tr>
<td>Error (%)</td>
<td>-0.2</td>
<td>-0.6</td>
<td>-1.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Improvement 3% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>77.48</td>
<td>62.96</td>
<td>23.85</td>
<td>4.56</td>
</tr>
<tr>
<td>Exact Value</td>
<td>75.31</td>
<td>62.11</td>
<td>24.01</td>
<td>4.42</td>
</tr>
<tr>
<td>Error (%)</td>
<td>2.9</td>
<td>1.4</td>
<td>-0.6</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Continuous Life Annuity at 5% pa Interest

<table>
<thead>
<tr>
<th>Age</th>
<th>20</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Table</td>
<td>18.80</td>
<td>17.87</td>
<td>11.53</td>
<td>3.47</td>
</tr>
<tr>
<td>Improvement 1% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>19.10</td>
<td>18.25</td>
<td>11.94</td>
<td>3.55</td>
</tr>
<tr>
<td>Exact Value</td>
<td>19.10</td>
<td>18.25</td>
<td>11.93</td>
<td>3.55</td>
</tr>
<tr>
<td>Error (%)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Improvement 2% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>19.39</td>
<td>18.64</td>
<td>12.44</td>
<td>3.67</td>
</tr>
<tr>
<td>Exact Value</td>
<td>19.38</td>
<td>18.62</td>
<td>12.36</td>
<td>3.63</td>
</tr>
<tr>
<td>Error (%)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.6</td>
<td>1.1</td>
</tr>
<tr>
<td>Improvement 3% pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation</td>
<td>19.63</td>
<td>19.03</td>
<td>13.08</td>
<td>3.87</td>
</tr>
<tr>
<td>Exact Value</td>
<td>19.63</td>
<td>18.96</td>
<td>12.82</td>
<td>3.72</td>
</tr>
<tr>
<td>Error (%)</td>
<td>0.0</td>
<td>0.4</td>
<td>2.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>
Further Remarks on Risk Sources Measuring: The Case of a Life Annuity Portfolio

Mariarosaria Coppola,* Emilia Di Lorenzo† and Marilena Sibillo‡

Abstract§

The paper considers a model that allows the actuary to measure the riskiness connected to the randomness of projected mortality tables in evaluating a portfolio of life annuities, obtaining a measure to reflect the risk associated with the randomness of the projection. The coherence of the risk parameters with the specific nature of the considered risk sources is also discussed.

Numerical examples illustrate the results, showing the importance of the risk components in terms of the number of policies and comparing measure tools obtained by means of two procedures.

Key words and phrases: investment risk, insurance risk, longevity risk, random projected mortality tables

*Mariarosaria Coppola, Ph.D., is assistant professor of financial mathematics at the University of Salerno. She is author of several papers in actuarial mathematics.

Dr. Coppola’s address is: Universita’ di Salerno, Dipartimento di Scienze Economiche e Statistiche via Ponte don Melillo, 84084 Fisciano (Salerno), ITALY. Internet address: mrcoppola@unisa.it

†Emilia Di Lorenzo is professor of financial mathematics at the University of Naples. She is author of several papers in actuarial mathematics and related fields.

Dr. Di Lorenzo’s address is: Dipartimento di Matematica e Statistica, Facolta’ di Economia Universita’ degli Studi di Napoli “Federico II”, via Cintia, Complesso Monte S.Angelo 80126 Napoli, ITALY. Internet address: diloremi@unina.it

‡Marilena Sibillo, Ph.D., is associate professor of financial mathematics at the University of Salerno. She is author of several papers in actuarial mathematics and related fields.

Dr. Sibillo’s address is: Universita’ di Salerno, Dipartimento di Scienze Economiche e Statistiche via Ponte don Melillo, 84084 Fisciano (Salerno), ITALY. Internet address: msibillo@unisa.it

§The authors thank the anonymous referees and the editor for their comments that have improved the presentation of this paper.
1 Introduction

As a life insurance business can be viewed as a dynamic risk process, actuaries must have at their disposal tools to control the various factors affecting this risk process. The two most important risks life insurers face are investment risk and demographic risk. The investment risk is the risk to the market value of the insurer's assets due to the random movements of the financial market. [See, for example, Beekman and Fuelling (1990) and (1991), Frees (1998), Parker (1994a) and (1994b), and Zaks (2001).] The demographic risk is the risk of premature death (in the case of life insurance) or excessive longevity (in the case of annuities).1

Focusing on the annuity business, demographic risk consists of two components: (i) the insurance risk, which is due to the random deviations of the number of deaths from their expected values; and (ii) the longevity risk, which is due to improvements in mortality rates.2 The longevity risk combines the effects of two phenomena: (i) "rectangularization," which refers to the higher concentration of deaths around the mode of the curve of deaths; and (ii) "expansion," which refers to the increase in the mode of the curve of deaths over time. Insurance risk can be viewed as a pooling risk because it decreases as the number of policies in-force increases, while longevity risk is a non-pooling risk because it is not affected by the number of policies in-force.

Given the potentially adverse impact of the longevity risk on the stability of a portfolio of annuities, mortality tables used to value annuity contracts must take into account the anticipated improvements in future mortality, i.e., a mortality projection. In other words, tables should be constructed based on anticipated decreases in future mortality rates. Failure to include future improvements in mortality could cause a significant underestimation of future obligations.

Though several authors [e.g., Pitacco (1997), Marocco et al., (1998), Olivieri (1998), Olivieri et al., (1999) and Coppola et al., (2000)] have studied the longevity risk, the Coppola et al., (2000) paper is of particu-

---

1In the case of insurance with a death benefit, the effect of improving mortality is to delay the time of death thus postponing the payment of the death benefit. In the case of annuities, however, the effect of improving mortality is to increase the duration of the insurer's payments to the annuitant.

2Due to the constant advances in public health and safety and better nutrition, people in most societies around the world are living longer and healthier lives. In the developed economies, there is a trend of increasing sales of annuity contracts to pay for retirement. This combination of increasing longevity and increasing sales of annuity contracts requires actuaries to have a deeper understanding of mortality trends, and hence the longevity risk. [See, also, UP-Task Force (1996).]
lar interest to us. Coppola et al., identify and characterize the two risk factors for a life annuity portfolio and analyze the demographic risk by taking into account only the contribution of the longevity risk. They present a model of the global riskiness of the portfolio and provide an expression for the contribution of each risk component to the value of the entire portfolio.

Our current paper is a follow-up to the Coppola et al., (2000) paper. It uses a stochastic framework for the interest rates and a deterministic framework for the longevity risk. The basic underlying distribution of an individual's future lifetime is Weibull. The longevity risk is modeled by constructing three different projected mortality tables, each representing a particular scenario for improving mortality. This leads to a more general model of the present value of the portfolio. We propose measurement tools for determining the riskiness of the average cost per policy due to the randomness in choosing projected mortality tables, while still taking into account the effect of interest rates and random mortality deviations.

This paper is organized as follows: In Section 2 we review some valuation results already presented by Coppola et al., (2000) and introduce the stochastic model for interest rates. Section 3 presents equations to decompose the total riskiness taking into consideration the randomness of the projection together with the interest randomness and the random mortality deviations. In particular we obtain different equations following two procedures, both based on a risk decomposition using variance, but conditioning on different risk sources. The asymptotic behavior of the obtained risk parameters, when the portfolio's size tends to infinity, is also discussed. In Section 4 the main results of the paper are complemented by some numerical examples.

2 Portfolio Valuations

Let us consider a portfolio consisting of $c$ individuals age exactly $x$ each of whom has a whole life annuity-immediate policy (or contract) paying $1 per year for life. Let us introduce the following notation:

$$K_i(x) = \text{Curtailed future lifetime of the } i^{th} \text{ policy;}$$
$$Z_i = \text{Present value of the annuity contract for the } i^{th} \text{ policy;}$$
$$Z(c) = \text{Present value of the entire annuity portfolio of } c \text{ contracts;}$$
and
$$\delta(s) = \text{Stochastic force of interest at time } s \geq 0.$$
It follows that

\[ Z_i = \sum_{h=1}^{K_i(x)} \exp(-y(h)) \]  

(1)

where

\[ y(h) = \int_0^h \delta(s) ds; \]  

(2)

in addition,

\[ Z(c) = \sum_{i=1}^c Z_i. \]  

(3)

As in Coppola et al., (2000), the following assumptions are made:

(i) The times of death $K_1(x), K_2(x), \ldots, K_c(x)$ are mutually independent and identically distributed random variables;

(ii) Given the sequence $y(1), y(2), \ldots$, the $Z_1, Z_2, \ldots, Z_c$ are independent and identically distributed; and

(iii) The times of death $K_1(x), K_2(x), \ldots, K_c(x)$ and the interest rate process $\delta(s)$ are mutually independent.

The first two moments are given in Coppola et al., (2000) and are as follows:

\[
\mathbb{E}[Z_i \mid \{y(h)\}_{h=1}^\infty] = \sum_{h=1}^\infty h p_x e^{-y(h)}
\]

\[
\mathbb{E}[Z_i] = \sum_{h=1}^\infty h p_x \mathbb{E}[e^{-y(h)}]
\]

\[
\mathbb{E}[Z_i^2 \mid \{y(h)\}_{h=1}^\infty] = \sum_{h=1}^\infty h p_x e^{-2y(h)} + 2 \sum_{h=2}^\infty h p_x \sum_{r=1}^{h-1} e^{-y(r)-y(h)}
\]

\[
\mathbb{E}[Z_i^2] = \sum_{h=1}^\infty h p_x \mathbb{E}[e^{-2y(h)}] + 2 \sum_{h=2}^\infty h p_x \sum_{r=1}^{h-1} \mathbb{E}[e^{-y(r)-y(h)}]
\]

\[
\mathbb{E}[Z(c)] = c \sum_{h=1}^\infty h p_x \mathbb{E}[e^{-y(h)}], \quad \text{and}
\]

\[
\mathbb{E}[Z(c)^2] = \sum_{i=1}^c \mathbb{E}[Z_i^2] + \sum_{i,j=1, i \neq j}^c \mathbb{E}[Z_i Z_j].
\]
By assumptions (i), (ii), and (iii), we can write [see, Coppola et al., (2000)]
\[
E[Z_i Z_j] = E[E[Z_i Z_j | \{y(h)\}_{h=1}^\infty |] \sum_{h=1}^\infty \sum_{k=1}^\infty h p_x k p_x E[e^{-y(h)-y(k)}].
\]

It follows that
\[
E[Z(c)^2] = c E[Z_i^2] + \sum_{i,j=1}^c \sum_{h=1}^\infty \sum_{k=1}^\infty h p_x k p_x E[e^{-y(h)-y(k)}]
\]
\[
= c E[Z_i^2] + c(c - 1) \sum_{h=1}^\infty \sum_{k=1}^\infty h p_x k p_x E[e^{-y(h)-y(k)}].
\]

The first two moments of the average cost per policy of the portfolio under consideration, $Z(c)/c$, are
\[
E\left[\frac{Z(c)}{c}\right] = \sum_{h=1}^\infty h p_x E[e^{-y(h)}]
\]
\[
E\left[\left(\frac{Z(c)}{c}\right)^2\right] = \frac{1}{c} E[Z_i^2] + \frac{c - 1}{c} \sum_{h=1}^\infty \sum_{k=1}^\infty h p_x k p_x E[e^{-y(h)-y(k)}].
\]

### 2.1 The Stochastic Interest Rate Environment

In order to get a realistic description of the insurance environment, we consider the risk arising from fluctuations of the rate of interest process $\delta(t)$. The interest rate process is viewed as a sum of two components: a deterministic component, $i(t)$, which can be estimated on the basis of the company's investment policy, and a stochastic component, $X(t)$, which describes the deviations of the interest rate process from its expected values. Thus
\[
\delta(t) = i(t) + X(t).
\]

The $X(t)$ process is assumed to be an Ornstein-Uhlenbeck process, with parameters $\beta > 0$ and $\sigma > 0$, and initial position $X(0) = 0$. Ornstein-Uhlenbeck processes are characterized by the following stochastic differential equation
\[ dX(t) = -\beta X(t)dt + \sigma dW(t) \]

where \( W(t) \) is a standard Wiener process. See, for example, Arnold (1974) or Gard (1998) for more on stochastic differential equations.

The present value at time 0 of a payment of one monetary unit at time \( t \) is given by

\[ v(t)F(t) = e^{-\gamma(t)} = e^{-\int_0^t (i(s)+X(s))ds} \]

where

\[ v(t) = e^{-\int_0^t i(s)ds} \quad \text{and} \quad F(t) = e^{-\int_0^t X(s)ds} \]

are the deterministic and stochastic discounting factors, respectively. Using the fact that \( F(t) \) is log-normal and \( \mathbb{E}[X(t)] = 0 \), Coppola et al., (2000) demonstrate that:

\begin{align*}
\mathbb{E}[F(t)] &= e^{\frac{1}{2} \phi(t)} \\
\text{Var}[F(t)] &= e^{\phi(t)}[e^{\phi(t)} - 1] \\
\text{Cov}[F(h), F(k)] &= e^{\frac{1}{2} [\phi(h)+\phi(k)]}[e^{\phi(h,k)} - 1]
\end{align*}

(7) \hspace{1cm} (8) \hspace{1cm} (9)

where

\[ \phi(t) = \text{Var}\left[\int_0^t X(s)ds\right], \quad \text{and} \]

\[ \Phi(h, k) = \text{Cov}\left[\int_0^h X(s)ds, \int_0^k X(s)ds\right] \]

(10) \hspace{1cm} (11)

is the autocovariance function of \( F(t) \).

3 A Measure of Projection Randomness

To take into account the influence of the randomness in projections, we use a well known variance decomposition equation for estimating the importance of different risk sources in portfolio valuations.\(^3\) Coppola et al., (2000) obtain measurement tools for estimating the impact of some risk components, i.e., the insurance and the investment risks, when different projected mortality tables are used.

3.1 Conditioning on the Random Survival Function

Let $P$ denote the survival function used to construct the survival probabilities in the projected table. The variance of the average cost per policy can then be split in two components:

$$\text{Var}\left[\frac{Z(c)}{c}\right] = \text{Var}\left[\mathbb{E}\left[\frac{Z(c)}{c} \mid P\right]\right] + \mathbb{E}\left[\text{Var}\left[\frac{Z(c)}{c} \mid P\right]\right]. \tag{12}$$

The first term on the right side of equation (12) is a measure of the variability of $Z(c)/c$ due to the effect of the randomness of the projection. The second term measures the effect of the other risk components (random interest rates and random mortality deviations), the effects of the projection randomness having been averaged out. As

$$\text{Var}\left[\mathbb{E}\left[\frac{Z(c)}{c} \mid P\right]\right] = \text{Var}\left[\mathbb{E}\left[\frac{1}{c} \sum_{i=1}^{c} Z_i \mid P\right]\right] = \text{Var}\left[\mathbb{E}\left[\frac{K_i(x)}{c} \sum_{h=1}^{\infty} e^{-y(h)} \mid P\right]\right],$$

it is clear that $\text{Var}\left[\mathbb{E}[Z(c)/c \mid P]\right]$ is a measure of a systematic risk, which is independent of the size of the portfolio $c$. This agrees with the nature of the risk due to the randomness of projection. Thus we have the following definition

**Definition 1A:** $\text{Var}\left[\mathbb{E}[Z(c)/c \mid P]\right]$ is a measure of the projection risk.

The second term on the right side of equation (12) can again split as follows

$$\mathbb{E}\left[\text{Var}\left[\frac{Z(c)}{c} \mid P\right]\right] = \mathbb{E}\left[\text{Var}\left[\mathbb{E}\left[\frac{Z(c)}{c} \mid \{y(h)\}_{h=1}^{\infty}\right] \mid P\right]\right] + \mathbb{E}\left[\mathbb{E}\left[\text{Var}\left[\frac{Z(c)}{c} \mid \{y(h)\}_{h=1}^{\infty}\right] \mid P\right]\right]. \tag{13}$$

Now, it is easy to see that

$$\mathbb{E}\left[\text{Var}\left[\mathbb{E}\left[\frac{Z(c)}{c} \mid \{y(h)\}_{h=1}^{\infty}\right] \mid P\right]\right] = \frac{1}{c^2} \mathbb{E}\left[\text{Var}\left[\mathbb{E}[Z(c) \mid \{y(h)\}_{h=1}^{\infty}] \mid P\right]\right] = \frac{1}{c^2} \mathbb{E}\left[\text{Var}\left[c \sum_{h=1}^{\infty} p_x e^{-y(h)} \mid P\right]\right] = \sum_{h=1}^{\infty} \sum_{k=1}^{\infty} \mathbb{E}[h p_x k p_x] \text{Cov}[e^{-y(h)}, e^{-y(k)}].$$
Also, observing that \( \mathbb{V}ar[\mathbb{E}[Z(c)/c|\{y(h)\}_{h=1}^\infty]] \) is a measure of the variability of \( Z(c)/c \) due to the effect of the stochastic discounting factors, the effect due to the randomness of mortality having been averaged out [see, also Coppola et al., (2000)] and is independent of \( c \). This leads to the following definition

**Definition 2A.** \( \mathbb{E}[\mathbb{V}ar[\mathbb{E}[Z(c)|\{y(h)\}_{h=1}^\infty]|P]] \) is a measure of the portfolio's investment risk.

From equation (3) we have

\[
\mathbb{E}[\mathbb{E}[\mathbb{V}ar\left(\frac{Z(c)}{c}\right)|\{y(h)\}_{h=1}^\infty]|P]] = \frac{1}{c^2} \mathbb{E}[\mathbb{E}[\mathbb{V}ar[Z(c)|\{y(h)\}_{h=1}^\infty]|P]].
\]

We observe that \( \mathbb{E}[\mathbb{V}ar\left(\frac{Z(c)}{c}\right)|\{y(h)\}_{h=1}^\infty]|P]] \) is a measure of the variability of \( Z(c)/c \) due to the randomness of the number of deaths for a given mortality table. Notice the effect of pooling risks: as \( c \) tends to infinity, this measure tends to zero.

**Definition 3A.** \( \mathbb{E}[\mathbb{V}ar\left(\frac{Z(c)}{c}\right)|\{y(h)\}_{h=1}^\infty]|P]] \) is a measure of the insurance risk.

### 3.2 Conditioning on the Interest Rate Process

The variance of \( Z(c)/c \) can be decomposed in another way:

\[
\mathbb{V}ar\left(\frac{Z(c)}{c}\right) = \mathbb{V}ar[\mathbb{E}[\frac{Z(c)}{c}|\{y(h)\}_{h=1}^\infty]] + \mathbb{E}[\mathbb{V}ar\left(\frac{Z(c)}{c}\right)|\{y(h)\}_{h=1}^\infty]].
\]

(14)

The first term on the right side of equation (14) provides a measure of the variability of \( Z(c)/c \) due to stochastic interest rates (discount factors), while the effect of the demographic components (projection and deviations) has been averaged out. As

\[
\mathbb{V}ar[\mathbb{E}[\frac{Z(c)}{c}|\{y(h)\}_{h=1}^\infty]] = \sum_{h=1}^\infty \sum_{k=1}^\infty \mathbb{E}[h p_x] \mathbb{E}[k p_x] \mathbb{C}ov[e^{-y(h)}, e^{-y(k)}],
\]

(15)

we note that \( \mathbb{V}ar[\mathbb{E}[Z(c)/c|\{y(h)\}_{h=1}^\infty]] \) does not depend on \( c \), the portfolio's size, as we can expect in the case of a systematic risk. This suggests the following alternative definition of investment risk:
Definition 2B. \( \text{Var}[\mathbb{E}[\frac{Z(c)}{c} | \{y(h)\}_{h=1}^\infty]] \) is a measure of the investment risk.

The second term on the right side of equation (14) can be split in turn as follows

\[
\mathbb{E}[\text{Var}[\frac{Z(c)}{c} | \{y(h)\}_{h=1}^\infty]] = \mathbb{E}[\text{Var}[\mathbb{E}[\frac{Z(c)}{c} | P] | \{y(h)\}_{h=1}^\infty]] \\
+ \mathbb{E}[\text{Var}[\frac{Z(c)}{c} | P] | \{y(h)\}_{h=1}^\infty]]. \tag{16}
\]

Now we observe that \( \mathbb{E}[\text{Var}[\mathbb{E}[Z(c)/c|P] | \{y(h)\}_{h=1}^\infty]] \), which is a measure of the variability of \( Z(c)/c \) due to the randomness of the projection, does not depend on \( c \), as we expect by virtue of the systematic nature of this kind of risk. In fact

\[
\mathbb{E}[\text{Var}[\mathbb{E}[\frac{Z(c)}{c} | P] | \{y(h)\}_{h=1}^\infty]] = \mathbb{E}[\text{Var}\left(\sum_{h=1}^\infty \frac{h \rho_x e^{-y(h)}}{y(h)} | \{y(h)\}_{h=1}^\infty\right)] \\
= \sum_{h=1}^\infty \sum_{k=1}^\infty \text{Cov}[h \rho_x, k \rho_x] \mathbb{E}[e^{-y(h)-y(k)}]. \tag{17}
\]

This suggests the following alternative definition:

Definition 1B. \( \mathbb{E}[\text{Var}[\mathbb{E}[\frac{Z(c)}{c} | P] | \{y(h)\}_{h=1}^\infty]] \) is a measure of the projection risk.

Because

\[
\mathbb{E}[\mathbb{E}[\text{Var}[\frac{Z(c)}{c} | P] | \{y(h)\}_{h=1}^\infty]] = \mathbb{E}[\text{Var}[\frac{Z(c)}{c} | \{y(h)\}_{h=1}^\infty | P]] \tag{18}
\]

where the variance is calculated with respect to the random deviations of mortality, \( \mathbb{E}[\text{Var}[Z(c)/c|P] | \{y(h)\}_{h=1}^\infty]] \) is just equal to the measure provided by Definition 3A. So we have an alternative definition of insurance risk:

Definition 3B. \( \mathbb{E}[\text{Var}[\frac{Z(c)}{c} | P] | \{y(h)\}_{h=1}^\infty]] \) is a measure of the insurance risk.

\(^4\)Note that Definitions x.A and x.B are alternative definitions of the same concept, where x can be 1, 2 or 3.
At this point let us consider the difference between the two investment risk measures given in Definitions 2A and 2B respectively. Note that

$$E[\text{Var}[E[Z(c)|\{y(h)\}_{h=1}^\infty]|P]] - \text{Var}[E[Z(c)|\{y(h)\}_{h=1}^\infty]|P]] = \sum_{h=1}^\infty \sum_{k=1}^\infty \text{Cov}[h p_x, k p_x] \text{Cov}[e^{-y(h)}, e^{-y(h)}]. \quad (19)$$

Now, let us consider the difference between the two projection risk measures given in Definitions 1A and 1B

$$\text{Var}[E[Z(c)|P]] - E[\text{Var}[E[Z(c)|P]|\{y(h)\}_{h=1}^\infty]] = - \sum_{h=1}^{\omega-1-x} \sum_{k=1}^{\omega-1-x} \text{Cov}[h p_x, k p_x] \text{Cov}[e^{-y(h)}, e^{-y(h)}]. \quad (20)$$

4 A Numerical Example

Consider a portfolio of life annuities, each policy being issued to each of $c = 1000$ lives age $x = 65$. We assume that the underlying survival function for the group of insured lives follows a Weibull distribution, i.e.,

$$s(x) = \exp\left(-\left(\frac{x}{\alpha}\right)^\gamma\right), \quad x > 0$$

where $\alpha$ and $\gamma$ are positive constant parameters. The Weibull model is often used because it is simple and fits well the statistical observations and it easily represents mortality related to adult ages; see, for example, Kefitz and Beekman (1984) for specific details.

Specifically we assume that the current mortality is such that $\alpha = 82.7$ and $\gamma = 7.00$. The projected mortality are obtained by choosing $\alpha$ and $\gamma$ to reflect a survival function with mortality rates that are pessimistic (i.e., higher than expected), realistic (i.e., as expected), and optimistic (i.e., lower than expected). Table 1 displays the values of $\alpha$ and $\gamma$ for the various projections.

Moreover, we consider projected survival tables by choosing the parameters $\alpha$ and $\gamma$ corresponding to a survival function for contemporaries and to three projected tables with increasing survival probabilities. [See Olivieri (1998)]
Table 1
Parameters Used for the Projections

<table>
<thead>
<tr>
<th>Parameters Used for the Projections</th>
<th>Contemporary Mortality Table</th>
<th>Pessimistic Mortality Projection</th>
<th>Realistic Mortality Projection</th>
<th>Optimistic Mortality Projection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>82.7</td>
<td>8.00</td>
<td>85.2</td>
<td>9.15</td>
</tr>
<tr>
<td></td>
<td>83.5</td>
<td>8.00</td>
<td>85.2</td>
<td>9.15</td>
</tr>
<tr>
<td></td>
<td>87</td>
<td>10.45</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

With regards to the force of interest rate process, we calculate the parameters \( \beta \) and \( \sigma \) as described in Coppola et al., (2000). In particular, recall that the stochastic process \( X(t) \) defined in Section 2.1 represents the deviations of the force of interest from its expected value. Thus, the differences between actual observed rates and their forecasted values are used to estimate \( \beta \) and \( \sigma \) by means of the covariance equivalence principle. [See, for example, Pandit and Wu (1983), or Parker (1994a or b).]

For this example, however, our illustrations, the Italian short term (three months) bond series for the period 1993-1996 is used. It turns out that the constant deterministic component is \( i = 0.09 \), and the parameters of the deviation process, \( X(t) \), are \( \beta = 0.11 \) and \( \sigma = 0.005 \).

Table 2 displays certain values for the average cost per policy for each type of mortality table. The insurance system scenario is such that the probability of choosing a type of projected mortality table is 0.2, 0.6, 0.2 for the pessimistic, the realistic, and the optimistic projections, respectively. Table 3 shows the variance decomposition for each definition.

In particular, according to the first variance decomposition presented in Section 3.1, we obtain the values in the first column on the basis of Definitions 1A, 2A, 3A. In the second column, according to the second variance decomposition presented in Section 3.2, the values are obtained by means of Definitions 1B, 2B, 3B.

Concerning the mean value, investment risk, and variance, we can note that the values in Table 3 are greater than the corresponding values in Table 2 for the realistic projection and smaller than those for the optimistic projection. On the contrary, the insurance risk in Table 3 is smaller than the insurance risk in Table 2 for the realistic projection, and it is greater than insurance risk in Table 2 for the optimistic projection. Moreover, in Table 3 a new set of risk parameter appears, which is the projection risk. Note the contribution of the projection risk is higher than the insurance risk.
Table 2

Properties of $Z(c)/c$ for $c = 1000$ and $x = 65$

<table>
<thead>
<tr>
<th>Mortality Type</th>
<th>Contemporary</th>
<th>Projected Mortality Table Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Value</td>
<td>Pessimistic</td>
</tr>
<tr>
<td>Mean Value</td>
<td>7.11024</td>
<td>7.33410</td>
</tr>
<tr>
<td>Variance</td>
<td>0.42786</td>
<td>0.46240</td>
</tr>
<tr>
<td>Investment Risk</td>
<td>0.42088</td>
<td>0.45613</td>
</tr>
<tr>
<td>Insurance Risk</td>
<td>0.00698</td>
<td>0.00627</td>
</tr>
</tbody>
</table>

Table 3

Variance Decompositions

<table>
<thead>
<tr>
<th></th>
<th>First Decomp.</th>
<th>Second Decomp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Value</td>
<td>7.65833</td>
<td>7.65833</td>
</tr>
<tr>
<td>Projection Risk</td>
<td>0.04693</td>
<td>0.04601</td>
</tr>
<tr>
<td>Investment Risk</td>
<td>0.51442</td>
<td>0.51534</td>
</tr>
<tr>
<td>Insurance Risk</td>
<td>0.00538</td>
<td>0.00538</td>
</tr>
<tr>
<td>Variance</td>
<td>0.56672</td>
<td>0.56672</td>
</tr>
</tbody>
</table>

Finally, we note also that the differences between the projection and the investment risk measures are very small in the two decomposition procedures.

5 Closing Comments

We have considered a model for a portfolio of identical life annuities under the assumption that both mortality and interest rates are random, and the rate of return consists of two components: a deterministic one, which takes into consideration the existing investments of the company, and a stochastic one, which is modeled by an Ornstein-Uhlenbeck process.

The main part of the paper analyzes the effect of the randomness of the projected mortality rates in the valuation of an annuity portfolio. This study points out the importance of the systematic risk component due to the randomness of the survival functions used in constructing the mortality tables. Further information about the analysis of the pro-
jection risk could be obtained by scenario testing on different mortality tables.

In the context of a life annuity portfolio, in which the three risks under consideration are the mortality risk, the projection risk and the financial risk, the equations we have derived are easy to implement by practitioners. Moreover in this way the overall variance is obtained simply adding the three contributions.

References


A Note on the Parallelogram Method for Computing the On-Level Premium

David P.M. Scollnik* and Wai Man Sara Lau†

Abstract‡

This paper discusses the differences appearing in the descriptions of the parallelogram method for the determination of earned premium at current rate levels given by McClenahan (1996) and Brown and Gottlieb (2001). It observes that the former is consistent with the method of extending exposures while the latter is not. An illustration is provided. This paper also discusses two other approaches to the determination of the earned premium.

Key words and phrases: earned premium, extending exposures, ratemaking

1 Introduction

For the purpose of ratemaking, it is often necessary to determine the dollars of earned premium at current rates. The method of extending exposures (also known as the extension of exposures technique) simply re-rates each policy using the current rate manual and the existing distribution of earned exposures. This is the best method when detailed

*David P.M. Scollnik, A.S.A., Ph.D., is an associate professor of actuarial science and statistics at the University of Calgary. He holds a Ph.D. in statistics from the University of Toronto. Dr. Scollnik's address is: Department of Mathematics and Statistics, University of Calgary, 2500 University Drive N.W., Calgary, Alberta T2N 1N4, CANADA. Internet address: scollnik@math.ucalgary.ca and <http://www.math.ucalgary.ca/~scollnik>

†Wai Man Sara Lau is a graduate student in the Department of Mathematics and Statistics. Ms. Lau's address is: Department of Mathematics and Statistics, University of Calgary, 2500 University Drive N.W., Calgary, Alberta T2N 1N4, CANADA. Internet address: wm­lau@math.ucalgary.ca

‡This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors would like to thank the editor for helpful comments that have improved the presentation of this paper.
data and appropriate rating software are both available. When this is not the case, the so-called parallelogram method can be used instead. The parallelogram method adjusts calendar year earned premiums to reflect the effect of all rate changes made since these earned premiums were written. McClenahan (1996, pages 42-44) and Brown and Gottlieb (2001, pages 73-76) describe these two methods in some detail.

Careful readers of McClenahan (1996) and Brown and Gottlieb (2001) will note that, whereas the geometrical interpretations given to the parallelogram method initially appear to be the same in both sources, the implementation details regarding the definition and calculation of the on-level factors differ. Practitioners and students taking the SOA Course 5 and/or CAS Exam 5 will benefit from an explicit mention of the discrepancy along with an illustration and discussion clarifying the discrepancy. This paper also discusses two other approaches to the determination of the earned premium.

2 Illustration

Suppose that the experience period consists of three calendar years, Z, Z + 1, and Z + 2, during which the earned premiums were 1250, 1575, and 1620, respectively. Suppose $P$ is the rate level effective 1/1/Z - 1 and rate increases (applied to newly issued policies or renewals) were introduced as follows:

- $+25\%$ effective 7/1/Z, and
- $+28\%$ effective 4/1/Z + 1.

Figure 1 shows the rates in effect during the experience period, under the standard assumptions that all policies have a one-year term and policy issue dates are uniformly distributed over time. Under these assumptions, the parallelogram method can be used to determine the proportion of policies in each year that were written at the various premium rate levels. Using the methodologies in McClenahan (1996) or Brown and Gottlieb (2001), the reader can verify that these proportions are as given in Table 1.

\footnote{These standard assumptions can be modified using techniques available in the casualty actuarial literature as in Miller and Davis, 1976.}
Figure 1
Experience Period and Rate Changes

Table 1
Proportion of Policies Within Each Calendar Year Written at the Premium Rates Effective On 1/1/Z – 1, 7/1/Z, and 4/1/Z + 1

<table>
<thead>
<tr>
<th>Calendar Year</th>
<th>1/1/Z – 1</th>
<th>7/1/Z</th>
<th>4/1/Z + 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>0.875</td>
<td>0.1250</td>
<td>0.00000</td>
</tr>
<tr>
<td>Z + 1</td>
<td>0.125</td>
<td>0.59375</td>
<td>0.28125</td>
</tr>
<tr>
<td>Z + 2</td>
<td>0.000</td>
<td>0.03125</td>
<td>0.96875</td>
</tr>
</tbody>
</table>
2.1 Using the Brown and Gottlieb (2001) Method

At this stage the parallelogram method as described in Brown and Gottlieb (2001) proceeds by determining the multiplicative factors that should be applied to each premium band within each calendar year in order to calculate the earned premium at the current rate level. These rate promotion factors are given in Table 2 and are illustrated in Figure 2.

<table>
<thead>
<tr>
<th>Calendar Year</th>
<th>1/1/Z - 1</th>
<th>7/1/Z</th>
<th>4/1/Z + 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>1.25 \times 1.28 = 1.6</td>
<td>1.28</td>
<td></td>
</tr>
<tr>
<td>Z + 1</td>
<td>1.25 \times 1.28 = 1.6</td>
<td>1.28</td>
<td>1</td>
</tr>
<tr>
<td>Z + 2</td>
<td></td>
<td>1.28</td>
<td>1</td>
</tr>
</tbody>
</table>

The next step is to determine the weighted average on-level factor for each calendar year as follows:
Scollnik and Lau: A Note on the Parallelogram Method

<table>
<thead>
<tr>
<th>Year</th>
<th>On-Level Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>0.875 \times 1.6 + 0.125 \times 1.28 = 1.56</td>
</tr>
<tr>
<td>Z + 1</td>
<td>0.125 \times 1.6 + 0.59375 \times 1.28 + 0.28125 = 1.24125</td>
</tr>
<tr>
<td>Z + 2</td>
<td>0.03125 \times 1.28 + 0.96875 = 1.00875</td>
</tr>
</tbody>
</table>

The proportions in Table 1 are used as the weights in the calculations above. The on-level factor corresponding to a particular calendar year can be interpreted as the weighted average of the required multiplicative factors needed to bring that year’s earned premium to the current rate.

Table 3
Development of Earned Premium at Current Rates
Using the Methodology in Brown and Gottlieb (2001)

<table>
<thead>
<tr>
<th>Calendar Year</th>
<th>Earned Premium</th>
<th>On-Level Factor</th>
<th>Earned Premium at Current Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>1250</td>
<td>1.56000</td>
<td>1950.000</td>
</tr>
<tr>
<td>Z + 1</td>
<td>1575</td>
<td>1.24125</td>
<td>1954.969</td>
</tr>
<tr>
<td>Z + 2</td>
<td>1620</td>
<td>1.00875</td>
<td>1634.175</td>
</tr>
<tr>
<td>Total:</td>
<td></td>
<td></td>
<td>5539.144</td>
</tr>
</tbody>
</table>

The earned premiums for the different calendar years under consideration are reported in the second column of Table 3. The estimated earned premiums at the current rate level are developed in the fourth column of Table 3 using calculated on-level factors.

2.2 Using the Method Described in McClenahan (1996)

The traditional methodology described in McClenahan (1996) also begins with the determination of the proportions in Table 1. Instead of developing the rate promotion factors in Table 2, however, one determines the relationship each premium rate class within each calendar year bears to the earliest rate in effect at the beginning of the period under examination.

For instance, the earned premium in calendar year Z +1 that was written between 7/1/Z and 4/1/Z + 1 was written at a level that is equal to 1.25 times the rate that was in effect on 1/1/Z – 1. These relations or factors are reported in Table 4 and are illustrated in Figure 3.
Table 4
Relation of the Written Premium Rates
Within Each Calendar Year to the Earliest Premium Rate

<table>
<thead>
<tr>
<th>Year</th>
<th>1/1/Z - 1</th>
<th>7/1/Z</th>
<th>4/1/Z + 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>1.25</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Z + 1</td>
<td>1.25</td>
<td>1.25</td>
<td>1.25 × 1.28 = 1.6</td>
</tr>
<tr>
<td>Z + 2</td>
<td>-</td>
<td>1.25</td>
<td>1.25 × 1.28 = 1.6</td>
</tr>
</tbody>
</table>

Figure 3
Relation of the Written Premium Rates
Within Each Calendar Year to the Earliest Premium Rate
The next step is to determine the weighted average of these factors for each calendar year as follows:

<table>
<thead>
<tr>
<th>Year</th>
<th>Weighted Average Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>0.875 + 0.125 × 1.25 = 1.03125</td>
</tr>
<tr>
<td>Z + 1</td>
<td>0.125 + 0.59375 × 1.25 + 0.28125 × 1.6 = 1.3171875</td>
</tr>
<tr>
<td>Z + 2</td>
<td>0.03125 × 1.25 + 0.96875 × 1.6 = 1.5890625</td>
</tr>
</tbody>
</table>

As before, the proportions in Table 1 are used as the weights in the calculation of the weighted average factors above. These weighted average factors are not yet the on-level factors as they are traditionally defined. [For example, as in McClenahan (1996).]

Rather, the on-level factor for a given calendar year is determined by dividing the current rate level (i.e., 1.6P in this example) by P and by the weighted average factor for the year under consideration. So, the on-level factors for this example are given as follows:

<table>
<thead>
<tr>
<th>Year</th>
<th>On-Level Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>1.6/1.03125 = 1.551515</td>
</tr>
<tr>
<td>Z + 1</td>
<td>1.6/1.3171875 = 1.214709</td>
</tr>
<tr>
<td>Z + 2</td>
<td>1.6/1.5890625 = 1.006883</td>
</tr>
</tbody>
</table>

It is evident that these on-level factors differ from the ones constructed using the methodology given in Brown and Gottlieb (2001), as do the resulting estimates of the earned premium at current rates given in Table 5.

### Table 5

**Development of Earned Premium at Current Rates Using the Traditional Methodology as in McClenahan (1996)**

<table>
<thead>
<tr>
<th>Calendar Year</th>
<th>Earned Premium</th>
<th>On-Level Factor</th>
<th>Earned Premium at Current Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>1250</td>
<td>1.551515</td>
<td>1939.394</td>
</tr>
<tr>
<td>Z + 1</td>
<td>1575</td>
<td>1.214709</td>
<td>1913.167</td>
</tr>
<tr>
<td>Z + 2</td>
<td>1620</td>
<td>1.006883</td>
<td>1631.150</td>
</tr>
<tr>
<td>Total:</td>
<td></td>
<td></td>
<td>5483.711</td>
</tr>
</tbody>
</table>
3 Interpretation of the Results

3.1 Practical Interpretation of the Results

In Section 2, we stated that the parallelogram method can be used to determine what proportion of policies in each year were written at the various premium rate levels. The differences we have observed in the two methodologies arise from the fact that McClenahan (1996) interprets these proportions as proportions of earned exposures in a year, whereas Brown and Gottlieb (2001, page 75) implicitly develop them as proportions of earned exposures but then use them as proportions of earned premium dollars. Clearly, both cannot be correct under the assumption that policy issues are uniformly distributed within a calendar year during which a rate change occurs.

Suppose that an additional 1000 earned units of exposure were discovered in the books for each of calendar years \( Z, Z+1, \) and \( Z+2 \), in the context of the previous illustration, and we wanted to determine the addition to our estimate of the current earned premium. Using the method of extending exposures, this value is simply \( 3 \times 1000 \times 1.6 = 4800 \).

If we assume that policy issues were uniformly distributed within each year, then the additional earned premiums in calendar years \( Z, Z+1, \) and \( Z+2 \) are given by 1031.25, 1317.1875, and 1589.0625, respectively. Using the on-level factors developed in Section 2.2, we find that McClenahan’s method estimates the addition to current earned premium as follows:

\[
1031.25 \times 1.551515 + 1317.1875 \times 1.214709 + 1589.0625 \times 1.006883 = 1600 + 1600 + 1600 = 4800.
\]

This result is consistent with that given by the method of extending exposures.

On the other hand, using the on-level factors developed in Section 2.1 we find that Brown and Gottlieb’s method estimates the addition to current earned premium as follows:

\[
1031.25 \times 1.56 + 1317.1875 \times 1.24125 + 1589.0625 \times 1.00875 = 1608.75 + 1634.96 + 1602.97 = 4846.68.
\]
This result is not consistent with that given by the method of extending exposures. It also demonstrates that 1000 units of earned exposure in a calendar year will yield different additions to current earned premium, depending on the calendar year to which it is assigned. This is not the answer that most practitioners are looking for.

This illustration clarifies the fact that the parallelogram methodology described in McClenahan (1996) is the one that should be used under the assumption that policy issue dates are uniformly distributed over time. When the policy issue dates are uniformly distributed, McClenahan's method yields results that are consistent with the method of extending exposures, and with traditional methods in the casualty actuarial literature, such as Kallop (1975) and Miller and Davis (1976).

The on-level factors defined in Brown and Gottlieb (2001) are appropriate if the dollars of earned premium income are uniformly distributed over any calendar year so that the proportions given by the parallelogram method were proportions of earned premium. This is not the assumption made in Brown and Gottlieb (2001, pages 73 and 76) (cf. Brown, 1993, page 74), however, nor can it ever be consistent with the assumption of uniform policy issues in a period containing a rate change.

3.2 Mathematical Interpretation of the Results

To explain the difference in the two methods in a more formal fashion, consider the following: Suppose that there are \( n(k) \) premium bands in the calendar year \( Z + k \), and the base premium at the start of the experience period (i.e., on 1/1/Z) is \( P \). Then, for \( i = 1, 2, \ldots, n(k) \), define the following:

- \( p_i^{(k)} \) = Premium for band \( i \) in calendar year \( Z + k \);
- \( f_i^{(k)} \) = Factor applied to \( P \) to give \( p_i^{(k)} \), i.e.,
- \( p_i^{(k)} = f_i^{(k)} \times P \);
- \( p^{(\text{cur})} \) = Current premium rate at the end of the experience period;
- \( g_i^{(k)} \) = Factor applied to \( p_i^{(k)} \) to give \( p^{(\text{cur})} \), i.e.,
- \( p^{(\text{cur})} = f_i^{(k)} \times g_i^{(k)} \times P \); and
- \( A_i^{(Z)} \) = Proportion of policies written in band \( i \), with

\[ A_1^{(k)} + \cdots + A_{n(k)}^{(k)} = 1. \]
The on-level factor for calendar year $Z + k$ can now be given as follows: The McClenahan (1996) on-level factor, $OLF_M^{(k)}$, is given by

$$OLF_M^{(k)} = \frac{p^{(\text{cur})}}{p} \frac{1}{\sum n(k) A_i^{(k)} P_i^{(k)}} = \frac{p^{(\text{cur})} / p}{\sum n(k) A_i^{(k)} f_i^{(k)}} = \left[ \sum_{i=1}^{n(k)} \frac{A_i^{(k)}}{g_i^{(k)}} \right]^{-1}.$$

On the other hand, the Brown and Gottlieb (2001) on-level factor, $OLF_{BG}^{(k)}$, is given by

$$OLF_{BG}^{(k)} = \sum_{i=1}^{n(k)} \frac{A_i^{(k)} P_i^{(k)}}{P_i^{(k)}} = \frac{p^{(\text{cur})} \sum_{i=1}^{n(k)} A_i^{(k)}}{p} \sum_{i=1}^{n(k)} f_i^{(k)} = \sum_{i=1}^{n(k)} A_i^{(k)} g_i^{(k)}.$$

It is evident that $OLF_{BG}^{(k)}$ is not equal to $OLF_M^{(k)}$ in general.

In Section 3.1, we observed that $OLF_{BG}^{(k)}$ would be appropriate if the dollars of earned premium income were uniformly distributed over any calendar year so that the proportions given by the parallelogram method were proportions of earned premium. In this case,

$$A_i^{(k)} = \frac{A_i^{(k)} P_i^{(k)}}{\sum_{i=1}^{n(k)} A_i^{(k)} P_i^{(k)}},$$

which implies

$$P_i^{(k)} = \sum_{i=1}^{n(k)} A_i^{(k)} P_i^{(k)},$$

for $i = 1, 2, \ldots, n(k)$. Under this assumption, we have

$$OLF_{BG}^{(k)} = \sum_{i=1}^{n(k)} \frac{A_i^{(k)} P_i^{(k)}}{P_i^{(k)}} = \frac{p^{(\text{cur})} \sum_{i=1}^{n(k)} A_i^{(k)}}{\sum_{i=1}^{n(k)} A_i^{(k)} P_i^{(k)}} = \frac{p^{(\text{cur})} / p}{\sum_{i=1}^{n(k)} A_i^{(k)} f_i^{(k)}} = OLF_M^{(k)}.$$

This demonstrates that $OLF_{BG}^{(k)}$ is equal to $OLF_M^{(k)}$, and hence its usage will reproduce the method of extending exposures, only in a very special (and unrealistic) case.

### 4 Further Discussion and New Approaches

Of course, the assumption (as in McClenahan, 1996) that policy issue dates are uniformly distributed over time is unlikely to be consistent with actual experience. When the actual past levels of exposures are known and available, the procedure described in Miller and Davis (1976)
Scollnik and Lau: A Note on the Parallelogram Method

can be used to determine the premium adjustment factors and on-level premiums. When this is not the case, the following variation on the parallelogram method might be used instead.

The basic idea behind this variation is to use the observed earned premium in each calendar year in order to better approximate the true underlying levels of exposure and then reprice these exposure levels at the current premium rate. The standard assumption that policy issue dates are uniformly distributed over time is replaced with the assumption that they are uniformly distributed at a constant rate between any two adjacent premium rate change dates. This twist on the parallelogram method will be illustrated in the context of the continuing example from Section 2.

Let $P_1$ be the initial premium rate and let $E_1$ denote the constant policy issue rate in effect prior to 7/1/$Z$. Then, $P_1 \times E_1$ is the annual rate at which earned premium was being generated prior to 7/1/$Z$. Let $E_2$ denote the constant policy issue rate in effect between 7/1/$Z$ and 4/1/$Z+1$ and $E_3$ the rate thereafter. Then the earned premium in each of the three calendar years under consideration should satisfy these relations:

<table>
<thead>
<tr>
<th>Year</th>
<th>Earned Premium</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$</td>
<td>$1250 = P_1 \times (E_1 \times 0.875 + E_2 \times 1.25 \times 0.125)$</td>
</tr>
<tr>
<td>$Z + 1$</td>
<td>$1575 = P_1 \times (E_1 \times 0.125 + E_2 \times 1.25 \times 0.59375 + E_3 \times 1.6 \times 0.28125)$</td>
</tr>
<tr>
<td>$Z + 2$</td>
<td>$1620 = P_1 \times (E_2 \times 1.25 \times 0.03125 + E_3 \times 1.6 \times 0.96875)$</td>
</tr>
</tbody>
</table>

Solving this system of linear equations yields the values:

\[
\begin{align*}
P_1 \times E_1 & = 1195.162602, \\
P_1 \times E_2 & = 1307.089431, \\
P_1 \times E_3 & = 1012.220528. 
\end{align*}
\]

Hence, the earned premium at the current rate (i.e., $1.6 \times P_1$) in each calendar year is as follows:

<table>
<thead>
<tr>
<th>Year</th>
<th>Earned Premium at Current Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$</td>
<td>$1934.645529 = P_1 \times 1.6 \times (E_1 \times 0.875 + E_2 \times 0.125)$</td>
</tr>
<tr>
<td>$Z + 1$</td>
<td>$1936.266717 = P_1 \times 1.6 \times (E_1 \times 0.125 + E_2 \times 0.59375)$</td>
</tr>
<tr>
<td>$Z + 2$</td>
<td>$1634.296290 = P_1 \times 1.6 \times (E_2 \times 0.03125 + E_3 \times 0.96875)$</td>
</tr>
<tr>
<td>Total:</td>
<td>$5505.208536.$</td>
</tr>
</tbody>
</table>