ARTICLES

Product Innovation in Financial Services: A Survey
Christopher O'Brien ................................................................. 5

Phased Retirement for Defined Benefit Plan Participants
Patricia L. Seahill and Jonathan Barry Forman ......................... 43

The Actuarial Value of Life Insurance Backdating
James M. Carson and Krzysztof M. Ostaszewski .......................... 63

Decision Tree Analysis of Terminated Life Insurance Policies
Robert Keng Heong Lian, Yuan Wu, and Hian Chye Koh............... 79

A Comparative Study of Parametric and Nonparametric Estimators of Old-Age Mortality in Sweden
Peter Fledelius,Montserrat Guillen,Jens Perch Nielsen, and Kitt Skovsfj Petersen .................................................. 103

Estimation of Complete Period Life Tables for Singaporeans
Siu-Hang Li and Wai-Sum Chan .................................................. 129

Approximating the Bias and Variance of Chain Ladder Estimates Under a Compound Poisson Model
Janagan Yogaranpan, Sue Clarke, Shauna Ferris, and John Pollard... 147

Rapid Calculation of the Price of Guaranteed Minimum Death Benefit Ratchet Options Embedded in Annuities
Eric R. Ulm .................................................................................. 169

Credibility Theory and Geometry
Elias S.W. Shiu and Fuk Yum Sing................................................ 197
Journal of Actuarial Practice

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 referee reports. The referee process is expected to take three to four months (depending on the length of the paper).

See inside back cover for instructions to authors.
Product Innovation in Financial Services: A Survey
Christopher O'Brien*

Abstract†

This paper considers product innovation in insurance and other financial services, an area where actuaries have an important role. It considers the proposition that there is no unique formula for success and that what works well in one situation may not work well in another. It first examines the sources of ideas for new products and, in particular, the role played by consumers, which is generally regarded as weak. It then looks at how ideas are implemented, with particular importance attributed to cross-functional teams and the formality of the product development process. Then it considers how success is measured (with the indirect as well as direct benefits of a development) and the factors that may distinguish success from failure. The paper concludes that there is no unique formula for success, but that there are some shared characteristics of firms that are good innovators; it is comforting to find that there are guidelines that firms can follow to improve their chances of success.

Key words and phrases: product development process, insurance, customer needs

*Christopher O'Brien, M.A., F.I.A., A.S.A., is director of the Centre for Risk and Insurance Studies at Nottingham University Business School. He has an M.A. in economics from Cambridge University, England. He qualified as a Fellow of the Institute of Actuaries in 1975 and worked in the life insurance industry for 28 years, being chief actuary and a director of Royal & SunAlliance U.K. life operations in 1998-99. He moved to his role in Nottingham, in 2000, where his research interests include the management of life insurance firms, especially performance measurement and risk management.

Mr. O'Brien's address is: Centre for Risk and Insurance Studies, Nottingham University Business School, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, ENGLAND. E-mail: christopher.obrien@nottingham.ac.uk

†The author gratefully acknowledges financial support from the Financial Services Research Forum. The comments from anonymous referees are also appreciated; the usual disclaimer applies.
1 Introduction

Consider the following questions:

• Is there a unique secret for product innovation?
• Or perhaps what works well in one situation may not work well in another?

In this paper, we survey the answers to these questions; if innovation tends to be successful in some situations but not others, we explore the factors that help lead to success. In particular, some firms appear to be more efficient and successful innovators than others: what lies behind this?

The paper analyzes the forces behind new products, how they are put into effect, and how they have an impact. In doing so, it recognizes that products are indeed different and that success factors can differ accordingly.

Product innovation covers both changes to existing products and more extensive changes. The evidence is that the former category is the more common. Some product development effort will not involve innovation, which implies something new. Johne (1993), looking at U.K. life and general insurers, found that most development work was updating existing products. Stern and Whittemore (1998), in a survey of U.S. life insurers, found that 85% of all initiatives in 1996 were small changes or line extensions. Ennew (1995b) refers to "product line stretching" or product proliferation having traditionally accounted for much of the product development activity in financial services. Strieter et al. (1997), in a survey of product managers in U.S. banks, found that they spend 70% of their time managing existing products, 5% on product improvement, only 5% on new product development, and 20% on other marketing activities.

We usually regard innovation as a good thing. We need to be conscious, however, of the view that innovation can be excessive. It is certainly true that products can be complex. Abroe (1999) referred to long-term care products in the U.S. having a complex sales environment. Products typically have multiple benefit options and riders; it is not uncommon for a product to have more than 100 benefit options based on the possible combinations of benefit period, elimination period, and optional riders. Sandler (2002), considering medium- and long-term savings products in the U.K., concluded that there was a proliferation of products, some being fundamentally the same but marketed as different, and the sheer volume of products was overwhelming. From time
to time, regulators have attempted to address this issue through mandated product standardization, for example regarding Medicare supplement policies in the U.S. We shall bear this problem in mind when considering the impact of product innovations.

The paper is intended for those who have an interest in the management of product development, with the subject being relevant not only to product actuaries, but also to others such as senior management and product managers. It is assumed that most readers will have had experience in developing products. To enhance our understanding of the management issues, however, the paper also considers the evidence available from not only surveys of financial services firms, but also some more general surveys.

The main topics covered in the following sections of the report are the three "I"s of innovation:

- **Ideas**: Where do the ideas for development come from and why?
- **Implementation**: How are ideas implemented? What are the key features?
- **Impact**: How do we determine the success of a development and what distinguishes the successful from the less successful?

2 Ideas

In this section we consider the first of the three "I"s of innovation: ideas. Is there some unique secret for successful products? Firms are responsible for their products, and we consider how firms come up with new ideas, noting that factors from both the supply side and demand side are relevant. There are several other parties who can have an input to generating ideas.

2.1 Firms

2.1.1 Firm Size and Associated Factors

We first consider whether the structure of an industry can affect the propensity of firms to innovate.

Cohen and Levin (1989) considered the view that in an industry that was more concentrated, innovation was more likely (e.g., because firms could more easily appropriate the fruits of their work) or less likely (e.g., because firms developed inertia). They concluded that the majority of
studies find a positive relationship between market concentration and research and development expenditures. There are a number of significant industry effects. Furthermore, the causality may be that past innovation leads to concentration rather than high concentration being the cause. It is therefore possible, they conclude, that market concentration may exercise no independent effect on R&D intensity.

They also considered different views about the effect of firm size on innovative activity. Economies of scale is one of the arguments presented in favor of large firms being expected to be more innovative; small firms requiring finance for innovation may also suffer capital market disadvantages. Alternatively, as firms grow large, efficiency in innovation may be undermined through loss of managerial control and inability to reward those individuals who have key roles in innovation.

Murray (1976), in a survey of U.S. insurers, found that innovations sprung from firms of all sizes. Indeed, Cohen and Levin (1989) found the evidence on the effect of firm size to be inconclusive. In certain circumstances, however, size may well be relevant; for example, if there are expensive technological developments, large firms may find these easier to implement. [See Edgett, 1993, in a survey of U.K. building societies (residential mortgage providers).]

The age of firms may be important; new firms may be especially innovative. In the U.K., Her Majesty's Treasury (2001) said that new entrants have begun to challenge existing firms to change their ways.

2.1.2 Adversity

Nickell, Nicolitsas, and Patterson (2001), in a general survey, found significant evidence that innovation in management techniques was encouraged by adversity; not only did firms need to change to avoid financial difficulties, but the slack demand gave them the resources to innovate.

We can also see examples in financial services firms striving to innovate when they have been in difficulty. Low interest rates have made many traditional insurance products expensive for insurers, who have been forced to put more effort into new products, including unit-linked policies; for example, in China, the first unit-linked policy was introduced in 1999 (Zhou, 2000).

A further area where insurers are expected to come under pressure is the increased availability of genetic testing, which could threaten the viability of protection products. But this could also be a stimulus to new products; one possibility is a combination of long-term care and
pension, although regulatory resistance to genetic testing being used in conjunction with insurance may hinder developments.

2.1.3 Strategy

Successful innovation may depend on whether innovation is a key element in a firm's strategy. The Design Council (2000) had an initiative to find the most innovative products and services in 1995-99 in the U.K. and concluded: "The lasting impression of these companies is of the sheer energy and commitment with which innovation is pursued. ... They are out at the leading edge, driven by a deep-rooted, company-wide passion for innovation."

Edgett (1993) referred to studies on tangible new product development, where a strong strategic focus was related to successful innovation. In his own work on building societies, he found a general lack of strategic focus on product innovation, perhaps related to the relative informality with which the activity is undertaken (see 3.5). By having a strong strategic focus combined with good development practices, a firm's success rate can be increased.

2.1.4 Organizational Structure, Individuals, and Creativity

Bharadwaj and Menon (2000) investigated the hypothesis that innovation is a function of individual efforts and organizational systems to facilitate creativity. In a general survey of U.S. firms, they found evidence to support this, especially from organizational systems (such as a formal idea generation program), although it was individual creativity that was emphasized by firms in the financial services sector.

Adams, Day, and Dougherty (1998) identified organizational learning barriers to new product development (ambiguity, compartmentalized thinking, and inertia) and identified steps to overcome them. An important step that they identify is the use of cross-functional approaches.

Johne and Pavlidis (1996) considered banks introducing derivative products and found that three of the four most active innovators were structured on a product basis. The fourth, and the less active, banks were organized on a functional input basis. The most active also placed significantly more emphasis on marketing than the less active, consistent with the findings of other work. It was also noted that the less active innovators frequently place heaviest emphasis on getting technical features right before selling operations started in earnest; in contrast,
the most active, while accepting that technology is important, regarded it as insufficient on its own.

It is useful to review a study of insurers. Johne and Davies (1999) consider approaches to stimulating change in mature (general) insurance companies. They use the analysis of Peters and Waterman (1982) to develop hypotheses that innovation will be more common where:

- **Strategy:** a more balanced mix of innovation types is pursued, and there was a greater emphasis on major as opposed to incremental innovations;

- **Style:** there is an emphasis on a participative transformational leadership style;

- **Shared values:** there is closer agreement on objectives;

- **Structure:** formal levels in the organizational hierarchy are fewer; there is a greater customer focus to the formal structure; and there is less centralization of decision-making by top management;

- **Systems:** formalization and centralization of systems is relatively low; but standardization in the use of cross-functional teamwork is higher, reflecting an atmosphere that had been freed from the bureaucratic strangulation that characterizes more mature organizations;

- **Staff:** a higher proportion of staff is working on front-end business activities; and

- **Skills:** there is a wider range of trained functional specialists.

They studied eight mature U.K. insurers (mostly foreign-owned) and found that the "pacesetters" had chief executives who adopted a dictatorial style of management: they were destructive, not showing a caring attitude to incumbent staff, and were breaking down existing organizational structures. After this phase, however, they went into "building-up" mode, where they hired new specialist staff and changed to a far more participatory style. Pacesetter chief executives introduced new systems of management that displayed features of new-style organizations, with the emphasis not on strategy and structure but on shared values, staff, and skills. Appreciation of marketing played a key role, and they insisted that market opportunities provided a direction for organizational change.
2.1.5 Competitors

McGoldrick (1994) referred to the importance for financial services firms of gaining ideas from competitors. A firm may prefer to be a "copycat" developer; once the innovator has blazed the trail, the mechanics of design and, where necessary, regulatory approval or filings become easier. Edgett (1993) described the norm in building societies as "me-too" products.

The reference in Section 2.1.2 to unit-linked life policies is an example of the spread of products internationally. There can, however, be barriers to overcome. Products designed to meet a local regulatory or tax environment are not intended to be migrated elsewhere. Customer inertia may mean that countries have their own likes and dislikes, and it can be difficult to break in; Cruickshank (2000) shows how a number of countries have adopted different directions in innovations in banking services, some being more innovative than others. In some instances the local traditions are important; for example, some continental European countries have traditionally favored participating life insurance products, and it has taken some time for unit-linked products to get off the ground. It may be that common technological developments will help speed the transmission of technology-based products, except that where the technology is replacing human intervention, we would expect differences in how countries respond to this. Notwithstanding these problems, there have been some successes; for example, critical illness products were invented in South Africa and are now important elsewhere.

2.1.6 Reinsurers and Merchant Banks

We also note that reinsurers and merchant banks promote new ideas to financial services firms; they are relying on those firms to sell the products to retail customers.

2.1.7 Market Potential

Notwithstanding the criticism that the financial services industry has taken insufficient account of customers needs (see Section 2.2), there clearly are areas where innovations have taken place as firms see a market for a new product. For example, internet insurance, while technologically-driven, falls into this category. The Faculty and Institute of Actuaries (2001) indicates that we can expect to see demographic changes driving a growing range of disinvestment products; Shifman
I

12

Journal of Actuarial Practice, Vol. 11, 2004

(1999) emphasizes the opportunities for insurers in the area of immediate annuities, while Wadsworth, Findlater, and Boardman (2001) demonstrate that this is already happening as new types of annuity products develop.

Products may have their life cycle, as referred to by Ennew (1995a) and Diacon and Watkins (1995). The phases of introduction, growth, maturity, and decline imply that there are incentives for innovation as the decline stage approaches.

2.1.8 Comments

Several authors have summarized the main motivations for innovation. For example, Friedwald (1991), reviewing product development in life insurance, put forward the following categories:

- Defensive (to address an old product nearing the end of its useful life);
- Aggressive (to meet new markets or new demands);
- Legislative (i.e., driven by laws or regulations); and
- Financial (to make better use of capital, to optimize the company's tax position or reduce exposure to risk).

Johne and Davies (2000) concluded that mature general insurers emphasized "market innovation," i.e., new ways of reading and serving markets, concerned with entry into market segments that were new to the company. Also important were product innovation, to ensure that appropriate offers are available to serve chosen markets, and process innovation, to reduce costs. Dixon (1990) referred to profit-maximization as the role of company management, and factors such as new technology, changing patterns of demand, and tax issues that could lead to product innovations. In addition he referred to the "marketing edge" that can come from a firm that is quick to spot and develop a new opportunity. Furthermore, a product that is unique can be charged at a higher premium than otherwise (although copycats mean this may not last for long). Last, he mentioned the validity of developing new products that can demonstrate to the market that the firm has a go-ahead image. On the other hand, there were dangers of developing products merely to respond to a request from the sales force. Insurance managers often regard themselves as facing a conundrum: is their customer the end-customer or the distribution force?
McGoldrick (1994, page 200) says it is rare for there to be one overriding source or innovation for a product innovation: "Most often we see an evolutionary process, driven by competitive forces, technological change and, hopefully, retailers' perceptions of customer needs." He goes on to say that the impetus for a differentiating factor may have an internal or external source. It may come from specialist in-house research or from an agency. It could be a junior employee who works closely with the customer. It could be regulation. Or it could be competitors.

While there is a variety of research results, we can highlight the following conclusions:

- Adversity is a stimulus to innovation (though not necessarily successful);
- There is no one type of firm that is successful at innovation, although innovation activity may be a particular characteristic of new firms;
- A positive factor is if a firm emphasizes innovation in its strategy and products in its organizational structure; and
- The stimuli to innovation include the ability to identify market potential as well as supply-side factors.

2.2 Customers

2.2.1 Customer Influence

Customer influence in financial services is arguably low. Knights, Sturdy, and Morgan (1994) have indicated that although financial services firms appear to have paid more attention to marketing, this still is not very significant. De Brentani (1993) has argued that financial services firms have not taken advantage of customers as an important source of ideas. McGoldrick (1994) indicates that some firms are very weak at harnessing customers' ideas. Abercromby and Hall's (1994) survey of U.K. life insurers found that direct customer research rarely occurs in developing new products. Akamavi, Thwaites, and Burgess (1999) indicate that financial services firms have a greater need to involve customers in new product development. Oliver (2000) commented on unit-linked policies with charging systems of "mind-boggling complexity and confusability." Sandler (2001, 2002), reviewing medium- and long-term retail saving, concluded that the general picture is one of weak consumer influence.
In a 2002 report, PwC Consulting refer to the market for retail financial services being increasingly commoditized as new channels erode entry barriers. Although this is more prevalent in banking products compared to life insurance, the title of PwC's report ("Simplify to Succeed") may have a valuable message. Simpler products can be helpful, given the low level of financial understanding of most of the population. Given that many life assurance products will be on a firm's systems for 20 years or more, it can be a material advantage if the firm avoids complications that will make processing more difficult—relevant for both the computers (where systems will probably have to be rewritten before all the policies come to the end of their lifetime) and the insurer's staff.

2.2.2 Customers and Innovation

Mohammed-Salleh and Easingwood (1993) found from interviews and a questionnaire survey of financial institutions that test marketing is rarely conducted as part of product development. There can be a number of valid reasons for this, including the difficulty of producing test market conditions and the threat that details of the new product may be leaked to competitors, in a situation where copying can be cheap and quick.

We should not assume that non-financial firms always carry out proper market assessments with customer inputs, as this is not the case (Adams, Day, and Dougherty, 1998). Nevertheless, the evidence does not lead to confidence that customers have an input to innovation. There is a contrary point, however: perhaps the really excellent firms know their business so well that they can take short cuts that others cannot afford to take. While there are risks in this, it is the sensible taking of risks that leads to superior performance for shareholders.

Section 2.1.7 shows that market potential can lead to innovation. Ennew (1995b) refers to the launch of a new telephone banking service (First Direct) in 1989 as an example of the successful anticipation of changing customer needs. In the survey by Abercromby and Hall (1994) of U.K. life insurers, all respondents said that meeting an iden-
O'Brien: Product Innovation

tified market need was the most important reason for launching a new product.

2.3 Intermediaries

The large proportion of insurance sales though intermediaries means they may have an important role in the industry. Iqbal (1988) explains the philosophy of Liberty Life of South Africa as regarding as its customers not the public but its wide network in the broking fraternity, its full-time agents, and branch managers. It attached by far the most importance to the brokers and agents from whom it obtained the greatest intelligence. Coupled with a commitment to innovation, this has helped the company's growth.

The following example from Iqbal (1988) shows the importance of distribution: the relevance of the form of distribution (direct sales or broker) and the way in which office size can be significant through its distribution power. He describes how a small unit-linked office introduced the first unit-linked whole life policy in 1973. Volumes of business were small. In 1977 the largest direct selling office began the same idea, with a number of new product features. Sales picked up after 12 months of sluggish sales. Then a new broker-only office introduced a variation on the product, but "its sales were modest because intermediaries had not yet accepted unit-linked whole life plans as readily as direct salesmen."

Gupta and Westall (1993, pages 32-33) describe the introduction of universal life policies in the U.S.:

The product was originally introduced as a simpler product than conventional whole life which would have more consumer appeal, and hence be easier to sell ... Because it was easier to sell, margins and commissions were lower, but there was no attempt to change the distribution from that used for more complex life assurance. Thus there was a mismatch between the product and distribution, and the product was not a success. After some time of non-success the product was withdrawn, and a new more complex version with higher margins and commission was introduced. Thus the product and distribution were matched and the product became a success.

Some would question whether more complexity and higher costs represented a success for consumers. The ability to link the new prod-
uct with distribution system is clearly crucial, and the example illustrates the problems if the focus is on the end-customer to the exclusion of the distribution channel.

Bradshaw (1995) refers to the salesman as the single greatest source of information for product development; however, his timescale is the date of the next commission payment, so there is the potential for conflict with the longer-term perspective of the company. Oliver (2000) refers to addition of a loanback facility to personal pensions, sales through independent financial advisers having required this. However, he comments that rather than being a helpful innovation, this was a great complication and rarely used.

Her Majesty's Treasury (2001) said that it has been possible to over-engineer products to appeal more to advisers than customers. This typically introduces such complexity that it can be almost impossible for people to select financial products without advice. The Faculty and Institute of Actuaries (2001) also commented that the active intermediary market is one reason why innovation in charging structures is restricted.

Milton (1996), in a survey of U.S. life insurers, showed that the main current driver for product development was agents; however, companies wished to see a much higher role accorded to consumers. This would also need better market intelligence, which many thought was deficient.

2.4 Government

2.4.1 Regulatory Restrictions

Clearly there are some areas where government regulates the form of products it will permit to be marketed. This constrains innovation. It means that there are opportunities for firms when regulations change and opportunities for firms to anticipate and influence trends in regulation.

The degree of product regulation varies between countries. For example, in the U.S., the National Association of Insurance Commissioners (NAIC) adopted the Variable Life Insurable Model (VLI) regulation, which establishes certain mandatory policy design characteristics and policy provisions. The regulation also covers the qualifications of a company to conduct VLI business, operations of VLI separate accounts, reserve requirements, and the information to be sent to applicants and policyholders (Black and Skipper, 2000, page 103).
More generally, U.S. states require (or will accept) life insurance contract forms that contain, in substance, provisions as recommended by NAIC, relating to, for example, the grace period, apportionment of dividends, surrender values, and options and policy loans (Black and Skipper, 2000, page 952).

In many countries of the European Union there were rules requiring companies to submit proposed new policies to the regulator before approval. This was widely regarded as putting a brake on innovation. For example, the German motor insurance industry had standardized contracts, regulated premiums, and uniform calculation methods. Finsinger, Hammond, and Tapp (1985) highlighted the greater division of risk categories in the less regulated U.K. market compared to Germany, where product variety was less. When new EU Directives were implemented in 1994, there was greater freedom for insurers, with the abolition of the prior approval conditions. We then saw German firms moving to introduce new rating factors such as the age of the car, the mileage covered, and whether the car was garaged (Wein, 2002).

Finsinger, Hammond, and Tapp (1985) also referred to a much freer market for life insurance in the U.K. compared to Germany. The U.K. market used a greater number of risk factors and greater product variety. They said (page 92): "Due to the strict regulation in the West German life insurance market there is much less innovation in establishing new products as contract terms are standardized by the regulatory agency." The constraints were reduced in the 1990s.

There are also stimuli to change arising from the regulations on how to calculate provisions for certain product types. We are expecting changes in insurance accounting as the International Accounting Standards Board works to establish a new standard for accounting for insurance contracts; this may lead to behavioral responses, including changes in the direction of product development effort.

2.4.2 Market Failure Reasons for Intervention

Governments may intervene in product development as a result of a view that the market has not performed adequately. This is not a paper on regulation as such, but it is worth emphasizing these concerns.

Her Majesty's Treasury (2001) indicated (paragraph 22): "Too many [U.K.] financial services firms have been cynical about their customers' interests—more interested in devising creative ways of hiding profit centers than building real value for their customer." This, it argued, has led to complex products with confusing detail hidden in the small print. A recent intervention in the U.K. has been the introduction of
stakeholder pensions, a cash accumulation product where 75% of the fund at retirement is to be converted into an annuity. The product design was essentially done by the government, who imposed a maximum charge of 1% p.a. of the fund. The aim is to provide a simpler and better value-for-money pension product (with no front-end loading) than had typically been available. The new rules on participating policies under stakeholder pensions have led some companies to develop new types of product. For example, Leahy (2001) describes how one office decided not to offer the guarantees traditionally offered under participating policies, arguing that such guarantees were costly to provide and doubting whether policyholders really want, at the outset, a spot guarantee applicable at one point in time, perhaps many years ahead.

2.4.3 Taxation

Government tax policy may lead directly to the introduction of new products. For example, in the U.S., the Tax Reform Act of 1984 created a niche market for single premium life products, only for this to be ended with a new definition of “modified endowment contracts” in legislation in 1988. This led to companies raising the interest rates they were offering, and “Aggressive companies like Executive Life pushed the junk-bond mania beyond the limits and failed” (Baldwin, 1994, page 4).

In France, Predica was instrumental in developing *bons de capitalization*. These were long-term contracts mostly in the form of single premium policies and, until 1988, were the only products that enabled French savers to accumulate capital and to get tax relief at the end of the period. In the U.K. the government has introduced new product types with tax advantages designed to promote saving (Tax Exempt Special Savings Accounts, Personal Equity Plans, Individual Savings Accounts).

Financial services firms may be uncomfortable with some of the intrusions by government. It is up to them, however, to make the most of it; for example, by designing features of their products around the basic requirements; by identifying appropriate market needs that can be met; and by using the optimum service delivery mechanism. For a number of firms, these products have been a significant part of their sales.

2.4.4 Other Policy Objectives

Governments may also have objectives relating to health and pension provision or other benefits (such as unemployment benefit) and rely on providers to play a part in this.
For example, U.S. insurers have health insurance products that provide coverage that supplements the benefits provided by governmental health insurance plans. For example, the Medicare wraparound policy provides benefits that cover the deductibles and coinsurance amounts that individuals must personally pay under Medicare (Health and Skipper, 2000).

In Germany there were reforms of the pension system in 2001, which has led to the opportunity for life insurers to develop Riester policies. There were also incentives to provide private disability cover as a result of reductions in public disability benefits available from the state pension system. Daykin and Lewis (1999) refer to developments in a number of countries where there has been a cutback in state pension benefits and encouragement for private sector financial services firms to enter the market with products designed accordingly.

2.5 Technology

Technology plays an important role in innovation, both from a marketing and administrative perspective. For example, universal life insurance in the U.S. could not be marketed without the aid of computer-generated customized illustrations; neither could it be administered effectively if insurers relied on manual as opposed to computerized procedures.

The internet is now bringing further innovations, although there have been a number of comments that insurers have been slow in adopting e-commerce, e.g., Bukowski (1999). Rakovska (2001) refers to general insurance having made a number of advances, but internet development is slower where financial services depend on the aid and skill of individual agents. In life insurance, Pugh (2003) refers to term insurance being most commonly sold on the internet, this being a relatively simple product; there are several attempts to sell annuities but hardly any for universal life, where the complexity requires agent intervention.

It is not just a case of the internet being an alternative distribution system: it needs other changes. Underwriting term insurance sold through the internet has led to the development of new underwriting tools, the goal being to have the trade-off between price and convenience at an acceptable level to minimize anti-selection (Pugh, 2003).

The above innovations could not have arisen without the technological revolution. We may speculate that further changes in technology will lead to continued innovation in product design, terms, distribution, and administrative processes. It is also plausible to think that geopolitical regions less technologically advanced can learn from the growing
pains of those more advanced, although the cultural and regulatory circumstances of the former can be expected to produce outcomes relevant for their situations rather than necessarily replicating what has happened before.

We should also remember, from a slightly different perspective, the influence of developments in financial economics, including the Black-Scholes and Merton papers on option pricing in the 1970s. Miller (1992) comments: "The extent to which academic thinking and criticism prefigured the great wave of financial innovations of the 1970s and 1980s is still too little appreciated."

An example of a new derivative-based product is a guaranteed equity product, which can be offered by life insurers or other financial institutions. This type of contract has in mind the wishes of many customers to receive the benefits of equity returns, but with the safeguards of a guarantee. Nevertheless, the amount foregone by customers to pay for these benefits may be judged high in relation to what are thought to be typical levels of risk-aversion that lead to the wish for guarantees (Cantor and Sefton, 2002). Similarly, we note the comment by Brizeli (1999) that Canadian life offices offering a segregated fund with a guaranteed maturity benefit (GMB) find that few reinsurers are prepared to accept the GMB risk at a marketable price. This exemplifies that the process of understanding and meeting customers' needs is not easy.

2.6 Comments

We have seen from the above that there are different ways in which the ideas for innovations can arise. There may be external stimuli from government and technology as well as initiatives coming from firms themselves. And in some cases the driver will be some new aspect of market potential. It is worth highlighting a number of points from the discussion:

- It is common for financial services firms not to involve customers deeply in the product innovation process, hence with risks that planned sales will not materialize;

- The success of an innovation may well depend on how the firm copes with distribution issues;

- Some regulation can stifle innovation, but there is also the potential to produce innovation geared to the particular circumstances of the rules, including tax rules; and
O'Brien: Product Innovation

Technology has been a large stimulus to innovation, involving new designs rather than merely replicating existing processes and products.

We can also see here the potential for innovation in one area does not always translate into other areas of the industry. Clearly there are cases where a product is developed to meet local regulatory needs, including tax incentives and to link in with public benefit systems, and these cannot be easily replicated. Furthermore, banks and general and life insurers interact with their customers in different ways. Therefore, innovations may be acceptable in one context but not another. For example, a bank may offer a new product where it has discretion to vary the terms of the contract over time. Customers' experience of banks may be such that this is (or is not) acceptable, whereas they may take a different view of an insurer introducing such a product because of different past experiences of contacts with insurers. Therefore, the relationship between the customer and the financial services firm (including its distribution network) may affect the acceptability of an innovation.

3 Implementation

3.1 Implementation Processes

We here consider how firms implement new ideas and attempt to identify the factors that are more likely to lead to success.

The general marketing literature describes the key stages in the product development process. For example, Booz, Allen, and Hamilton (1982) refer to:

- New product development strategy, followed by
- Idea generation, then
- Screening, then
- Development and testing, and finally
- Launch.

McGoldrick (1994) presents a more detailed process for financial products.

Stern and Whittemore (1998) describe the following steps for use by insurers:
• Company objective and strategic direction, followed by
• Idea generation/screening, then
• Preliminary design, then
• Detailed design and economics, then
• Detailed systems development, then
• Go to market, and finally
• Performance tracking.

Ritzke (1992) refers to product development in small insurance companies, which tend to make greater use of external resources such as reinsurers, auditors, and actuarial consultants. This shows that firms need to adapt processes to their own circumstances.

Product development processes have also been changing rapidly: a U.S. survey by Milton (1996) showed that 40% of life insurers had made significant changes to their process within the previous year.

3.2 Speed

Two surveys indicate that implementation of a new product idea is quite slow. In the U.S., Milton (1996) showed that the average time for implementing a new life insurance product from idea to selling was six to twelve months in most companies. The need for prior approval in most of 50 separate jurisdictions for insurers carrying on business nationwide can be a significant factor, although this could change if federal life insurance charters become available in the future. The process has been longer still in the U.K.: Abercromby and Hall (1994) showed that the length of time from the idea/concept being accepted by senior management (not from the idea itself) to the first policy being issued was twelve to eighteen months in the majority of companies. In both the U.S. and the U.K., there were concerns that these time lags should be shorter.

Milton (1996) identified the following features of firms that were taking under six months:

• They were product-driven companies with a strong belief that ongoing product development was the best way to compete;
• They had stronger senior management support than average for product development;
• They had a formal implementation plan in which marketing materials were developed early, and there was good communication between departments;
• Most of the product development efforts were used for simple line extensions and price revisions; and
• They were more satisfied than average with respect to their strategic planning process and product strategy.

The benefits of rapid product development are largely intangible, however, according to Drew's (1995) survey of U.S. banks and insurers. Such intangible benefits include projecting a more innovative image with customers and appearing to be more competitive. The correlation between speed and revenues was, however, marginal. Competitive factors and customer pressures were driving the search for quicker practices. His recommendations to managers were:

• Commitment to speed must come from the top and be promoted throughout the organization;
• A proactive approach to technology is needed;
• People must be motivated and rewarded;
• Strategies and goals for accelerating products to market must be developed; and
• New style organizational structures must be created, and a new mindset of fast paced competition must be developed.

Nevertheless, for a genuinely new product, a shorter development time may be critical to avoid being preempted by competitors. Daniel and Tomkin (1999) illustrate this with First Direct, the telephone bank in the U.K.

3.3 Cross-Functional Teams

Many writers have highlighted the benefits of cross-functional teams in product innovation. Ittner and Larcker (1997), in a general survey, reported that cross-functional teams could increase the amount and variety of information to design products and help spot problems earlier in the process. Griffin (1997) found that cross-functional teams can be beneficial in reducing product development cycle times, although there were interactions with other factors.
Sethi, Smith, and Park (2001), in a survey, reported on 141 cross-functional product development teams and found that innovativeness was positively related to:

- The team having its own identity, rather than retaining old functional objectives, ideas, and stereotypes (so-called high superordinate identity);
- Encouragement to take risks;
- Customers' influence; and
- Active monitoring of the project by senior management. Beyond a moderate level, social cohesion among team members has a negative effect on innovativeness (groupthink often arises in highly cohesive groups).

Olson et al. (2001), in a general survey, stressed that the importance of functional co-operation varied by stage of project and with the degree of innovativeness. Henard and Szymanski (2001) found, from a questionnaire survey, that functional diversity can play an important role in some tasks, such as the generation of ideas, but may be less important in some other areas.

Johne (1993) reports the increasing use of project or venture teams by general insurers, though in a number of companies these were being used by top management as a mechanism to drive through changes within and between traditional fiefdoms.

Bradshaw (1995, pages 6–7) considers the potential conflicts: “Marketing would want a new product to be the most innovative to create the most impressive press coverage. Sales would want the cheapest, best, quickest to get commission out of ... . Actuarial [departments] go for complicated products.” Perhaps the key is balancing the roles so that there is no undue domination by any one and having high superordinate identity in the project team.

3.4 Internal Marketing to Distribution Channels

Cross-functional teams help the development process, but success can also depend on the developers selling what they are doing to the distribution channels and other parts of the firm. Johne and Pavlidis (1996) found that the more innovative banks (for derivatives) go to great lengths to explain developments to the dealers and brokers who will be selling their new product. In a general survey, Atuahene-Gima (1997)
O'Brien: Product Innovation

stresses that the sales force's commitment to a new product cannot be taken for granted and suggests a number of factors that influence this: factors include their perception of the firm's commitment to new products and their problem-solving style.

Strieter et al. (1997) reported several product managers in banks expressing concern about co-ordination problems when introducing new products. Banks use many locations, and when branch offices and affiliates did not understand the product or the reason for its introduction, the chance of success was lower. The areas identified as problematic included lack of training for and educating of line offices, the changing banking environment, and having a diverse international sales force. Several product managers said they had the general responsibility for new product introductions without the necessary authority to implement sub-programs effectively. This is evidence that different parts of the financial services sector need to cope with their specific distribution issues if they are to be successful innovators.

3.5 Formality

Is innovation a formal or an informal process, and is this related to success?

Vrakking and Cozijnsen (1993), in a general survey, describe the classical approach to innovation as an individual process, ungovernable and uncontrollable; more or less accidental; and unpredictable. They contrast the modern approach which views it as a multi-disciplinary group process; guided and controllable; more than just adaptation of an existing product; a process by jumps and starts, but predictable. They set out the four main phases of innovation as generation of ideas, initiation, implementation, and incorporation. They discuss these in some detail. For example, the conditions for successful implementation were:

- Demonstrate the need for the innovation;
- Make the innovation integral for your organization (department);
- Use a step-by-step approach;
- Ensure a fast execution of every step; irreversibility; and sound management of the project;
- Reserve in advance the resources for implementation;
• Create an effective form for the broad participation of your department;
• Make sure there is a balance between substantive and process innovations; and
• Make sure you take enough time to implement the innovation.

There is evidence that product innovation in financial services is often informal. Edgett (1993), studying building societies, indicated that only 13% had written guidelines for new product development; 76% had informal or ad hoc approaches. There can be benefits, however, from cumulative experience in product development. Riek (2001), in a general survey, emphasizes the merits of checklists to capture the lessons of new product development and ensure that minor tasks are not missed; this complements the cumulative experience of the project team. This links with the finding of Johne and Pavlidis (1996) that the administrative procedures of the more innovative banks (in derivatives) were more standardized and formal than those of the less active innovators.

Johne (1993) found that product development in insurance had become more systematic, with clear evidence of a sophisticated approach to analyzing markets and greater marketing department input. Poor market information usage, however, contributed to the failure to envisage more radical amendments to what was offered. Stern and Whittemore (1998), considering U.S. insurers, found that efficient companies create a project plan and implementation standards based on internal experience and best practice derived from other companies. Companies can improve their product development efficiency by devising benchmark performance measures and identifying best practice product development and project management techniques.

De Brentani (1993) examined 106 new services from 37 financial services firms, about half of which were successful. He found that the most significant positive factors were:

• A supportive, high-involvement environment, with good communication throughout and support from top management; and
• A formal and extensive launch program.

Milton (1998) refers to good product developers having a deliberate process that they use to generate new ideas, balancing reactive approaches (such as listening to customers and agents) and proactive approaches (conducting market research and competitive analysis). Gold-
stein (2002) describes one life office's product development process, emphasizing the benefits of a structured approach.

Formality appears beneficial, but we should recognize that product development is not a routine process. Bradshaw (1995, page 13) describes it as "a dynamic affair, running from here to there, troubleshooting, communicating, motivating, learning, deciding, informing."

Some parts of the process may benefit from formality more than others. Avlonitis, Papastathopoulou, and Gounaris (2001) concluded that the degree of formality did not seem to bear a significant impact on achievement of management objectives in relation to new-to-the-market services (where perhaps there was less experience to produce a formal process), though it helped for other innovations.

We should also emphasize that many new products do not succeed. Boulding, Morgan, and Staelin (1997), in a general survey, say that senior managers often remain committed to a losing course of action. They suggest either commitment to a predetermined decision rule or introduction of a new decision-maker at the time of the stop/no stop decision. Including this as a formal part of the plan was also highlighted as a success factor by Cooper (1999).

We may also add that formality can play a part in ensuring that cross-functional teams are established, and that there is effective selling of the development to the distribution channels. Successful firms can determine what is needed in innovation and can formalize and continue doing it.

4 Impact

4.1 Success Criteria

In looking to find what works well in an innovation, what do firms consider to be the objectives they are seeking from product innovation? This may not be straightforward. Johne (1993) indicates a reluctance on the part of insurers to use explicit formal criteria for evaluating products. When aims were articulated by top management, these were typically in terms of sales targets over what had been achieved in the past.

Bradshaw's (1995) case study of a whole life and critical illness product included the comment (page 20): "There was no explicit sales target. However, there was a corporate target of doubling business within three years up to the end of 1995 and an implicit target of at least maintaining our share of the IFA [independent financial advisor] market in these
fields ... These seem as 'accurate' as any other projected sales figures I have seen."

There can be a number of criteria for evaluating product development projects. One might suggest that the key test is the impact of the innovation on the market value of the firm. In a general survey, Chaney, Devinney, and Winer (1991) find that the announcement of a new product increased the share price of a firm by about 0.75% on average over a 3-day period. In practice, however, more concrete measures related to the innovation are required; a longer-term perspective is appropriate. Share price is not relevant to some types of organization.

Hultink and Robben (1995), in a general survey, find that, in the short term, firms emphasize product-level measures of success, such as speed to market and whether the product was launched on time. The efficiency of the product innovation process itself is important, and Stern and Whittemore (1998) describe some significant differences between U.S. insurers in this respect. In the long term, Hultink and Robben find that the focus is on customer acceptance and financial performance, including attaining goals for profitability, margins and return on investment. Four factors were perceived as equally important for short-term and long-term success: customer satisfaction (the most important), customer acceptance, meeting quality guidelines, and product performance level.

Johne (1993) considers success criteria. Market share success says that customers are responding to the new product; but in addition to market-based measures, a firm needs supply measures to ensure that, as a supplier, the firm is managing to meet customer needs profitably.

This raises the issue of whether the marketing and financial managers both have an input to setting objectives and monitoring success. Marketeers may not be aware of the benefits of using actuarial techniques where applicable. Indeed, marketing managers may also have a different stance from consumers.

Edgett and Snow (1996) describe the performance measures used by the Canadian financial services industry (banks, insurance companies, trusts, and credit unions). These measures were derived from a questionnaire study addressing measurement issues for customer satisfaction, product quality, and new product success. For each of these three sections respondents indicated which of several possible measurement approaches they used. Table 1 shows the proportion of respondents using the most frequently used measures and presents the measures regarded as most helpful, based on the average score given by the respondents.
<table>
<thead>
<tr>
<th>Table 1</th>
<th><strong>Product Innovation Performance Measures</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Most Frequently Used</td>
</tr>
<tr>
<td><strong>Customer Satisfaction</strong></td>
<td>Increase in number of customers (92%)</td>
</tr>
<tr>
<td></td>
<td>Increase in portfolio dollars (85%)</td>
</tr>
<tr>
<td></td>
<td>Complaint measurements (80%)</td>
</tr>
<tr>
<td></td>
<td>Market share (75%)</td>
</tr>
<tr>
<td><strong>Product Quality</strong></td>
<td>Increase in sales (96%)</td>
</tr>
<tr>
<td></td>
<td>Increase in income (88%)</td>
</tr>
<tr>
<td></td>
<td>Reduced operating costs (85%)</td>
</tr>
<tr>
<td></td>
<td>Increase in market share (84%)</td>
</tr>
<tr>
<td><strong>New Product Performance</strong></td>
<td>Sales growth against objectives (96%)</td>
</tr>
<tr>
<td></td>
<td>Total sales (units/revenue) (94%)</td>
</tr>
<tr>
<td></td>
<td>Number of new customers (94%)</td>
</tr>
<tr>
<td></td>
<td>Profitability (93%)</td>
</tr>
</tbody>
</table>

**Notes:** The results in parentheses in Column 3 are based on a scale from 1 (extremely helpful) to 7 (not at all helpful).
Some of the conclusions from the study were:

- Most institutions have begun to use multiple measures for determining success and failure;
- Most are unhappy with the measurement techniques in use;
- Most need more customer input;
- Clearer objectives need to be set for new products; and
- Benchmarking needs to be applied to the measures of success.

4.2 Indirect Benefits

Firms may also gain from indirect benefits when they innovate. Eas­ningwood and Percival (1990) studied 18 examples of product innovation in financial firms; all were at least minor successes, with the bias toward the more successful. The firms identified six indirect benefits, in order of importance, were:

- Corporate reputation: e.g., Bank of Scotland's introduction of its Home and Office Banking System is thought to have improved its image as well as that of Scottish banking in the financial world, indicating that clear technological advances were taking place;
- Existing customers buying more existing products: managers argued that the new product adds to the range and so makes it more likely that a customer will see the firm as able to satisfy all his/her financial needs. For example, Cornhill Insurance Group found that the introduction of an investment-linked flexible unit-based permanent health insurance package stimulated broker interest in its existing products;
- Improved new product development capability: the system developed to launch the new product can provide a platform to help future new products; the other benefit is the extra expertise gained;
- New customers buying existing products: an area where large organizations with a significant number of product lines can have an advantage, although they do not always prove to be successful;
- Improved loyalty of existing customers: American Express helped to reinforce the loyalty of its members to the company when it
built ATMs in railway stations and airports throughout the world—travelers need financial services when on the move, and the convenient provision of these services helped tie members to the company; and

- Helping redirect the company in a new direction: a new product can help the company emphasize markets in which it was not strong (or not present). For example, a switch in the market's perception of Thomas Cook was achieved through its provision of a foreign exchange service.

The survey found that the overall value of indirect benefits was only a little less than the direct financial return from the product. There was also a clear association between the success of the product and the number of high indirect benefits.

Daniel and Tomkin (1999) considered three innovations in banking: Bank of Scotland’s screen-based banking, First Direct, and Mondex. They identified, in addition to product-level benefits, some firm-level benefits. An example was the development of new competencies; for example, Bank of Scotland developed a sales force to close the sale, which could be used in other situations. Others were:

- Bank of Scotland; access to new market (England), bank seen as innovative, perceived positive impact on share price, platform for future development;
- First Direct (which introduced telephone banking in the U.K.): seen as innovative, dominant brand established, referral sales, building loyalty; and
- Mondex (a smart card producer): seen as a technology innovator; major brand established, formed key alliances, a platform for future card-based developments.

Such benefits should be recognized in deciding whether to go ahead with, and measuring the success of, a product development. Daniel and Tomkin warned of the difficulties because company-wide benefits can be unexpected or affect intangibles (e.g., reputation) or arise over time frames longer than that used for the project evaluation.

4.3 Distinguishing Features of Success

Here we review some surveys that have looked specifically at what distinguishes successful from unsuccessful innovations. There are two
particular concerns in these evaluations. First, much of the literature is from surveys of marketing managers. Others may have a different perspective. For example, the possible conflicts between marketing and financial directors have not been explored fully. Second, a product may be successful in the short term, but fail over a longer time frame, depending on the economic environment.

Adams, Day, and Dougherty (1998), in a general study, refer to research having demonstrated that the top success factors for new products are:

- A differentiated product that offers superior customer value; and
- A strong market orientation reflected in a thorough understanding of customers' needs and wants, the competitive situation, and the market environment.

Cooper (1999), in a general survey, delineated eight common denominators of success:

- Up-front homework pays off;
- Build in the voice of the customer;
- Seek differentiated, superior products;
- Demand sharp, stable, and early product definition;
- Plan and resource the market launch, ... early in the game;
- Build tough go/kill points into your process;
- Organize around true cross-functional project teams; and
- Build an international orientation into your new product process.

Easingwood and Storey (1991) surveyed marketing managers in U.K. financial services firms to give their views on 77 new financial products identified from the trade press: 64 were judged to be successful, 13 unsuccessful. Respondents to the survey rated 43 attributes (in 9 factor groups) on a 9-point scale (1 = very much hindered the success of the product; 9 = very much helped the success of the product). The significant (at the 1% level) differences in scores between successful (S) and unsuccessful (U) products are shown in Table 2.
Table 2
Attribute Scores of Successful and Unsuccessful Products

<table>
<thead>
<tr>
<th></th>
<th>Mean Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
</tr>
<tr>
<td>There was quality in the delivery of the service</td>
<td>7.0</td>
</tr>
<tr>
<td>The organization had a reputation for quality</td>
<td>7.7</td>
</tr>
<tr>
<td>A good fit between product, delivery system, and organizational structure</td>
<td>6.8</td>
</tr>
<tr>
<td>Product was considered a quality product compared to competitive products</td>
<td>7.1</td>
</tr>
<tr>
<td>Product had a strong brand image</td>
<td>6.9</td>
</tr>
<tr>
<td>Communication strategy was consistent with marketing strategy</td>
<td>6.8</td>
</tr>
<tr>
<td>Delivery was supported by an extensive branch network</td>
<td>6.9</td>
</tr>
<tr>
<td>There was investment in the training of the staff</td>
<td>7.0</td>
</tr>
<tr>
<td>The product offered unique benefits to the customer</td>
<td>6.9</td>
</tr>
<tr>
<td>The product was considered innovative</td>
<td>7.0</td>
</tr>
<tr>
<td>Product was conceived quickly and implemented in response to market</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Notes: The differences in scores between successful (S) and unsuccessful (U) products are significant at the 1% level. Factors correlating with overall success were overall quality, differentiated product, product fit and internal marketing, and use of technology.
We have seen the success in telephone banking, followed by its use in personal lines general insurance. In the U.K., the latter was introduced most successfully by Direct Line. They then established Direct Line Life, for telephone selling of term insurance, but sales were negligible and the company closed. This is a dramatic example of an innovation succeeding in one area, failing in another. This reflects the genuine differences between banking, general insurance, and life insurance products. Life insurance is more complex than other financial services, and the need for agent intervention is more crucial. The different distribution systems of banking, general, and life insurance are bound to influence innovation. While we can have a generic factor for innovation, which is that there needs to be a good fit between the product and the distribution systems, this has different implications in the various subsets of the industry. We have seen similar issues with the internet, which is most easily adapted for banking and then general insurance.

5 Conclusions

5.1 Background Factors

What does the literature tell us about why innovation succeeds or fails? First, we note some background factors that can be influential in either encouraging or discouraging innovation. A strict regulatory environment can reduce firms' incentives and motivation to innovate. On the other hand, if regulations encourage competition and there are profit incentives for firms that can distinguish themselves, innovation is more likely (Finsinger, Hammond, and Tapp, 1985; Wein, 2002).

Innovations may be stimulated by some specific regulatory event, such as new product terms or designs being permitted, new tax rules, or changes in public sector social security programs that interact with private sector provision (Section 2.4). There may also be some adverse event that leads companies to change their products (Section 2.1.2). Technology also can lead firms to produce new product designs, processes, and delivery systems.

5.2 Firm-Specific Criteria

Given the above background factors, there are some common themes that help explain why some firms are more successful than others. We highlight four reasons, which are linked.
First, some firms have a passion for innovation, which is an integral part of their strategy (Section 2.1.3). Not only will innovation be a key objective of such firms, but also the resulting new products can be expected to be more different compared to firms where innovation is a lesser priority. The surveys in Section 4.3 indicated that differentiated products were a distinguishing feature of success. Furthermore, these firms will also be determined to implement using development processes that are known to work rather than designed ad hoc.

Second, the successful firms recognize the importance of internal co-ordination, with a good fit between the product, the delivery system, and the organization (again as evidenced in Section 4.3). This means having cross-functional teams to develop the product. Such teams are, of course, common, but we have also seen differences in the way they can work, with, for example, benefits if firms ensure that old functional objectives do not intervene to cause friction (Section 3.3).

Third is the adage of remember the customer. Success depends on both sales and costs, but if the customer is forgotten and the sales do not appear, the development is a waste (Oliver, 2000). We have seen that success is, at least in part, a matter of sales. The surveys in Section 4.3 demonstrate that product differences make a significant difference to the likelihood that the innovation will be a success. We should also bear in mind, however, that if there is a proliferation of similar products, the customer perspective could be that the benefits are barely worthwhile.

Last, we emphasize the benefit to firms if they have a formal product development procedure (Section 3.5). Up-front homework, planning, and tough go/kill points are examples of steps that can enhance the likelihood of success. Perhaps too many unsuccessful products go to market because there were not appropriate go/kill points in the development process. Formality can also include having formal objectives, which may include indirect benefits.

We also see benefits if firms are innovating regularly. One aspect of this is establishing a reputation as an innovative firm. In addition, innovation is a learning experience. Successful innovative firms can learn how to carry out internal co-ordination. These firms formalize the innovation process. Success is more likely if firms have innovation at the heart of their strategy and give priority to the needs of customers.

5.3 Success Factors

What works well in one situation may not work in another. In some cases we find products developed to meet particular regulatory or tax circumstances, where they are of narrow application. We still have
countries where particular likes or dislikes may hinder rapid international innovation. For example, many continental European countries have a long-established preference for participating life insurance policies. Whatever the merits of unit-linked products, the inroads they have made into those countries have been slow.

Rapidly increasing technological developments will help innovation transmission. There are, however, differences between subsets of the financial services industry that react differently to innovation. The different distribution channels used pose particular issues.

We have seen that use of the telephone and internet was earlier for banking and general insurance than for life insurance. The subsets of the financial services industry have adopted solutions for their customers that reflect historical differences, although these have genuine effects: for example, a bank with a large branch network will take this into account in deciding how to develop new products, and must incorporate its distribution system in the development process (Johne and Pavlidis, 1996; Strieter et al. 1997).

On the other hand, life insurance is complex, and its long-term nature may lead customers to seek agent advice, especially when they may not be in regular contact with their provider in the future (unlike banking), and where they may suffer penalties if they subsequently decide to terminate the contract. This has led to telephone and internet sales of life insurance being slower to develop than in the case of banking and general insurance.

At the end of the day, customers determine whether a product is worthwhile. The surveys in Section 4.3 indicated the characteristics of products, in particular those differentiated with high quality and customer value, that made innovations worth the effort and the risk involved. While this is gratifying, we should also recall the role of the intermediaries, who may have a strong role in the development process, which may lead to products over-engineered for their benefit (Gupta and Westall, 1993; Her Majesty’s Treasury, 2001), with resulting question marks about whether the development was worthwhile (Oliver, 2000).

The surveys reviewed in this paper demonstrate that innovation can take place in a number of different ways. There is indeed no unique secret for innovation. We have also seen that what works well in one situation may not work elsewhere. It is also clear that there are several common factors that tend to lead to success, and there are benefits to firms in identifying and adopting these best-practice principles.
References


Abroe, M. "Long-Term Care Topics." Product Development News 48 (1999): 1-7. (Published by the Society of Actuaries.)


Phased Retirement for Defined Benefit Plan Participants

Patricia L. Scahill* and Jonathan Barry Forman†

Abstract†

The demographic makeup of the U.S. workforce is changing. The population between ages 55 and 64 is projected to increase significantly by 2020, but employment rates for this age group have not been increasing. Employers will likely need to encourage critical employees in this age group to delay retirement. Phased retirement is one tool for delaying retirement, while also not continuing full-time employment, so it can be a compromise for employers and employees.

Both Congress and two administrative agencies have begun to consider changes in pension laws and regulations that would be needed to accommodate phased retirement for employers who sponsor defined benefit plans. This paper discusses some of the impediments in the current legal framework and changes that could be made without diluting participant protections. This paper also discusses aspects in the actuarial calculation of retirement benefits that impact the financial neutrality of a phased retirement program.

Key words and phrases: ERISA, delaying retirement, pension law, actuarial neutrality, financial neutrality, part-time work

*Patricia L. Scahill, FSA, JD, is a plan review consultant at American United Life Insurance Company®.
Ms. Scahill's address is: American United Life Insurance Company®, One American Square, P.O. Box 368, Indianapolis, Indiana 46206-0368, U.S.A. Internet address: Pal_Scahill@aul.com

†Jonathan Barry Forman, J.D., is a professor of law at the University of Oklahoma.
Mr. Forman's address is: College of Law, University of Oklahoma, 300 Timberdell Road, Norman, Oklahoma 73019, U.S.A. E-mail: jforman@ou.edu

‡The views expressed in this paper are those of the authors and they do not necessarily reflect the views of their employers.
1 Introduction

As America ages, the demographic makeup of the workforce will change. The U.S. Bureau of the Census projects that, between 2000 and 2020, the number of people between ages 55 and 64 will grow 73.5%, going from 24,276,000 in 2000 to 42,107,000 in 2020. At the same time, the population from age 25 to 54 is projected to remain level. From 1995 to 2002, employment rates remained level for men ages 55 to 61 and rose for women ages 55 to 61, as well as for both men and women ages 62 to 64. Labor force participation rates, however, are much lower for both men and women ages 55 to 64 than for those ages 25 to 54. In 2001, 91% of men and 76% of women ages 25 to 54 participated in the labor force compared to 68% of men and 53% of women ages 55 to 64. If the current labor force participation rates continue, the pool of available workers will decline as the population ages. Consequently, employers will need to find ways to retain their productive older workers.

Factors influencing the employment rate among people age 55 and older include economic conditions, Social Security benefits, and the prevalence and design of private pensions (Purcell, 2002). Since the repeal of mandatory retirement, phased, or gradual, retirement is beginning to replace cliff retirement where a person retires from the workforce and does not return. According to Watson Wyatt (1999a, page 2) "phased retirement is any arrangement that enables employees approaching normal retirement age to reduce their work hours and job responsibilities for the purpose of gradually easing into full retirement."

Many older Americans are staying in or re-entering the workforce in part-time and contingent work situations; see Herz (1995), Quinn (1999), and Wiatrowski (2001). Sixteen percent of the companies participating in a 1999 Watson Wyatt survey offered phased retirement programs (Watson Wyatt, 1999b, page 9). According to one estimate, roughly one-third of older workers leave their long-held career jobs and begin new jobs that serve as a bridge to full retirement. In another 1999 Watson Wyatt survey, phased retirement was more prevalent at firms in which workers have an average age of 45 or higher (Watson Wyatt, 1999a, page 3).

1The statistics cited above are taken from Purcell (2002).
2Mandatory retirement is still allowed for certain highly compensated employees.
Clearly, both employers and employees are interested in phased retirement, but, unfortunately, the U.S. pension system was not designed with an eye toward phased retirement. Many companies face serious legal impediments to establishing an effective phased retirement program. Congress and the administrative agencies charged with overseeing ERISA are aware of at least some of these obstacles. In 2000, one of the working groups of the Department of Labor's ERISA Advisory Council focused on phased retirement. Representative Earl Pomeroy (D-N.D.) and Senator Charles Grassley (R-Iowa) introduced legislation that would have changed federal pension law to allow qualified retirement plans to provide in-service distributions once an employee reaches age 59 1/2 or 30 years of service. In 2002, the IRS solicited "comments on issues relating to 'phased retirement' arrangements under qualified defined benefit plans."6

This paper discusses the impact of phased retirement on benefits provided by a traditional final average pay defined benefit pension plan. It also presents some of the legal, administrative, and public policy concerns raised by phased retirement. An earlier paper by Scahill and Forman (2002) explored in depth the impact of phased retirement on benefit amounts under various payout patterns. They compare common offsets for benefits paid against continued accruals with an actuarially neutral approach that avoids excessive offsets when only part of the benefit is being paid out during phased retirement. That research is not reproduced in this paper.

2 Overview of Phased Retirement

2.1 What Is Phased Retirement?

The definition of retirement is not simple. It is not just the time when an employee stops working and begins receiving retirement benefits. It has become a more complex activity. People often work while receiving retirement benefits. Long-term employees may retire from one career and go on to another career. Some choose to work less—phasing out of their full-time jobs. Other employees leave their career job and work for another employer, usually part time. This job is used to bridge

---


5The Phased Retirement and Liberalization Act (S. 2853/H.R. 4837) (2000). The bill was not voted out of committee in either the House or the Senate.

the transition from full-time work to full retirement and is referred to as a bridge job. Bridge jobs are often different from the person's career job, perhaps requiring different skills in a different industry. Current impediments to in-service distributions from defined benefit plans during phased retirement may force workers to use a bridge job as the phased retirement vehicle rather than a reduced work schedule on the career job.

This paper focuses on the type of phased retirement in which an employee works a reduced schedule on the career job prior to full retirement from that job. It does not discuss other retirement arrangements such as bridge jobs.

2.2 The Importance of Phased Retirement

Phased retirement is not a new phenomenon. It is expected to increase in importance for the U.S. workforce as the large cohort of baby boomers begins to reach retirement age. The baby boomer generation is often defined as those born from 1946 through 1964. The oldest of the baby boomers have already begun to reach age 55—a common age for early retirement eligibility in defined benefit plans. They will begin reaching age 65 in 2011. With increased longevity and more healthy years, many baby boomers will have an active life well beyond age 65.

EBRI's 2001 Retirement Confidence Survey reported that 26% of retirees say they have worked either full time or part time since they retired (ERBI, 2001, page 1 of "Retirement in America" fact sheet). Not all employees will have other sources of income, such as investment income, to supplement their earned income during phased retirement, so they may need to access at least a portion of their pension as they ease into full retirement. The current U.S. pension system does not facilitate phased retirement, especially for participants in defined benefit plans who want to begin receiving pension benefits prior to normal retirement age while continuing to work. The conflict between part-time work and phased retirement is an example of unintended consequences in U.S. pension law. Legislative and/or regulatory changes that allow employers and workers to structure in-service access to retirement benefits will be necessary if phased retirement is to become an attractive alternative to a significant segment of baby boomers.

2.3 Individualized Phased Retirement Arrangements

One advantage of phased retirement is that it allows employees and employers to negotiate individualized part-time work schedules during
phased retirement. One employee may want to gradually decrease the hours worked each year while another employee may prefer to begin with a significant drop in the full-time work schedule and to fully retire after a few years of that reduced schedule. Individualized phased retirement work schedules are no more complicated for the employer to administer than different part-time schedules for workers.

Based on business needs, the employer might designate certain positions as available for part-time hours for phased retirees. In other situations, the employer may want to retain a valuable employee, so it makes business sense to accommodate the employee's desire to ease into retirement with a reduced work schedule. On the other hand, if an employee is only a marginal employee, the employer has little or no motivation to negotiate a special work schedule or transfer the employee to a part-time position. It is legal for an employer to differentiate between employees based on work performance, but it is not legal for an employer to discriminate on the basis of any protected classification, including age.

One possible change in pension regulations, would be to allow payment of partial benefits prior to normal retirement, but the participant must be working a reduced schedule. If the employer refuses a marginal employee's request to phase into retirement by working a reduced schedule, the employee might claim his or her ERISA rights were violated because the employer's refusal to allow phased retirement interfered with early access to retirement benefits. The employer should maintain careful documentation about when and why phased retirement working arrangements are or are not permitted in order to be successful in any such legal challenge.

3 Financial Neutrality of Payouts

This discussion of financial neutrality of phased retirement payouts begins with the premise that phased retirement should be beneficial to both the employer and the employee. One way to assess the impact of a phased retirement arrangement is to weigh the cost and benefit of the arrangement. The primary benefit to the employee is flexibility in designing the transition from full-time work to full retirement and being able to work a reduced schedule at the career job rather than being forced to use a bridge job. The primary cost to the employee is reduced income that results from a reduced work schedule. The employee can use personal savings or in-service retirement benefits to help offset the reduction in compensation during phased retirement.
From the employer's perspective, the main benefit of allowing part-time work during phased retirement is retaining a valued employee.\(^7\) The balance between that benefit and any cost of an individualized phased retirement arrangement will help determine whether a particular arrangement makes business sense. The cost, if any, of paying in-service retirement benefits depends on whether those payments prior to normal retirement are subsidized.\(^8\) Other costs the employer might incur are outside the scope of this paper.\(^9\)

The key to financial, or actuarial, neutrality in pension payouts is for the plan to make a full actuarial reduction for early retirement distributions as well as a full actuarial increase for benefits accruing during continued employment after normal retirement. Actuarial assumptions must also be consistent in the calculation of early retirement reductions, delayed retirement increases, and conversion from the normal payout method to optional payout methods to achieve this actuarial neutrality.\(^10\)

If a participant is entitled to a certain monthly lifetime annuity beginning at normal retirement, the benefit is reduced if it commences at an earlier date because the participant will receive more benefit payments.\(^11\) It is common for defined benefit plans to pay actuarially subsidized benefits to participants who retire prior to normal retirement. These subsidized benefits may have been part of a workforce management program that encouraged employees to retire early as a way of creating opportunities for younger workers through turnover.\(^12\) A full actuarial reduction for early commencement and a full actuarial increase for delayed commencement refer to the situation where the actuarial value of the benefit payouts is the same regardless of when

\(^{7}\)Because the employer is not required to permit a full-time worker to change to a part-time schedule, the employer has no motivation to make phased retirement available to marginal, or even average, workers.

\(^{8}\)Subsidized early retirement benefits are discussed below. It is highly unusual for the actuarial adjustment to payments that commence after normal retirement age to be subsidized.

\(^{9}\)These other costs could include the cost of benefits such as life or health insurance.

\(^{10}\)If a defined benefit plan pays lump sums to phased retirees, this actuarial neutrality may not be possible because of actuarial assumptions currently mandated for lump sum calculations.

\(^{11}\)Under a lifetime annuity, benefits are payable until death. Regardless of when payments begin, payments cease upon the participant's death. As a result, the younger the participant is when benefits begin, the more benefit payments will be received during the participant's lifetime.

\(^{12}\)These subsidized benefits likely have continued from a prior era of generous pension benefits if they remain in the plan. From the authors' experience, plan sponsors today are no longer adding subsidized early retirement benefits to plans.
they begin. Both subsidized early retirement benefits and subsidized delayed retirement benefits are more valuable than the corresponding benefit commencing at normal retirement. Participants receive more valuable benefits by commencing payments at the age when the subsidy is the highest. As a result, the cost of the benefits to the pension plan is highest when the participant maximizes the value of benefits by timing payments to begin when the subsidy is the highest.

If retirement benefits are financially neutral, there will be no financial impact on the employer if the employee decides to supplement his or her phased retirement income with pension plan distributions. This financial, or actuarial, neutrality is achieved when the present value of the expected pension payments does not change because the employee decides to phase into retirement and begins receiving in-service distributions rather than fully retiring immediately. Actuarial, or financial, neutrality also means that the plan is neither better off nor worse off financially because of the in-service distributions an employee receives during phased retirement.

If a participant terminates under a pension plan and is eligible to begin receiving pension distributions at early retirement, normal retirement, or any time between, the employer does not participate in the participant's decision of when to begin pension payments. Benefits paid prior to normal retirement may be subsidized, but the employer does not discourage the employee from receiving these distributions. Similarly, once the phased retirement pattern is negotiated, the employee is free to decide when to commence pension distributions within the constraints of the law. If in-service payouts are permitted prior to normal retirement and those benefits are subsidized, the cost of offering the flexibility of phased retirement to employees who are under the normal retirement age will be higher because of the increased cost of the subsidized retirement benefits. On the other hand, if the retirement benefits are not subsidized, there will be no cost to the employer if the employee receives in-service distributions prior to normal retirement.

4 Final Average Pay Benefit Issues

Most defined benefit plans base benefits on compensation, and most of those plans use a variation of final average pay. As discussed elsewhere in this paper, regulatory and/or legislative changes will be needed for these payments to be available. Of the defined benefit plans surveyed by the Bureau of Labor Statistics, 76% used some form of final or final average compensation in their benefit formulas; see Employee
plan's benefit formula might be 1% of average pay multiplied by credited service. If a five-year averaging period is used, average pay could be the average of the highest five consecutive compensation amounts or it might use the highest five consecutive compensation amounts of the final ten years.\(^{15}\)

Although it is not true in all cases, most employees receive salary increases throughout their working career. As a result, pay in the years immediately preceding retirement would produce the highest average. If the employee begins working a reduced work schedule just before retirement, pay received during the year will be lower than if the employee had continued working full time. Because those final years would produce the highest average if the participant continued working full time, the employee will have a lower final average compensation as a result of phasing into retirement. If the plan defines final average compensation as the average of the highest five compensation amounts during the employee's entire working career, the final average itself will not decline during phased retirement. It will not be as large, however, as it would have been if the final years had been full-time years.

The definition of final average pay clearly has a significant impact on the effect of phased retirement on the retirement benefits payable from a final average pay plan. Internal Revenue Code §401(a)(4) regulations have special provisions for employees working less than full time in a safe-harbor-design plan using final average compensation.\(^{16}\) These rules allow the plan to drop years or months in which the participant works fewer than a specified number of hours. These drop-out rules would only help a participant who returns to full-time work prior to retirement.

Because phased retirement should be structured to benefit both the employee and the employer, it seems unfair not to reflect pay increases in final average pay used to determine the benefit amount. To be sure the worker gets the benefit of pay rate increases during phased retirement, the plan could annualize pay during phased retirement years similar to the approach some plans use for any participant who does not work a full-time schedule. From the authors' experience, the most common approach is to annualize pay when the participant receives a

\(^{15}\)Plans that integrate benefits with Social Security (i.e., use permitted disparity described in IRC §401(l)) are required to use consecutive compensation amounts in determining final average pay. See §1.401(a)(4)-3(e)(2)(ii)(E).

\(^{16}\)§1.401(a)(4)-3(e)(2)(ii)(D).
partial year of service when not working a full-time schedule.\footnote{17} If the plan credits a partial year of service for a year in which a participant works fewer than a threshold number of hours, a participant working part time while phasing into retirement would receive a partial year of service.\footnote{18} In order to avoid double prorating, the plan would then annualize compensation for that year.\footnote{19} Other approaches are available to assure the phased retiree receives credit for pay increases while phasing into retirement in the calculation of final average compensation (Scahill and Forman, 2002).

5 Public Policy Issues In IRS Notice 2002-43

In 2002, the Internal Revenue Service and the Treasury Department requested “comments on issues relating to ‘phased retirement’ arrangements under qualified defined benefit plans.”\footnote{20} The Notice acknowledges that both employees and employers are interested in encouraging older, more experienced workers to remain in the workforce and phased retirement is one approach to offering a smoother transition from full-time work to full retirement.

The Notice raises a concern that allowing earlier access to retirement income could increase the possibility of the person outliving retirement savings. Phased retirement can provide additional time to save prior to full retirement. On the other hand, if the person needs to access retirement savings during phased retirement, phased retirement will begin the payout of those retirement savings sooner. If retirement savings are converted to a lifetime annuity, early distribution will not increase the risk of outliving retirement savings. Early distribution as a lifetime annuity increases the risk of inadequate retirement income because the distribution is reduced for early commencement.\footnote{21}

\footnote{17}{The authors have encountered plans sponsored by health care industry employers using this approach.}

\footnote{18}{Some plans credit a full year of benefit accrual service for a year in which the participant earns 2,000 or more hours and credit a fraction of a year equal to hours worked divided by 2,000 for a year in which the participant works at least 1,000 hours but fewer than 2,000 hours. Many other service crediting options are available.}

\footnote{19}{If a full-time employee works 2,000 hours and the phased retiree works 1,500 hours, pay for that year for the phased retiree would be annualized by multiplying pay received by 1.333 (2,000 / 1,500).}

\footnote{20}{Internal Revenue Service Notice 2002-43, 2002-27 IRB 38.}

\footnote{21}{As discussed above, benefits that begin prior to normal retirement are generally reduced to reflect the fact that the person will receive benefits over a longer period of time. The actuarial reduction is required to maintain actuarial neutrality and not in-}
The authors are pleased that the IRS and the Treasury Department are interested in finding ways to encourage employees and employers to find mutually beneficial phased retirement arrangements as well as finding ways to protect retirees from the risk of outliving retirement savings or having inadequate retirement income. The following are some of the specific issues raised in Notice 2002-43:

- The primary purpose of qualified retirement plans is to provide benefits after retirement. Under what circumstances would allowing defined benefit plan participants to begin receiving in-service retirement income distributions prior to normal retirement be consistent with this purpose?

- Should rules allowing in-service distributions consider the extent to which the participant has reduced his or her work schedule?

- If in-service distributions prior to normal retirement are allowed, how should additional benefits that accrue during continued employment be calculated?
  
  - How should reductions in compensation be addressed?
  - How should early retirement subsidies\(^{22}\) be taken into account?

- How should the definition of uniform benefits under nondiscrimination testing be changed?

- What guidance would be needed concerning qualified joint and survivor annuities and qualified preretirement survivor annuities requirements?

- What guidance should be provided concerning anti-backloading\(^{23}\) and maximum benefit\(^{24}\) limitations?

- How should phased retirement be distinguished from the situation in which an employee terminates retirement and is rehired as a consultant or independent contractor?

---

\(^{22}\) For the definition of early retirement subsidy, see discussion concerning actuarially neutral benefits in Section 3 above.

\(^{23}\) IRC §411.

\(^{24}\) IRC §415.
6 Is Phased Retirement Good Public Policy?

Workers currently have the option of easing into retirement without changing jobs, but there are pitfalls inherent in the current legal framework if an employer who sponsors a defined benefit plan offers phased retirement, including in-service pension benefits. Is it good public policy to change the law or issue regulations to support phased retirement? On the one hand, one could argue that providing workers with more opportunity to manage the end of their career is good public policy. Rather than forcing employees to change jobs in order to access their retirement benefits, employees would be able to continue their career job at a reduced schedule and receive a portion of their retirement benefits. The law would need to be changed or regulations would need to be issued to make this option a realistic one.

Some may be concerned that employers will force out older workers. Does phased retirement increase the risk that older workers who are not ready to reduce their work schedule will, instead, be forced out altogether? There is nothing inherent in phased retirement that increases the opportunity for age discrimination.

Whether to allow workers who are phasing into retirement access to a full distribution from the retirement plan is another public policy issue. Participants may need access to pension benefits in order to subsidize reduced earnings during phased retirement. As discussed earlier, benefits that commence prior to normal retirement are generally reduced to reflect the fact that they will be paid over a longer period. The periodic lifetime benefit is smaller if payments begin at a younger age. If the full accrued benefit is payable at the beginning of phased retirement, the participant faces a significant risk of inadequate income after full retirement. If the participant subsidizes the reduced pay\textsuperscript{25} during phased retirement by receiving the full accrued benefit rather than just a portion of the accrued benefit, the employee will have a significant reduction in pension income upon full retirement when earned income stops. On the other hand, if the participant only receives a portion of the accrued benefit during phased retirement, the participant will be able to receive a larger pension distribution upon full retirement when the untapped portion of the accrued benefit becomes payable in addition to the portion payable during phased retirement. This additional pension distribution will almost certainly be less than the income during phased retirement, but it lessens the income reduction upon full retirement.

\textsuperscript{25}Pay will be reduced during phased retirement because the employee is no longer working a full-time schedule.
In exchange for removing some of the current legal obstacles to a flexible phased retirement program, the government would likely require that phased retirement be available on a nondiscriminatory basis. Employers would be faced with the issue of whether a phased retirement program is retaining primarily highly skilled and effective workers or ones who are no longer effective. Employers who offer early retirement incentive programs face the same type of problem. The solution to this problem does not lie in the particulars of the retirement program, but instead it lies in effective workforce management. Just as employers are not required to employ anyone who wants to work for them, requiring employers to accept an individualized phased retirement program from any employee wanting to phase into retirement would interfere with business management.

The authors believe phased retirement is good public policy as long as the law is changed to facilitate phased retirement programs and protections are put in place to prevent abuse.

7 Basic Legal Considerations

Workers who elect phased retirement and who do not want to begin early distributions from the pension plan are free to take phased retirement under current pension law and regulations.

There are many legal considerations that impact a phased retirement program that seeks to allow participants access to in-service pension benefits prior to normal retirement age. We will discuss some of the major ones that affect defined benefit plans:

- Paying partial retirement benefits before full retirement,
- Offsetting continuing benefit accruals by the value of in-service distributions, and
- In-service distributions before the plan's normal retirement age.

Paying Partial Benefits Before Full Retirement: Although there is nothing specific in ERISA that prohibits defined benefit plans from paying partial benefits, there are a number of obstacles that may make these benefits impractical. For example, an employee taking phased retirement might want to receive 50% of his accrued benefit while working 50% of a full-time work schedule. ERISA and the Internal Revenue Code

\[\text{See Section 6 above for a discussion of problems caused by paying the full retirement benefit prior to full retirement.}\]
and related regulations refer to commencement of benefits, calculation of accrued benefits, spousal consent, etc. as they apply to the full pension. The statute and related regulations do not discuss paying a portion of the benefit beginning at one date and then paying the full benefit at a later date.

One question is how to increase the remaining portion of the accrued benefit for the period of phased retirement after normal retirement age.\textsuperscript{27} If the benefit were not actuarially increased, the participant would need to be given a suspension of benefits notice for the portion of the benefit for which payment is delayed. If the benefit were actuarially increased, the employer would need guidance on how the increase would be calculated. Would it apply to the full accrued benefit or to only the portion not in pay status? The actuarial increase must apply to the entire accrued benefit in order to achieve actuarial neutrality as defined in this paper. See Scahill and Forman (2002) for a detailed demonstration of one method for achieving actuarial neutrality.

**Offsetting Continued Accruals for Value of In-Service Distributions:** ERISA and the Internal Revenue Code prohibit the discontinuance of benefit accruals or a reduction in the rate of benefit accrual because of the attainment of any age.\textsuperscript{28} Proposed regulation §1.411(b)-2 provides details on the calculation of the accrued benefit after normal retirement age if in-service benefits are being paid out, but it only pertains to continued benefit accruals beyond normal retirement age.\textsuperscript{29}

The challenge for sponsors who want to design a balanced phased retirement program is how to offset for partial annuity distributions. If the entire additional benefit accrual were offset by the annuity value of the benefits paid, it is likely that no further benefits would accrue after partial distributions commence. The increase in the benefit ultimately paid out at full retirement over the benefit payable at the beginning of phased retirement might only be the elimination of the early retirement reduction. See Scahill and Forman (2002) for a detailed demonstration of various offset alternatives.

There are alternative ways to design the offset if the law and/or regulations accommodate these alternatives. If only 50% of the accrued benefit is being paid out prior to full retirement, the offset might apply only to half of the additional benefit accrual. As a result, the participant would continue to accrue at least 50% of what would have been accrued if no distributions had been received. This approach achieves actuarial neutrality.

\textsuperscript{27}Both the DOL and IRS have specific rules that apply to benefits that are not paid out while an employee continues working after normal retirement.

\textsuperscript{28}IRC §411(b)(1)(H) and ERISA §204(b)(1)(H).

\textsuperscript{29}See Example 3 of §1.411(b)-2 for a detailed discussion of these calculations.
neutrality. If the plan uses a full actuarial reduction before normal retirement and a full actuarial increase after normal retirement, the plan does not experience an actuarial gain or loss as a result of paying in-service benefits prior to full retirement.

**In-Service Distributions Before Normal Retirement Age:** Under current law, a defined benefit plan cannot make in-service distributions before the plan's normal retirement age. Many defined benefit plans use age 65 as the normal retirement age. Employees who want to begin phased retirement before the plan's normal retirement age are not able to use pension benefits to supplement earned income during phased retirement. Two-thirds of the companies participating in the Watson Wyatt phased retirement survey favor eliminating the restrictions on paying in-service distributions before normal retirement as a way to facilitate phased retirement (Watson Wyatt, 1999b, page 3).

Participants who want to maintain their prior standard of living during phased retirement will likely need personal savings, in addition to access to a portion of their retirement income, to supplement their pay during phased retirement. As discussed earlier, it does not seem to be good public policy to allow access to the full retirement benefit while the participant phases into retirement by working a reduced schedule. Pension benefits will not be sufficient to replace the reduction in compensation during phased retirement. Employees already face the need to have personal savings available during retirement to maintain their pre-retirement standard of living because Social Security and the employer's pension benefit rarely combine to replace 100% of the person's income just prior to retirement. The need for personal savings to use during phased retirement is no different.

8 **Impact of Phased Retirement**

8.1 **On Participant Protections**

One of the purposes of ERISA was to provide protection to participants. Some of the areas of protection could be impacted by phased retirement.

---

30Treas. Reg. §1.401-1(b)(1)(i) states "[a] retirement plan within the meaning of section 401(a) is a plan established and maintained by an employer primarily to provide systematically for the payment of ... benefits to his employees ... after retirement." In PLR 8137048, the IRS applied this regulation and concluded that an employee may not receive a distribution from a pension plan before normal retirement while still an active employee.
Disclosure: Effective communication about the plan lets participants understand and take advantage of the benefits offered—it is one of ERISA's participant protections. Plan sponsors will be challenged to provide understandable information about phased retirement because of the many choices available to the participant. Additional communication material may be needed to explain phased retirement options. The complexity of the communication materials depends on the flexibility of the phased retirement options available to participants. Because phased retirement is an individual arrangement, the communications will need to be tailored to the participant's particular situation. It will be important to disclose the impact, if any, of reduced pay and credited service on the ultimate retirement benefit. The participant also needs to understand the impact of in-service distributions on the ultimate annuity amount. Helping the participant assess the relative value of various options will help the participant make the best personal choice.

The Economic Growth and Tax Relief Reconciliation Act (EGTRRA) of 2001 enhanced the notice requirements for plans reducing the rate of future benefit accruals. Although these requirements will not apply to phased retirement, they provide useful guidance on the types of communication that could be helpful to employees considering phased retirement. Under the EGTRRA disclosure rules, the average participant should be able to understand the communication, and it must give enough information to explain the impact of the provision on the participant.

Software that allows participants to model their benefits under various phased retirement scenarios can be helpful for participants who are comfortable using these tools. In other situations, the sponsor could use a workbook or a series of benefit exhibits to help participants understand the effect of phased retirement on their retirement benefits.

Benefit Accrual Rules: The benefit accrual rules look at the rate of benefit accrual throughout the full employment period. Their basic purpose is to prevent backloading of benefits, and the demonstration

31 Examples of some of the choices are how much to reduce the full-time work schedule, whether (or when) to commence retirement plan distributions, and the payment method for those distributions.


33 I.R.C. §419b; ERISA §204.

34 Backloading refers to benefit accruals that increase steeply either as service increases or after a certain number of years of service. For example, a benefit formula providing 0.25% of average pay for each of the first 20 years of service and 2% of pay for each of the next 5 years of service would be considered a back-loaded formula. After 25 years of service, 5% of average pay would have been earned during the first 20 years.
of compliance of the benefit formula with the rules is typically based on a full-time employee. As a result, a plan that allows phased retirement should not have problems satisfying one of the accrual rules. Participants will continue earning benefit accrual service as long as they work the required number of hours, assuming the plan uses hours to credit service. 

Nondiscrimination Protection: The mechanical nondiscrimination rules can create problems for employers who try to accommodate employees who want to phase into retirement. The 2000 ERISA Advisory Council’s Working Group on Phased Retirement recommended the following nondiscrimination test alternatives to the Secretary of Labor:

- Permit a facts and circumstances test for phased retirement provisions in a pension plan, as an alternative to passing the mechanical nondiscrimination test.

- Develop safe harbors and/or special rules addressed to phased retirement programs that accommodate their special characteristics.

8.2 On Spousal Protections

The primary areas of spousal protection are the following ERISA requirements (Forman, 2000):

- Spousal consent for certain forms of benefit payment, and

- Amount of qualified surviving spouse annuity (QJSA) and qualified pre-retirement spousal death benefit (QPSA).

Spousal Consent: Spousal consent is only an effective protection if the spouse understands the impact of waiving the qualified joint and survivor annuity (QJSA). If the participant works a reduced schedule of employment and 10% of average pay would have been earned during the final five years of employment. This formula backloads the benefit accrual because it provides a much larger value for later years of service.

35Plans that use elapsed time for service credits will credit a full year of service for each full year during phased retirement. Plans requiring a certain number of hours for a year of service may credit less than a year of service during phased retirement, depending on the hours actually worked.


37See page 3 of the reference in footnote 4.

38A qualified joint and surviving spouse annuity (QJSA), as defined in §417(b), is an annuity that pays the surviving spouse no less than 50% and no more than 100% of the amount payable while the participant is living and receiving benefits.
during phased retirement, but he or she does not elect to receive any pension benefits before full retirement, spousal consent will not be affected by phased retirement.

If the participant elects to receive benefits during phased retirement, spousal consent would be required if the benefit were not payable in the form of a QJSA when phased retirement benefits begin. Upon full retirement, the original spousal consent would continue to apply to the additional benefit that will be payable unless the plan requires a new spousal consent. The requirement of multiple spousal consents may be confusing to the spouse, so the plan sponsor should try to ensure that the spouse understands that the initial consent only applies to the initial partial benefit.

**Amount of Qualified Joint Survivor Annuity (QJSA) and Qualified Preretirement Survivor Annuity (QPSA):** If participants elect phased retirement in a final average pay plan and the final average pay decreases during phased retirement, the ultimate retirement benefit may be lower than if the participant continued working full time. Therefore, the QJSA as well as the QPSA will be lower as a result of lower annual pay during phased retirement.

Although it is not reasonable to expect the spouse to have the right to consent to a reduced work schedule as part of phased retirement, there is an erosion of some of the spousal protections on account of phased retirement.

9 **Deferred Retirement Option Plans (DROPs)**

Some public sector retirement plans include deferred retirement option plans (DROPs) that allow workers to continue working and have retirement benefits deposited into a separate account that earns interest. The participant receives the value of the DROP account upon full retirement, generally no more than five years after electing to have benefits deposited into the DROP.

DROPs can be structured to apply once the participant has become eligible for unreduced benefits or to apply also to participants who are eligible for an early retirement subsidy. If the DROP is only available to participants who are eligible for unreduced benefits, the DROP effectively lets the participant take the unreduced benefit without having to retire. In this situation, DROPs would be attractive to participants

---

who do not need retirement income as a supplement during phased retirement.

If the DROP applies to participants eligible for subsidized early retirement benefits, it allows the participant to receive that subsidy without having to terminate employment. The subsidized benefit is deposited in the DROP and earns interest until retirement. At retirement, the subsidized early retirement benefit would be the monthly benefit payable to the participant. As long as the earnings on the DROP are sufficient to protect the value of the early retirement subsidy, the participant will end up with more valuable lifetime benefits because the participant will receive the value of the early retirement subsidy. Even though the benefits paid out after retirement are reduced as if the participant had retired early, the value of the DROP could more than compensate for the cost of the early retirement reduction in the lifetime benefit.

10 Conclusion

Phased retirement provides employees with important options for managing the end of their working careers. It provides employers with a way to retain valuable knowledge workers who no longer want to work full time. It is important for U.S. pension law and regulations to be modified to facilitate phased retirement, but those changes should include safeguards to protect workers and spouses as they make decisions that will have a lifetime financial impact.

References


Scahill and Forman: Phased Retirement


The Actuarial Value of Life Insurance Backdating

James M. Carson* and Krzysztof Ostaszewski†

Abstract

Backdating is a common (and legal) practice in the U.S. whereby a life insurance contract bears a policy date that is prior to the actual application date. This practice often results in the opportunity for some insureds to reduce the annual premium paid. Using cash flow projections and U.S. mortality, lapse, and interest rate data, we provide a model of the actuarial value of term life insurance backdating. Results indicate that the benefits to the applicant of backdating a term life insurance policy increase as the applicant age (and hence premium) increases. Increasing mortality, lapse, and interest rates, as well as increasing the length of the backdated period decreases the potential benefits of backdating. Finally, backdating appears to serve as a substitute for a finer partitioned pricing structure in the life insurance industry, as a risk-hedging mechanism for insurers, and as a risk-arbitrage tool for consumers.

Key words and phrases: insurance pricing, risk arbitrage, risk hedging, phantom surrender charge, incentive compatible contracting

*James M. Carson, Ph.D., C.P.C.U., C.L.U., holds the Midyette Eminent Scholar Chair in the College of Business at Florida State University. He serves as editor of the Journal of Insurance Issues and as president-elect of the American Risk and Insurance Association.

Dr. Carson’s address is: College of Business, Florida State University, Tallahassee, FL 32306-1110, U.S.A. E-mail: jcarson@cob.fsu.edu

†Krzysztof M. Ostaszewski, Ph.D, F.S.A., C.F.A., M.A.A.A., is the actuarial program director and professor of mathematics at Illinois State University. He is currently a Fulbright Senior Specialist and was a Fulbright Research Fellow in 1995.

Dr. Ostaszewski’s address is: Department of Mathematics, Illinois State University, Normal IL 61790-4520., U.S.A. E-mail: krzysio@ilstu.edu

The authors acknowledge the helpful comments of two anonymous referees, the editor, Randy Dumin, Kathleen McCullough, Ashwin Vasan, and participants at conferences held by S.O.A., A.R.I.A., W.R.I.A., and S.R.I.A. Dr. Ostaszewski acknowledges support of this research via a university research grant from Illinois State University.
1 Introduction

Life insurance backdating occurs when the insurance contract bears a policy date that is prior to the application date. From the applicant's perspective, the primary motivation for backdating is the reduction in premium that occurs because the premium is based on an age less than the applicant's life insurance age at the time of application. For example, suppose an insurer uses age nearest birthday. A person age 36 years and 9 months may be issued a policy that is backdated three months and one day in order to be charged the age 36 premium instead of the age 37 premium. Alternatively, for insurers that use age last birthday, a person age 36 years and two months may be issued a policy backdated two months and one day in order to be charged the age 35 premium instead of the age 36 premium. The obvious disadvantage of backdating is the necessity of paying premium for time that already has elapsed, i.e., from the backdated policy date to the actual application date.

To facilitate backdating, insurers often include, on the application for coverage, space for the agent/applicant to request that the policy be backdated. A survey (see Carson, 1994) yielded variations of the following comment from state insurance departments: "In researching the matter, it appears quite common in the industry for policies to be backdated." State laws in the U.S. typically allow backdating up to a maximum of six months. Thus, if birthdays and life insurance sales are assumed to be roughly evenly distributed throughout the year, only 50 percent of applications would be candidates for backdating. Therefore, if, for example, 40 percent of applications request backdating, this implies that up to 80 percent of the applications that are candidates for backdating actually request backdating. If, however, near future birthdays propel life insurance sales/purchases, then, for insurers using age nearest birthday, the 50 percent figure likely is a lower bound.

The question of whether to backdate essentially is a financial one: whether paying for lost time is offset by the right to pay lower premiums for the remaining life of the contract. Backdating appears to occur with significant regularity, as evidenced by discussions with U.S regulators, survey results, and examination of policy data from insurers. Surprisingly, however, little research exists on backdating, despite its potential for overcoming the effects of discrete (annual) life insurance pricing and serving as a potentially value-enhancing practice for the insured/policy owner.1

---

1 When measuring age of the insured person, two approaches are common (see, for example, Bowers et al., 1997): such age can be expressed as a real number, for example
An applicant who backdates and keeps the policy in force for a relatively short period of time will have less opportunity to reap the benefits from backdating. Thus, the decision to backdate may be seen as a signal to the insurer that the applicant plans to keep the policy for a relatively long period of time.2

Carson (1994) discusses life insurance backdating with respect to agents, insurers, and consumers and provides an analytical model for determining the value of backdating that accounts for interest. The goals of the present paper are to extend previous research by providing an actuarial model for the value of backdating that additionally incorporates assumptions on mortality and policy lapse rates. The following sections provide a conceptual framework and numerical examples of backdating, details of the model, data, results, and conclusions.

2 Model for Backdating

2.1 Conceptual Framework

Conceptually, backdating may be appropriate if the present value of premiums to be paid as a result of backdating is less than the present value of premiums to be paid based on the applicant's current age. An example of premium payments under a backdated versus a nonbackdated policy is in Table 1, which shows the annual premium payments for backdated (three months) and nonbackdated $250,000 annual renewable term (ART) insurance contracts issued to a 56 year old non-smoker (preferred risk) male who intends to hold the policy for six years. Premium data are for a large U.S. life insurer. In the present value calculations, the interest rate used is six percent, with annual

37.56 years (this is termed the continuous model) or as a whole number, for example 37 years (this is called the curtate or discrete model). While this terminology is not standard in economic literature, in this paper we refer to life insurance pricing based on the insured's age expressed as a whole number, as discrete pricing. This formulation can also be used for the age of the insured expressed in a unit of time shorter than a year, for example, a month. We term such shortening of the time unit used as a finer partition. Note also that discrete pricing exhibits elements of price discrimination of the form described in Nahata, Ostaszewski, and Sahoo (1990). To achieve a finer partitioned pricing structure, some single premium income annuity issuers interpolate rates monthly or daily, according to the actual age of the applicant.

2 The payment of an additional premium to reduce future premiums is similar to a residential mortgage borrower paying discount points (i.e., upfront interest) in order to obtain a lower interest rate (and thus lower monthly payments) on a mortgage loan. For more of the tradeoff between interest rates and discount points, see, for example, Stone and Zissu (1990), Yang (1992), and Brueckner (1994).
compounding. The next annual premium on the backdated contract is due in nine months, rather than in 12 months. Note that the six annual premiums on a backdated contract yield six years of coverage, less the number of backdated months. To achieve equal holding periods for the analysis, an additional number of months' coverage (three) for the backdated contract is purchased on a pro rata basis (no surcharge for partial year coverage, which yields a premium of 3/12 times $1,468 equals $367). This assumption is close to reality, because policyholders generally are able to switch the mode of payment (e.g., from annual to quarterly or monthly) after the issuance of the policy. The present value of premiums under each alternative equals $5,368 for the backdated contract and $5,633 for the nonbackdated contract, whereby each contract provides six full years of coverage.

Thus, this prospective insured would appear to benefit by $265 by purchasing coverage for time that already has elapsed, in order to gain the right to pay lower premiums over the next several years. Depending on several factors to be discussed, the benefit of backdating may be greater or less than that shown in this example; the benefit even may be negative (and thus a cost).

Continuing with the example above and taking the analysis from an annual to a monthly basis provides further understanding of the intricacies of backdating. That is, for the first nine months here, the insured enjoys a $62 premium savings ($825 versus $763). If the insured should die during this period, backdating will have been advantageous. At the end of the first nine months, however, the premium for the second year is due. If the insured dies during the next three months just after paying the second annual premium, backdating will not have been advantageous, as the cost of coverage would be $728 higher than without backdating ($825 versus $763 + $825/(1.06)^{9/12}$). For a contract that is backdated three months, this process continues for many years, as illustrated in Figure 1.

Figure 1 shows that insureds choosing to backdate must be aware of the true potential cost of backdating: the insured stands to gain from backdating during the first nine months of each policy year and stands to lose during the last three months of each policy year. This is due to the fact that future annual premiums for backdated policies will be

---

3 Although the example employs the ART plan for illustrative purposes, it should be noted that level term and universal life are more commonly sold in today's market, while a diminishing amount of ART life insurance is sold.

4 The results of the analysis will be biased in favor of backdating to the extent that this assumption is not valid. Alternatively, equal holding periods could be achieved by cancelling the nonbackdated contract prior to its expiration.
Carson and Ostaszewski: Life Insurance Backdating

Table 1
Backdated and Nonbackdated Scenarios

<table>
<thead>
<tr>
<th>Premium Due Date</th>
<th>Premium Amount</th>
<th>Nonbackdated Contract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Today</td>
<td>$ 763</td>
<td>Today</td>
</tr>
<tr>
<td>+ 9 months</td>
<td>$ 825</td>
<td>+ 12 months</td>
</tr>
<tr>
<td>+ 21 months</td>
<td>$ 903</td>
<td>+ 24 months</td>
</tr>
<tr>
<td>+ 33 months</td>
<td>$1,003</td>
<td>+ 36 months</td>
</tr>
<tr>
<td>+ 45 months</td>
<td>$1,130</td>
<td>+ 48 months</td>
</tr>
<tr>
<td>+ 57 months</td>
<td>$1,285</td>
<td>+ 60 months</td>
</tr>
<tr>
<td>+ 69 months</td>
<td>$ 367</td>
<td></td>
</tr>
</tbody>
</table>

Total Premiums Paid: $6,614  $6,276
Present Value at 6% $5,368  $5,633
Net Present Value = $5,633 - $5,368 = $265

Notes: The contract in this example is a $250,000 annual renewable term issued to a male age 56 classified as a preferred risk, nonsmoker. The policy is backdated three months. Premium data are from Best's Policy Reports, June 2000, for a large U.S. life insurer. The interest rate used for discounting in this example is six percent. The + 69 months premium of $367 is calculated as (3/12) times $1,468. This premium for three months of coverage is necessary to achieve equal holding periods (six years) for the comparative analysis.

due earlier (e.g., three months earlier) than for nonbackdated policies, creating what could be called a backdating phantom surrender charge. Further, this surrender charge can be more costly to the policy owner than the standard or regular surrender charge because the phantom surrender charge can apply even in the event of death, as is illustrated by the monthly NPV line in Figure 1. It is not for many years that the benefits of backdating are at all times positive.

2.2 Key Equations and Data

Equation (1) can be used to analyze the value of backdating. It considers annual premiums, the number of months by which the contract is backdated \( m \), and assumptions regarding the interest rate \( r \), the insured's backdated age \( x - 1 \), and the holding period (number of
years, \( y \)). The equation gives the present value of premiums paid on a backdated contract for a given holding period, \( \text{PVB}_{x-1,y} \), i.e.,

\[
\text{PVB}_{x-1,y} = P_{x-1} + \sum_{k=1}^{y-1} P_{x-1+k} \cdot \frac{m}{12} + \frac{m}{12} \cdot \frac{1}{1+(1+r)} \cdot P_{x-1+y} \cdot \frac{y-m}{12} \quad (1)
\]

where \( P_z \) is the premium due at age \( z \), and \( v = 1/(1+r) \) is the discount factor. The last term in equation (1) adjusts for the additional number of months' coverage that should be purchased in order to provide equal periods of coverage between the backdated and nonbackdated contracts. Equation (1) is expressed in number of months by which the contract is backdated, rather than number of days, although either would be acceptable (with appropriate adjustments to the equation).\(^5\) Note that premiums on the backdated contract begin at \( x - 1 \) and are consistent with those of the nonbackdated contract.

\(^5\)Tax considerations generally are not relevant to the analysis, as individual purchases of coverage are made with after-tax dollars. In a business setting, tax implications may require additional analysis.
Equation (2) computes the present value of annual premiums paid on a nonbackdated contract for a given holding period \( y \), i.e., \( \text{PVNB}_{x,y} \), which is given by

\[
\text{PVNB}_{x,y} = \sum_{k=0}^{y-1} P_{x+k}v^k. \tag{2}
\]

The net present value, NPV, is the difference between the first two equations and is given by equation (3). Observe that equation (3) gives the annual NPV of backdating, which is simply the weighted average of the monthly NPVs for any given year.

\[
\text{NPV}_{x,y} = \text{PVNB}_{x,y} - \text{PVB}_{x-1,y}. \tag{3}
\]

Equation (4) gives the actuarial net present value (ANPV) of backdating. This equation accounts for mortality, lapse, and interest rates. By accounting for mortality and lapse, equation (4) may be viewed as an analysis from a public policy perspective, as it is less common to think in terms of discounting for mortality and lapse for an individual. The term \( \left( k-1 \rho_x^{(\tau)} q_x^{(\tau)} \right) \) represents the probability that the policy owner will die or lapse during the year. The last term in equation (4) expresses the fact that those dying in the last policy year enjoy the same benefits of backdating (lower premiums) as those who survive to policy termination.

\[
\text{ANPV}_{x,y} = \sum_{k=1}^{y-1} \text{NPV}_{x,k} \times \left( k-1 \rho_x^{(\tau)} q_x^{(\tau)} \right) + \text{NPV}_{x,y} \times \left( y-1 \rho_x^{(\tau)} \right). \tag{4}
\]

Equations (1) through (4) are applied to life insurance premium data. Data are from A.M. Best (2000) for preferred risk, nonsmoking males aged (backdated/nonbackdated) 35/36 and 45/46, and $250,000 of annual renewable term insurance. The premium data used in the analysis are shown in Table 2. The NPVs of backdating for one month, three months, and six months are analyzed with respect to holding periods up to 30 years.

---

6Equation (4) could be adapted for monthly decrements and premium payments, as opposed to decrements that occur at the end of the year and premium payments at the beginning of each year (with an adjustment in the last year for the backdated policy), although the results would be minimally affected by such a change.
Table 2
Annual Premium Data
For $250,000 Annual Renewable Term Insurance

<table>
<thead>
<tr>
<th>Age</th>
<th>Premium</th>
<th>Age</th>
<th>Premium</th>
<th>Age</th>
<th>Premium</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>$258</td>
<td>49</td>
<td>$523</td>
<td>63</td>
<td>$1,915</td>
</tr>
<tr>
<td>36</td>
<td>$270</td>
<td>50</td>
<td>$550</td>
<td>64</td>
<td>$2,180</td>
</tr>
<tr>
<td>37</td>
<td>$283</td>
<td>51</td>
<td>$583</td>
<td>65</td>
<td>$2,473</td>
</tr>
<tr>
<td>38</td>
<td>$295</td>
<td>52</td>
<td>$620</td>
<td>66</td>
<td>$2,793</td>
</tr>
<tr>
<td>39</td>
<td>$308</td>
<td>53</td>
<td>$663</td>
<td>67</td>
<td>$3,140</td>
</tr>
<tr>
<td>40</td>
<td>$325</td>
<td>54</td>
<td>$710</td>
<td>68</td>
<td>$3,515</td>
</tr>
<tr>
<td>41</td>
<td>$343</td>
<td>55</td>
<td>$763</td>
<td>69</td>
<td>$3,918</td>
</tr>
<tr>
<td>42</td>
<td>$368</td>
<td>56</td>
<td>$825</td>
<td>70</td>
<td>$4,348</td>
</tr>
<tr>
<td>43</td>
<td>$388</td>
<td>57</td>
<td>$903</td>
<td>71</td>
<td>$4,805</td>
</tr>
<tr>
<td>44</td>
<td>$408</td>
<td>58</td>
<td>$1,003</td>
<td>72</td>
<td>$5,290</td>
</tr>
<tr>
<td>45</td>
<td>$428</td>
<td>59</td>
<td>$1,130</td>
<td>73</td>
<td>$5,983</td>
</tr>
<tr>
<td>46</td>
<td>$448</td>
<td>60</td>
<td>$1,285</td>
<td>74</td>
<td>$6,875</td>
</tr>
<tr>
<td>47</td>
<td>$470</td>
<td>61</td>
<td>$1,468</td>
<td>75</td>
<td>$7,900</td>
</tr>
<tr>
<td>48</td>
<td>$495</td>
<td>62</td>
<td>$1,678</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Premium data are from Best's Policy Reports, June 2000, for preferred risk nonsmoking males for a large U.S. life insurer. Premium for age 75, however, is extrapolated from the previous years' premiums.

3 Main Results

Applying equation (3) to the premium data in Table 2, the annual NPVs (for holding periods up to 30 years) of backdating a contract versus not backdating a contract are shown in Figures 2 through 5 below. For a 36 year old applicant, Figure 2 illustrates that the annual NPVs range from -$129 to $569, for holding periods up to 30 years. As shown in Figure 3, however, the actual benefit or cost of backdating depends upon the particular month in which the policy ends. For the 36 year old applicant that backdates by six months, Figure 3 shows that for the first six months of each year, the monthly NPV of backdating is positive. During the latter six months of each policy year, however, monthly NPVs of backdating are negative. For the 36 year old applicant who backdates a term policy by six months, the annual NPV line...
Figure 2: Annual NPV of Backdating: Age 36 Assumed Interest Rate: Six Percent Backdated by One, Three, and Six Months

shows that it takes almost 20 years before the weighted average of the monthly benefits and costs is positive.⁷

For the 46 year old applicant, Figure 4 illustrates that the annual NPVs range from -$200 to $1,933, for holding periods up to 30 years. As before, the actual benefit or cost of backdating depends upon the particular month in which the policy ends. For the 46 year old applicant that backdates one month, Figure 5 shows that for the first 11 months of each year, the monthly NPV of backdating is positive. During the last one month of each policy year, however, monthly NPVs of backdating are negative. For the 46 year old applicant who backdates a term policy by one month, the annual NPV line shows that it takes approximately two years before the weighted average of the benefits and costs is positive.⁸

⁷Note that the 6-month line of Figure 2 is the same as the annual NPV points in Figure 3.

⁸Note that the 1-month line of Figure 4 is the same as the annual NPV points in Figure 5.
Figures 2 and 4 clearly illustrate that the costs and benefits of backdating increase with age (premium). This finding is intuitively appealing, as the annual difference in mortality costs is greater at higher ages, and backdating to save age would be expected to have a larger impact on cost. These figures also illustrate that the benefit of backdating decreases as the number of months that a contract is backdated increases. For both ages examined, backdating an annual renewable term insurance contract by six months results in predominantly negative NPVs for holding periods of at least 13 years. The equations also are applied to 20-year level term insurance premium data. The resulting graphs are different than those shown here, but overall results are similar. The costs of backdating are somewhat larger (and the benefits somewhat smaller) based on the level term insurance premium data than the costs/benefits based on annual renewable term insurance premium data.

It is clear that the benefits of backdating are somewhat rear-end loaded. Even though a policy owner may intend to hold the policy for several decades, the potential for death or lapse often will make the
holding period shorter than planned. This nature of backdating begs the question of the likely holding period for a policy owner. To answer this question, we use mortality and lapse data. For this analysis, we obtained mortality data from the Life Table for the Total Population: United States, 1979-1981 (see Bowers et al., 1997). Because lapse rates vary across insurers (and across product lines), for simplicity we use lapse rates described in the Life Insurance Fact Book (1997) for ordinary life. Thus, equation (4) provides the actuarial present value of backdating by accounting for mortality and policy lapse. Applying equation (4) to the data, the actuarial present value of backdating for a 46-year-old male equals $72, $314, and $471, for backdating periods of six months, three months, and one month, respectively.

Note that the expected policy holding period (based on the mortality and lapse data described above) for the 46 year old male in the analysis
is 13 years. Thus, based on an expected policy holding period criterion, Figure 4 would suggest that the net present value of backdating (at the year-13 point) equals -$18, $180, or $310, for backdating periods of six months, three months, and one month, respectively. The actuarial value amounts are somewhat higher than the values indicated by the simple expected policy holding period criterion. The higher actuarial values stem from the nature of exponential growth of the benefits of backdating, and the relatively high dollar values that are factored into the actuarial present value calculation, but not into the expected value calculation.

4 Discussion

Our analysis indicates that discrete (annual) mortality pricing of life insurance results in the opportunity for some insureds to reduce the cost of term life insurance via backdating. Backdating typically is driven by the agent as opposed to the policy owner, and backdating likely
is the industry's response in lieu of a finer pricing structure. From a transaction cost perspective, allowing backdating may be less costly for an insurer than attempting to price more points on the age/price continuum.

While the benefits of backdating can be positive, the preceding analysis indicates that the benefits of backdating a term insurance contract generally are least likely to be positive in situations involving relatively long backdated periods and relatively short holding periods. The phantom surrender charge created by backdating (with the premium paid for time already elapsed) serves to align the interests of the insurer, agent, and policy owner in terms of policy persistency. In effect, backdating may serve as a bonding mechanism and as a signal that the policy owner intends to hold the policy longer than the typical policy owner. In this sense, backdating leads to superior incentive-compatible contracting between the various parties. Additionally, life insurance companies face significant risks due to surplus strain in early durations of life policies, and backdating transfers a (relatively small) part of that risk to consumers. From the perspective of the life insurance firm issuing the contract, backdating appears to be an indirect risk-hedging mechanism.

Backdating is a zero-sum game with respect to the insurer and the policy owner. Prior to policy termination, the winner from backdating is unknown and is not determined until the time of lapse/surrender or death. Figures 2 and 4 illustrate that the likelihood of benefiting from backdating (from a given policy owner's perspective) is maximized, ceteris paribus, with the shortest possible backdated period. Thus, backdating is, in a sense, risk arbitrage from the consumer's viewpoint: risk arbitrage, not in the sense that the contract must be held for some minimum amount of time to break even (as in Carson, 1994), but risk arbitrage in the sense that benefits could change quickly to costs (as shown most clearly in Figure 1) depending on the specific month of death or lapse.

This study's results suggest that regulatory concerns over potential problems related to backdating are valid because backdating will not be beneficial to all who backdate—i.e., those insureds that lapse or die soon after paying a renewal premium generally will be worse off from backdating. The results also indicate that prohibition of backdating is overly restrictive and would preclude beneficial transactions for many applicants. Because backdating may be beneficial or detrimental to the policy owner, insurers are wise to explain the potential costs and benefits of backdating to prospective insureds. Other legal or ethical issues
that might arise include whether it is an unfair trade practice to permit backdating for one applicant and not another similar applicant.

The actuarial present value of backdating suggests that backdating often is a value-enhancing practice. Higher mortality and lapse rates than those assumed here obviously would reduce the values associated with backdating. In addition, the choice of a particular interest rate has an important effect on the results of the analysis. Especially for holding periods greater than ten years, increasing (decreasing) the interest rate assumption results in lower (higher) NPVs of backdating. The magnitude of the effect of the interest rate assumption increases with the age (premium) of the applicant. Finally, the gains from backdating relate to the increase in premiums from year to year. Thus, to the extent that annual premium increases are similar between smoker/non-smoker or male/female insureds, no significant differences between smoker/nonsmoker or male/female insureds would be expected. Because annual premium increases become more pronounced at later ages, however, the potential benefits of backdating increase with age, especially beyond age 45.

5 Closing Comments

Life insurance backdating is similar to paying discount points to obtain a lower interest rate on a mortgage. Our analysis indicates that life insurance contract prices based on annual age differences result in the opportunity for some applicants to reduce their cost of coverage. In a sense, backdating is a market response to a pricing practice that does not distinguish between age differences less than one year. Backdating appears to serve as a substitute for a finer partitioned pricing structure in the life insurance industry, as a risk-hedging mechanism for insurers, and as a risk-arbitrage tool for consumers. While applicants realize the benefits of backdating immediately upon policy inception, these benefits quickly turn into costs for a number of months upon payment of each successive annual premium, and this cycle continues for many years. Thus, backdating is not a perfect substitute for a pricing structure with finer partitioning.

Findings indicate that the potential benefit of backdating tends to increase as the number of months by which the contract is backdated is decreased. Specifically, the annual NPVs of backdating a contract six months were predominantly negative for both ages examined (36 and 46) for holding periods of up to at least 13 years. For contracts backdated only one month and for later ages, however, the potential to
reduce the cost of coverage is substantial, even for relatively short holding periods. Increasing the assumed interest rate assumption (as well as mortality and lapse assumptions) decreases the costs and benefits of backdating. As discussed earlier, the potential benefits of backdating tend to increase with age of the applicant.

The equations presented here can be used to determine the financial and actuarial value of backdating a term life insurance contract. Future research on this topic might focus on the extent to which backdating for other types of life insurance contracts (e.g., cash value life insurance) differs from this analysis for term insurance.

References


Decision Tree Analysis of Terminated Life Insurance Policies

Robert Keng Heong Lian,* Yuan Wu,† and Hian Chye Koh‡

Abstract§

Statistical methods such as regression and survival analysis have traditionally been used to investigate the factors affecting the duration of terminated life insurance policies. This study explores a different approach: it uses a more recently developed data mining technique called decision trees. By sequentially partitioning the data to maximize differences in the dependent variable (duration in this study), the decision trees technique is good at identifying data segments with significant differences in the dependent variable. This identification can be useful when a company is trying to understand the factors driving or associated with the termination of life insurance policies. Decision trees also have an advantage over other techniques such as linear regression in their ability to detect nonlinear and other complex relationships that are more likely to exist in any practical data set.

Key words and phrases: data mining, life insurance policies, termination, persistence, lapses

*Robert Keng Heong Lian, M.Sc., A.S.A., is an associate professor emeritus at Nanyang Technological University. He was the general manager of British American Life Assurance in Singapore before joining the university. His research interests include health care financing and retirement planning.

Prof. Lian's address is: Nanyang Business School, Nanyang Technological University, Singapore, SINGAPORE 639798. E-mail: akhlian@ntu.edu.sg

†Yuan Wu, Ph.D., is an associate professor at Nanyang Technological University. His research interests include traffic accident analysis and application of statistics to business and finance.

Dr. Wu's address is: Nanyang Business School, Nanyang Technological University, Singapore, SINGAPORE 639798. E-mail: aywu@ntu.edu.sg

‡Hian Chye Koh is an associate professor and vice-dean of the Nanyang Business School. His research interests include data mining applications, statistical analysis and modeling in accountancy and business, and behavioral research.

Prof. Koh's address is: Nanyang Business School, Nanyang Technological University, Singapore, SINGAPORE 639798. E-mail: ahckoh@ntu.edu.sg

§The authors wish to thank the anonymous referee for helpful comments and suggestions that have improved this paper.
1 Introduction

From a management point of view, in order to contribute to the profitability of a business, quality sales are required. For life insurance, quality sales means policies sold that remain in force long enough to at least allow the life insurance companies to recoup their high acquisition costs. If a policy lapses (i.e., is terminated because of non-payment of premiums) too early, the cost can be high to all parties. For the policyholders, the premiums paid will be forfeited. For the company, on the other hand, the cost of acquisition will not be fully recovered. As a consequence, policyholders who persist (i.e., who keep their policies in force) must bear the unrecovered expenses, thus subsidizing those who terminate their policies. For agents, there will be loss of renewal commissions and, in extreme cases, maybe even a loss of their jobs. Therefore, it will be advantageous for life insurance companies to know the factors affecting the persistency of policies and work toward increasing persistency.

Motivated by the importance of persistency, the Life Insurance Association of Singapore (LIAS) has initiated a study on the termination of policies. In order to carry out the study, the LIAS has collected data on policies terminated because of non-payment of premiums. The data set includes individual factors such as policyholder age at purchase, gender, duration of policy at termination, type of policy, etc.

Several statistical methods such as regression analysis (Renshaw and Haberman, 1986, and Lian et al., 1993) and survival analysis (Lian, Wu, and Loi, 1998) have been used to analyze terminated life insurance policies. In the last decade, however, three new interrelated areas of data analysis have emerged that have emphasized obtaining more information from a data set: data warehousing, knowledge management, and data mining. Of particular interest to us is data mining, which aims to identify valid, novel, potentially useful, and understandable correlations and patterns in data (Chung and Gray, 1999). Data mining techniques such as neural networks and decision trees provide a different approach to predictive modeling. This study uses decision trees (also known as recursive partitioning algorithm) to analyze the LIAS data set.

2 Termination and Persistency of Policies

Although persistency is an important issue in insurance, few academic research studies have been done in this area. Instead, this issue
has been discussed in trade journals where authors have emphasized the agency force and described how it can help to reduce terminations.

Willet (1986), for example, stressed the importance of good personal communication with clients and prospects in creating goodwill, faith, trust, and high persistency as compared to a telephone recorder answering system. Ingram (1987), in examining the relationship between effort and compensation, found that agents chose activities that maximize return of their time and effort. He suggested that new concepts of life insurance sales be developed to focus on the winning of new clients and the building of value when clients repeatedly pay premiums and purchase additional policies. Santos (1992) believed that life insurance agents could prevent higher termination rates by maintaining an attitude of service, being thorough in determining a client's needs, and by presenting a clear, systematic explanation of the proposed solution in the spirit of trust and friendship. Santos added that agents should also listen, interact, and promise to stay in touch to monitor their clients' needs and to adjust their insurance programs.

Hall (1990) and Berlin and Powell (1990) expressed their concern on the importance of market research and how it could help to improve persistency. Hall (1990) urged life insurance researchers to stay attuned to factors such as unemployment caused by mergers, downsizings, and closures, which contribute to sharp declines in income and can affect terminations. Berlin and Powell (1990) emphasized the importance of the evaluation of persistency to monitor the overall quality of the industry's marketing and service efforts. The direction and degree of persistency rates reflect how well a company and its agents excel in matching products and services with the insurance needs of its policyholders.

Goodwin (1993) objected to insurers coercing agents (against their code of ethics) to encourage clients to retain insurance policies that are against their clients' best interests. He suggested that if a penalty is to be imposed on an agent for a policy termination, there should be a scale of commission forfeitures keyed to the age of the terminated policy. Sweeney (1996) opined that adverse publicity over improper sales practice had worked against persistency and had affected the sales of new products. Zinseer (1996) suggested that improvement in the persistency of existing business could have a 35% impact on pre-tax earnings. He attributed the main reason for the delay in insurers exploiting their existing customer base to the lack of skills and the scale of field officers who were responsible for managing customers.

Lautzenheiser and Barks (1991) and Howard (1997) focused on how product design could encourage policyholders to keep their policies longer. Lautzenheiser and Barks (1991) advocated that the design of
insurance products should provide features to cover life-cycle benefits and incentives to encourage policyholders to keep their policies to ensure that such benefits would be there when needed. Howard (1997) suggested that persistency could be improved by reinforcement visits, regular service and information, and termination could be minimized by product flexibility to meet changing policyholder insurance needs.

Purushotham (2001) presented the results of a LIMRA survey of company practices regarding the measurement of individual life persistency experience. Most of the responding companies have more than one tool in place for monitoring product persistency. Further, more than half of these companies conduct traditional actuarial studies on a regular basis. These traditional actuarial persistency studies involve the determination of lapse rates by policy count, face amount, and annualized premium level. Often lapse results are examined by product, policy year, issue age, premium payment frequency, and billing method.

In recent years, insurance companies have begun to recognize the impact of product persistency experience on company profitability. For example, Streiff (2001) expressed concern over shrinking annuity persistency and its adverse impact on profitability. He concluded that insurers need to recognize the problem and commit resources to solving it. Just as there is a team dedicated to new business (sales), there should also be a team dedicated to conservation. In a later paper, Streiff (2002) suggested that to increase persistency, insurers should carefully analyze past persistency patterns of their distributors in order to stop persistency problems before the product is sold. Further, in addition to meaningful contact and good service, it is important to make sure the product is performing as the policy owner expects.

While the papers cited above were not technical, there are a few papers that have used statistical techniques to analyze empirical data pertaining to the termination of policies. Renshaw and Haberman (1986) used lapse (or withdrawal) data with various policy characteristics to determine the reasons policyholders gave up their policies in the U.K. They applied a general linear model (GLM) to identify factors (such as type of policy, location of the office, and mode of payment) that could affect termination rates. They found that the main interaction involved policy duration and the type of policy.

Lian et al. (1993) applied multiple regression analysis and concluded that the variables age, gender, mode of payment, type of policy, and method of payment had a significant bearing on the duration of life insurance policies. In a later study, Lian, Wu, and Loi (1998) employed survival analysis to study the factors affecting the termination of policies and to evaluate the impact of each significant factor. The study
provided important information such as the average life expectancy of terminated policies, high risk-of-termination time periods, and incidence of survival rates by gender, types of policy, modes of payment, method of payment, and service status.

To summarize, few academics have analyzed policy terminations, while several trade and professional publications have discussed persistency and the minimization of terminations. Many of the findings and suggestions contained in these papers are based on anecdotal evidence instead of empirical analysis. This paper attempts to fill this gap in the literature by analyzing a large database of terminated policies, deriving objective (statistical) evidence of the factors associated with lapses, and generating useful rules and visual results to help insurance companies and their agents to better understand and minimize policy terminations.

3 Data Mining

3.1 The Literature

More companies are now using data mining as the foundation for strategies that help them outsmart competitors, identify new customers, and lower costs (Davis, 1999). Data mining is now widely used in marketing, risk management, and fraud control (Kuykendall, 1999). For example, an insurer may data-mine customer information to develop competitive rates, a retailer may analyze data mined from transactional systems to refine direct mailings and advertising campaigns to improve sales, while an auditing system may use data mining to help identify what could be fraudulent insurance claims. In a typical lapse-related application, data mining can identify customers who are likely to be profitable and who are likely to leave or churn, thus helping the insurer to target the valuable customers for extra value-added customer services, special offers, and loyalty incentives (Peacock, 1998).

To date, data mining has been applied to a wide range of business applications including direct mailings and advertising (Storey, 2001), customer relationship management (Lesser, Mundel, and Wiecha, 2000), fraud detection and fraud control (Kuykendall, 1999), surveillance (Holiday, 2001), risk management, (Berger, 1999), customer acquisition (Peacock, 1998), lifetime valuation (Drew et al., 2001), and others. Although there is widespread application of data mining in the commercial world, this interest is not matched in the academic world.
3.2 Data Mining Techniques and Decision Trees

Data mining can be defined as the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns of data (SAS Institute, 1998). The SAS data mining methodology comprises the following five stages: Sample, Explore, Modify, Model and Assess (also known as SEMMA):¹

- Sampling is desirable if the data for analysis are too voluminous for reasonable processing time or if it is desirable to avoid problems of generalization by dividing the data into different sets for model construction and model validation.

- Exploration and modification refer, respectively, to the review of data to enhance understanding (e.g., by examining the summary measures) and the transformation of data (e.g., to induce a linear relationship or a normal distribution). It is noted that not every data-mining project needs sampling or modification of the data. Exploration is usually useful, however, and done as a form of preliminary analysis.

- The modeling stage is the actual data analysis. Most data mining software include traditional statistical methods (e.g., regression analysis and discriminant analysis) as well as non-traditional statistical analyses such as neural networks and decision trees.

- Finally, the assessment stage allows the comparison of models and results from any data mining model by using a common yardstick (e.g., lift charts or profit charts).

This study uses SAS Enterprise Miner software for data analysis.²

Decision trees, one of the most commonly used data mining techniques, are often used to discover classification rules for a chosen attribute (i.e., target or dependent variable) by systematically subdividing

¹Another data mining methodology is provided by CRISP-DM (Cross-Industry Standard Process for Data Mining—see <http://www.crisp-dm.org>). This framework proposes the following stages: (1) business understanding, (2) data understanding and data preparation, (3) modeling, (4) evaluation, and (5) deployment. Business understanding is critical, as it identifies the business objectives and hence the success criteria of data mining projects. Deployment refers to the operationalization and implementation of a data mining model in the organization. The data, modeling, and evaluation stages in CRISP-DM are similar to SEMMA.

information contained in the data set. Consequently, a decision tree-based approach gives clear decision rules for target variables such as who is likely to purchase a certain product, how much profit or loss one can expect by targeting a particular subgroup, which customer is likely to churn/turnover ... etc. The greatest benefit of decision trees is their ability to generate clear and understandable "if... then..." rules that can be readily applied in a business process. Related to this is the ability to visualize decision tree results. As with all decision-making methods, however, decision trees should be used in conjunction with common sense—decision trees are just one important part of a decision-making tool kit.

4 Model Development

4.1 The Singapore Life Insurance Industry

Singapore is a premier insurance center in Asia. There are currently 13 insurance companies competing in the life insurance sector (as listed in the ASEAN Insurance Industry Directory\(^3\)). These include local insurance companies (e.g., NTUC Income and Asia Life), insurance companies associated with banks (e.g., UOB Life and Great Eastern Life) and foreign insurance companies (e.g., AIA and China Life).

The insurance industry in Singapore is regulated and supervised by the Monetary Authority of Singapore (MAS) (the central bank of Singapore) via a risk-based framework. In addition, parliamentary acts and MAS notices impact the life insurance industry. For example, the Financial Advisers Act implemented in October 2002 sets the minimum entry and examination requirements for financial advisers, while MAS Notice 318 sets the principles of appropriate market conduct for direct life insurers. Overall, MAS ensures that insurance companies in Singapore maintain a high standard of professionalism, financial management, and prudence.

Another unique feature in Singapore is the Central Provident Fund, which is a compulsory social security savings scheme. Employees can use their CPF funds to buy insurance policies such as term policies and health policies.

Based on data provided by MAS, the Singapore dollar amounts for new annual premiums for life insurance, new single premiums for life insurance, and new single premiums for annuity in 2002 are S$686.7

\(^3\)Available at: <http://www.insurance.com.my/zone_industry/directory>.
million, S$5.9 billion and S$602.6 million, respectively. Investment-linked policies accounted for 37% of total new single premiums. A total of 1.42 million policies were sold in 2002 with a life insurance coverage of S$27.9 billion. Of these policies, approximately 14% were whole life policies, 22% were endowment policies, 62% were health policies, and 2.0% were term policies.

Total individual annual premiums in force amounted to S$5.2 billion in 2002. There were 5.89 million individual policies in force with a life insurance coverage of S$240 billion. About 90% of the Singapore population is estimated to be covered by life insurance. The average annual growth in life insurance in the years 1992 to 2002 is 12.4%.

Life insurance in Singapore is sold primarily through the traditional agency distribution channel. The importance of agents is evidenced by the formation of the Committee on Efficient Distribution of Life Insurance in March 2000, which has a mandate to make recommendations to enhance the sales advisory process, product disclosure, and professional training requirements. In the past several years, insurance companies have also moved into direct marketing through strategic alliance with banks (bancassurance). This channel is now becoming increasingly important in view of the emergence of financial advisers in banks. Some insurance companies also sell their policies through credit card companies and through the Internet. The latter is likely to become an effective and economical means of distribution given the high information technology penetration in Singapore.

In the event of surrender of life policies, the surrender value is calculated using the U.K. model, which that looks at the market conditions at the point of surrender. Also, orphan policies (i.e., policies without agents) are quickly assigned to other agents by group managers. In recent years, the declining mortgage rates and re-financing rates appear to have affected the persistency of endowment policies.

Growth in the life insurance industry has been sluggish in 2003 because of economic uncertainties, the war in Iraq, and the effects of the SARS outbreak. These have contributed to the reluctance of the Singapore population to make long-term financial commitments. The outlook for 2004 is more positive, driven mainly by a more optimistic economic outlook.

4.2 The Data

The data for the study were supplied by the Life Insurance Association of Singapore, based on all the individual policies that are terminated because of non-payment of premiums. A total of 48,243 such
policies is included in the study, with an average duration of 2.58 years (standard deviation = 3.33 years). These include policies that have been terminated because of lapses (32,360 or 67.08%), surrender (11,705 or 24.26%), or conversion to reduced paid-up or extended term insurance (4,178 or 8.66%), but exclude policies terminated because of death, maturity, expiry, or conversion to permanent plans. For each policy, the following variables are provided in the LIAS data set: duration of policy, age, gender, type of policy, mode of payment, method of payment, status of service, and size of policy. The details are shown in Table 1.

Table 1
Variables Used to Study Duration of Terminated Policies

<table>
<thead>
<tr>
<th>Variable</th>
<th>Measurement</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>32,902 (68.2%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>15,341 (31.8%)</td>
</tr>
<tr>
<td></td>
<td>Type of Policy (TP)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole Life Par. (WLP)</td>
<td>16,211 (33.6%)</td>
</tr>
<tr>
<td></td>
<td>Whole Life Non-Par. (WLN)</td>
<td>7,386 (15.3%)</td>
</tr>
<tr>
<td></td>
<td>Endowment Par. (EP)</td>
<td>19,916 (41.3%)</td>
</tr>
<tr>
<td></td>
<td>Endowment Non-Par. (EN)</td>
<td>183 (0.4%)</td>
</tr>
<tr>
<td></td>
<td>Term (T)</td>
<td>3,916 (8.1%)</td>
</tr>
<tr>
<td></td>
<td>Others (TO)</td>
<td>631 (1.3%)</td>
</tr>
<tr>
<td></td>
<td>Mode of Payment (MP)*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Annually</td>
<td>11,237 (23.3%)</td>
</tr>
<tr>
<td></td>
<td>Semi-annually</td>
<td>3,187 (6.6%)</td>
</tr>
<tr>
<td></td>
<td>Quarterly</td>
<td>3,150 (6.5%)</td>
</tr>
<tr>
<td></td>
<td>Monthly</td>
<td>30,619 (63.6%)</td>
</tr>
<tr>
<td></td>
<td>Method of Payment (MPT)**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cash/Check (CC)</td>
<td>25,370 (53.3%)</td>
</tr>
<tr>
<td></td>
<td>GIRO</td>
<td>15,628 (32.9%)</td>
</tr>
<tr>
<td></td>
<td>Salary Deduction (SD)</td>
<td>5,799 (12.2%)</td>
</tr>
<tr>
<td></td>
<td>Banker's Order (BO)</td>
<td>593 (1.2%)</td>
</tr>
<tr>
<td></td>
<td>Others (MO)</td>
<td>189 (0.4%)</td>
</tr>
<tr>
<td>Status of Service</td>
<td>Serviced Policy (agent)</td>
<td>44,387 (92.0%)</td>
</tr>
<tr>
<td></td>
<td>Orphan Policy (no agent)</td>
<td>3,856 (8.0%)</td>
</tr>
<tr>
<td>Age at Purchase</td>
<td>In Years (Mean = 26.48, Std. Dev. = 9.95)</td>
<td></td>
</tr>
<tr>
<td>Size of Policy (SP)</td>
<td>In S$1000s (Mean = 36.05, Std. Dev. = 52.58)</td>
<td></td>
</tr>
</tbody>
</table>

*50 missing values; **664 missing values; Par. = Participating.

To ensure validity of the results, the data set is randomly partitioned into three parts, one part each for model construction, model valida-
tion, and model testing. Because the data set of 48,243 policies is large, a 70%, 20%, and 10% partitioning is used for model construction, validation, and testing, respectively. A significance level of 0.05 is used to guide the construction of the decision tree. For this study, the construction sample is used to construct the decision tree and the results are validated against the validation sample to avoid over-fitting the data. Finally, the reported $p$-values are computed with the test sample.

4.3 Decision Trees

The objective of decision trees is prediction and/or classification by dividing observations into mutually exclusive and exhaustive subgroups. The division is based on the levels of particular independent variables that have the strongest association with the dependent variable. In its basic form, the decision tree approach begins by searching for the independent variable that divides the sample in such a way that the difference with respect to the dependent variable is greatest among the divided subgroups. At the next stage, each subgroup is further divided into sub-subgroups by searching for the independent variable that divides each first-stage subgroup in such a way that the difference with respect to the dependent variable is greatest among the divided sub-subgroups. The independent variable selected need not be the same for each subgroup.

This process of division (or splitting in decision trees terminology) usually continues until either no further splitting can produce statistically significant differences in the dependent variable in the new subgroups or the subgroups are too small for any further meaningful division. The subgroups and sub-subgroups are usually referred to as nodes. Decision tree results can be converted into "if...then..." rules or graphically represented by a tree-like structure. These indicate the association of the independent variables with the dependent variable.

In the context of this paper, the decision tree shows the association between age, gender, type of policy, mode of payment, method of payment, status of service, and size of policy on one hand, and the average duration of terminated policies on the other. While univariate analysis can also be used to examine the relationship between duration and each independent variable to generate profiles that are more likely to experience policy termination at an aggregated level, decision trees lead to more powerful and richer results by:

1. Examining all independent variables at each split and selecting the particular variable that gives the best split;
2. Determining the threshold level to split interval variables such as age or size of policy;

3. Ordering the importance of the independent variables by their level or depth in the tree;

4. Segmenting the data into nodes that maximize the association between termination and the independent variables; and

5. Increasing usefulness of the results by generating rules and tree-like structures.

5 The Main Results

5.1 Decision Tree Structure

The structure of the decision tree, as summarized in Figure 1, consists of four levels. As mentioned earlier, a branch will not be split further if the difference in the mean duration of policy for all possible splits (based on all possible independent variables and all possible split thresholds) is not statistically significant at a 0.05 significance level. Two values are reported in the leading box of Figures 2 to 14. The first value refers to the mean duration of the particular branch and the second to the $p$-value for ANOVA. The abbreviations in Table 1 are used to describe the various branches.

At level 1 (see Figure 1), the overall sample is split by mode of payment into two branches: non-monthly and monthly. At level 2, the non-monthly branch is split by type of policy into four branches: EP, EN or WLN, WLP, and T or TO. The monthly branch is split by method of payment into three branches: CC, SD or MO, and GIRO or BO.

At level 3, for the non-monthly branch, three of the four branches at level 2 (i.e., EP, EN or WLN, and T or TO) are further split by size of policy whereas for the monthly branch, the three branches at level 2 are further split either by size of policy (CC and SD or MO branches) or type of policy (GIRO or BO branch). At level 4, only the monthly branch is further split either by type of policy (CC $\rightarrow$ SP branch at level 3) or age (SD or MO $\rightarrow$ SP, GIRO or BO $\rightarrow$ WLN or TO, and GIRO or BO $\rightarrow$ EN or EP branches at level 3).

5.2 Decision Tree Results

The results are presented in a top-down order to give due consideration to the relative importance of the variables. Generally, variables
Figure 1: Structure of the Decision Tree

Level 1:
Overall Sample
(Figure 2)

Level 2:
Non-Monthly
(Figure 3)

Level 3:
- EP
  (Figure 5)
- EN or WLN
  (Figure 6)
- WLP
  (Terminal node)
- Y or TO
  (Figure 7)

Monthly
(Figure 4)

Level 4:
- CC
  (Figure 8)
- SD or MO
  (Figure 9)
- GIRO or BO
  (Figure 10)

- SP
  (Terminal node)
- SP
  (Terminal node)
- SP
  (Figure 11)
- SP
  (Figure 12)
- WLP or TO
  (Figure 13)
- WLP or T
  (Terminal node)
- EN or EP
  (Figure 14)
appearing at a higher level in a decision tree are more important in determining the average duration of terminated policies than variables appearing at lower levels. Also, variables not included in the decision tree are not statistically significant.

5.2.1 Levels 1 and 2

As shown in Figure 1, the mode of payment is the first variable in the decision tree. The overall sample is split into two branches, that is, non-monthly mode policies and monthly mode policies, as shown in Figure 2. The ANOVA p-value is 0.0001.

The results indicate that mode of payment is the most important factor in determining the average duration of the terminated policies. This finding is reasonable, as non-payment of premium would directly lead to termination. Intuitively, policyholders paying premiums by the monthly mode have to consider continuation of payment more frequently than those who pay premiums on a non-monthly mode. Therefore, it is not surprising that mode of payment is selected by the decision tree algorithm as the most important factor in determining the duration of terminated policies.

5.2.2 Levels 2 and 3

As the decision tree branches down from level 2 to level 3, type of policy is selected as the next most important variable to split non-monthly mode policies (Figure 3) and method of payment as the next most important variable to split monthly mode policies (Figure 4). The splits in Figures 3 and 4 are statistically significant since the p-values are both equal to 0.0001.

Figure 2: The Decision Tree: Levels 1 and 2
Figure 3: The Decision Tree: Levels 2 and 3, Non-Monthly Branch

Figure 4: The Decision Tree: Levels 2 and 3, Monthly Branch
The results indicate that for non-monthly mode policies, different types of policies are associated with different average durations. One explanation for term and other policies (such as health insurance) being terminated earlier than all other types of policies (except whole life par) could be that they have no cash value to lose and are therefore more susceptible to competition and hence have a better chance of being replaced/terminated.

As for monthly mode policies, method of payment affects average duration significantly. The findings shown in Figure 4 are not surprising because salary deduction and GIRO (a popular type of electronic fund transfer service used in Singapore) can enhance the probability of premium payments, thus increasing the expected persistency.

The results presented above show that decision trees work on the strength of the effect that a current factor has on the preceding factor. Therefore, it is not surprising to see that different variables emerge for policies with different modes of payments, i.e., method of payment for monthly mode policies and type of policies for non-monthly monthly mode policies.

5.2.3 Levels 3 and 4

As the decision tree branches down from level 3 to level 4, for the non-monthly payment-mode branch, three of the four branches at level 3 are further split by size of policy (Figures 5 to 7); whereas for the monthly payment-mode branch, the three branches at level 3 are further split either by size of policy or type of policy (Figures 8 to 10). Again, the splits in Figures 5 to 10 are statistically significant, as all the p-values are 0.0001.

As shown in Figures 5 to 7 for non-monthly mode policies, a positive relationship exists between policy size and average duration for all types of policies except endowment par. Generally, policyholders with larger policies enjoy greater protection and have more reasons to hold on to them (see Figures 6 and 7). For endowment par, however, there is a negative relationship between the duration and size of policies (Figure 5). It is probable that because endowment par is the most expensive plan, it would be more difficult to keep a bigger policy in force.

As for monthly mode policies, size of policy is the next most important variable for cash/check and salary deduction and other methods of payment as shown in Figures 8 and 9. For GIRO and banker's order, type of policy is the next most important variable as shown in Figure 10.
Figure 5: The Decision Tree: Levels 3 and 4, Endowment Par (EP) Branch

Figure 6: The Decision Tree: Levels 3 and 4, Endowment or Whole Life Non-Par (EN or WLN) Branch
Lian, Wu, and Koh: Decision Tree Analysis

Figure 7: The Decision Tree: Levels 3 and 4, Term or Others (T or TO) Branch

Figure 8: The Decision Tree: Levels 3 and 4, Cash or Check (CC) Branch
Figure 9: The Decision Tree: Levels 3 and 4, Salary Deduction or Others (SD or MO) Branch

Figure 10: The Decision Tree: Levels 3 and 4, GIRO or Bank Order (GIRO or BO) Branch
Figures 8 and 9 indicate that the relationship between size of policy and average duration is non-monotonic (U-shaped) for cash/check method of payment and for salary deduction and others. For GIRO and banker's order payment, average duration varies by type of policy. These findings differ markedly from those for non-monthly mode policies. This suggests that simple rules of thumb cannot be applied to all situations.

5.2.4 Levels 4 and 5

At level 4 of the decision tree, only the monthly branch is further split either by type of policy (Figure 11) or age (Figures 12 to 14). It is noted that the p-values in Figures 11 to 14 are 0.0020, 0.0001, 0.0020, and 0.0001, respectively; therefore the differences of the average durations among various branches are statistically significant. Figure 11 shows that whole life non-par or other policies tend to last longer than whole life par or endowment par policies. This is probably because whole life non-par or other policies are cheaper and hence have a greater chance of keeping them in force.

Figures 12 to 14 indicate that generally policies payable on a monthly mode tend to have longer duration as the age of purchase of the policyholders increases. Apparently, older policyholders (at time of purchase) tend to value insurance policies more and keep them longer as they are more security conscious. (This finding does not apply to non-monthly mode policies and those of certain sizes in any significant manner.)

6 Closing Comments

The decision tree results can help insurance companies and their agents isolate and focus on the key variables and combinations of variables that increase persistency. On the flip side, they can avoid those undesirable combinations that decrease persistency. For example, the decision tree in Figure 1 shows that the most persistent policies (those with longest average duration) have an average of 6.27 years for policies with the following combination (interaction) of variables: non-monthly payment-mode policies (Figure 3), endowment non-par or whole life non-par (Figure 6), and size of policy exceeding $17,500. Such policies are expected to be more profitable. On the other hand, the shortest average duration is 0.97 years for policies with the following attributes: monthly payment-mode (Figure 4), cash/check payment (Figure 8), and size of policy between $10,500 and $20,500.
Figure 11: The Decision Tree: Levels 4 and 5, Size > $41,500 Branch

Figure 12: The Decision Tree: Levels 4 and 5, Size > $25,500 Branch
Figure 13: The Decision Tree: Levels 4 and 5, Whole Life Non-Par or Others (WLN or TO) Branch

Figure 14: The Decision Tree: Levels 4 and 5, Endowment Par or Non-Par (EP or EN) Branch
It is useful to know how age, gender, type of policy, mode of payment, method of payment, status of service, and size of policy are associated with the termination of policies—these associations were described earlier and illustrated in Figures 1 to 14. To gain full value from the decision tree results, however, the practical application of the results to an analytical model should be considered. In particular, life insurance companies and agents can focus on factors that are associated with the most significant improvement on the average duration of policies. Factors that are applicable to all policies such as mode of payment (i.e., non-monthly) and method of payment (i.e., non-cash or non-check) should be given greater emphasis. Measures such as premium discount, bonuses, and other incentives could be offered to induce policyholders to adopt a non-monthly mode and a non-cash (or non-check) payment method. These can have a positive effect on persistency. As for larger policies, which are already receiving size discounts (due to their lower unit cost of administration), there is an additional justification to do so: their better persistency.

While persistency of policies is critical to the profitability of life insurance companies, it is also the social responsibility of the companies to ensure that the policyholders are receiving value for their money. It is therefore important that insurance companies always keep track of the persistency of their policies and be guided by findings to take measures in ensuring that policyholders keep their policies in force for as long as possible. Longer average duration will not only yield greater profitability to the companies but also provide greater protection for the policyholders.

It should be noted that the data set used in this study is confined to terminated policies. The analysis is therefore restricted to the duration of terminated policies. If in-force policies are also included in the data set, a more comprehensive set of conclusions could be made.

Finally, it is hoped that this paper can make a contribution to the literature as well as the practice of life insurance companies and their agents.

References


Lian, Wu, and Koh: Decision Tree Analysis


A Comparative Study of Parametric and Nonparametric Estimators of Old-Age Mortality in Sweden

Peter Fledelius,* Montserrat Guillen,† Jens Perch Nielsen,‡ and Kitt Skovsø Petersen§

Abstract

A recent study of Swedish old-age mortality used a modified Gompertz-Makeham model with a linear hazard for the force of mortality. We propose an alternative model using smooth two-dimensional kernel hazard estimators and introduce a new estimator based on the multiplicative bias correction for the multivariate marker dependent hazard. The multiplicative bias correction

*Peter Fledelius, Ph.D., is business researcher in Royal & SunAlliance research department. He obtained his Ph.D. in finance from Aarhus Business School. His research areas include nonparametric statistical methods, mortality estimation and prediction, capital allocation and capital models regarding non-life insurance.

Dr. Fledelius’s address is: Aarhus School of Business & Codan, Codanhus, 60 Gammel Kongevej, DK-1790 Copenhagen V, DENMARK. E-mail: pfs@codan.dk

†Montserrat Guillen, Ph.D., is a full professor of insurance econometrics at the University of Barcelona. She has a Ph.D. in economics from the University of Barcelona. She is the director of a research group on risk in insurance and finance.

Dr. Guillen's address is: Department of Econometrics, University of Barcelona, Diagonal 690, 08034 Barcelona, SPAIN. E-mail: guillen@eco.ub.es.

‡Jens Perch Nielsen, Ph.D., is the director of research at RoyalSun & Alliance. He obtained his Ph.D. from the University of California (Berkeley). His research interests include nonparametric methods, statistical methods, life insurance and finance. He received the Danish doctorate in 2003.

Dr. Nielsen's address is: Codan, Codanhus, 60 Gammel Kongevej, DK-1790 Copenhagen V, DENMARK. E-mail: npj@codan.dk

§Kitt Skovsø Petersen, M.Sc., is an actuary at ATP PensionService. She obtained her B.Sc. and M.Sc. in Actuarial Science at the University of Copenhagen.

Dr. Petersen's address is: ATP PENSIONSERVICE, Kongens Vænge 8, 3400 Hillerød, DENMARK. E-mail: ksp@atp.dk

We thank Maria Undbergson for providing us the data; the Johan and Mimi Wessmann Foundation, the Danish Actuarial Association, and the Spanish CICYT SEC2001-3672 for financial support. We also thank the two anonymous referees and the editor for their helpful comments and suggestions.
appears to have great potential for estimating mortality rates at the highest ages. We also observe that mortality continues to increase at an exponential rate even in old-age.

Key words and phrases: longevity, expected remaining lifetime, kernel hazard estimation

1 Introduction

In a recent study of Swedish longevity, Lindbergson (2001) estimated old-age mortality by different parametric models and concluded that the Gompertz-Makeham model results in misspecified estimators of the force of mortality at advanced ages. She suggested a modified parametric model with a linear hazard to adjust for the misspecification of the standard Gompertz-Makeham curve. Similar misspecification of old age survival was shown by Vaupel et al. (1998). Using Swedish data, Vaupel et al. showed that the exponential Gompertz curve fits well until age 84, but for ages above 90 the Gompertz curve is unsuitable for describing human mortality.

In this paper we compare Lindbergson's results to those obtained by applying smooth two-dimensional kernel hazard estimators to the same Swedish data set. Moreover, we introduce a new estimator, which is based on the multiplicative bias correction for the multivariate hazard, and show it is suitable to analyze old age mortality. Early actuarial applications of kernel smoothing techniques include Copas and Haberman (1983), Ramlau-Hansen et al. (1987), Gavin, Haberman, and Verrall (1994) and (1995), and Nielsen and Voldsgaard (1996).

We take as our starting point the two-dimensional local constant mortality estimator, which was defined in Nielsen and Linton (1995) and applied to Danish and Spanish mortality experience data in Felipe, Guillen, and Nielsen (2001) and Fledelius et al. (2004). This estimator cannot be used to estimate old-age mortality because of the rapidly increasing mortality and rapidly decreasing exposure at old ages. The bias properties of the local constant estimator imply that it underestimates the true mortality significantly. The local linear estimator of Nielsen (1998a), however, does not have this weakness. We therefore conclude that it is important to use local linear estimators rather than local constant estimators while dealing with the problem of estimating old-age mortality.

Another estimator used is the multidimensional analogue of the one-dimensional multiplicatively bias-corrected kernel hazard estima-
tor due to Nielsen (1998b). The bias properties of the multiplicative bias correction method make it suitable for estimating old-age mortality. A study based on goodness-of-fit confirms this assertion.

Upon comparing our approach to Lindbergson's standard Gompertz-Makeham model we conclude that the quick and automatic method of the multiplicatively bias corrected multidimensional kernel hazard estimator is better because it does not require us to fix a parametric model. Moreover, in the nonparametric methodology the estimation procedure is straightforward, as no optimization is required.

2 Local Constant Versus Local Linear Estimation

Almost all observed individuals in actuarial data sets are either left truncated or right censored or both at the same time. Although actuaries traditionally have been able to take this into account, the best mathematical description of this type of model seems to be based on modern counting process theory, where left truncation and right censoring are natural elements of any model construction.

While the mathematical formulation of the stochastic intensity process below might seem overwhelming at first glance, it really only says that every individual has a mortality depending on age and calendar time and that death of this individual only can be observed while the individual is alive and included in the study.

Suppose we are observing n individuals. Let $N_i^{(n)}$ denote the number of observed failures for the $i^{th}$ individual in the time interval $[0, 1]$, $i = 1, \ldots, n$. Assume $N^{(n)} = (N_1^{(n)}, \ldots, N_n^{(n)})$ is an $n$-dimensional counting process with respect to an increasing, right continuous, complete filtration $\mathcal{F}_t^{(n)}$, $t \in [0, 1]$, i.e., one that obeys les conditions habituelles, see Andersen et al. (1992, page 60). The random intensity process of $N^{(n)}$, $\lambda^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_n^{(n)})$, is modeled as depending on chronological time and age, i.e.,

$$\lambda_i^{(n)}(t) = \alpha(X_i(t), t) Y_i^{(n)}(t), \quad (1)$$

where $\alpha(\cdot, \cdot)$ is the force of mortality (the functional form of which is unrestricted), $t$ is chronological time, $Y_i^{(n)}(t)$ is a predictable process taking values 0 or 1, while $X_i(t)$ is the age at time $t$ of the $i^{th}$ individual. $\{Y_i^{(n)}(t) = 1\}$ indicates the $i^{th}$ individual is under risk at time $t$.

First we define the local constant estimator of $\alpha(\cdot, \cdot)$, introduced by Fusaro, Nielsen, and Scheike (1993) and Nielsen and Linton (1995), as
\[
\hat{\alpha}(x,t) = \frac{O_{x,t}}{E_{x,t}},
\]  
(2)

where \(O_{x,t}\) and \(E_{x,t}\) are the smoothed number of deaths and smoothed exposure, respectively, at age \(x\) and time \(t\). The general form of \(O_{x,t}\) and \(E_{x,t}\) is:

\[
O_{x,t} = \sum_{i=1}^{n} \int_{T_0}^{T_1} K_{b_1}(x - X_i(s)) K_{b_2}(t - s) \, dN_i(s)
\]
(3)

and

\[
E_{x,t} = \sum_{i=1}^{n} \int_{T_0}^{T_1} K_{b_1}(x - X_i(s)) K_{b_2}(t - s) \, Y_i(s) \, ds
\]
(4)

where the kernel \(K\) is a one-dimensional kernel and \(\mathbf{b} = (b_1, b_2)\) a two-dimensional bandwidth-vector. Note that \(T_0\) and \(T_1\) refer, respectively, to the starting and the final calendar time points of the study, i.e., the study runs from \(T_0\) to \(T_1\). The time frame used in this paper runs from 1988 to 1997.

When the kernel is defined as

\[
K(x) = \begin{cases} 
1 & |x| < 1 \\
0 & \text{otherwise},
\end{cases}
\]

\(O_{x,t}\) denotes the actual observed number of deaths and \(E_{x,t}\) denotes the actual amount of exposure, both observed in the chronological time interval \([t - b_1, t + b_1]\) and in the age interval \([x - b_2, x + b_2]\). In this case the local constant hazard estimator is the well known "occurrence divided by exposure rate." In practice, however, smooth kernels are used to obtain realistically smooth mortality estimators.

The effect of the choice of the kernel function and the bandwidths has extensively been discussed in the literature; see, for example, Silverman (1986) or Gavin, Haberman, and Verrall (1994). The choice, however, seems to be rather unimportant as long as the kernel is continuous and smooth and has a simple mathematical expression.

Our analysis in this paper is based on the Epanechnikov kernels, i.e.,

\[
K_{b_1}(t - s) = 0.75I(|t - s| \leq b_1) \left[ 1 - \left( \frac{t - s}{b_1} \right)^2 \right]
\]
(5)
and

\[ K_{b_2}(x - X_i(s)) = 0.75I(|x - X_i(s)| \leq b_2) \left[ 1 - \left( \frac{x - X_i(s)}{b_2} \right)^2 \right], \]  
(6)

which are defined on a bounded support. An important feature of Epanechnikov kernels is that the resulting estimator has the well known occurrence divided by exposure construction. The occurrence and exposure data used throughout this paper have been constructed as described in Lindbergson (2001).

Let \( W_i(s) = (X_i(s), s), z = (x, t) \) and

\[ K_b(u) = K_{b_1}(u_1)K_{b_2}(u_2) \]

for \( u = (u_1, u_2) \). The local linear estimator is defined as

\[ \hat{\alpha}_L(x, t) = \sum_{i=1}^{n} \int K_{z,b}(z - W_i(s)) \, dN_i(s), \]  
(7)

where the corrected kernel \( K_{z,b} \) can be written as

\[ K_{z,b}(u) = \frac{K_b(u) - K_b(u)c_1^T D^{-1} c_1}{c_0 - c_1^T D^{-1} c_1}, \]  
(8)

with \( c_1 = (c_{11}, ..., c_{1d})^T \), where \( T \) denotes transpose, and \( D = \{d_{jk}\} \) is a \( d \times d \) matrix. The elements are

\[ c_0 = \sum_{i=1}^{n} \int K_b(z - W_i(s)) \, Y_i(s) \, ds, \]  
(9)

\[ c_{1j} = \sum_{i=1}^{n} \int K_b(z - W_i(s)) \left( z_j - W_{ij}(s) \right) \, Y_i(s) \, ds \]  
(10)

and

\[ d_{jk} = \sum_{i=1}^{n} \int K_b(z - W_i(s)) \left( z_j - W_{ij}(s) \right) \left( z_k - W_{ik}(s) \right) \, Y_i(s) \, ds \]  
(11)

where \( z_j \) is the \( j^{th} \) element of \( z \) and \( W_{ij}(s) \) is the \( j^{th} \) element of \( W_i \).

From Nielsen (1998a), the equations for the bias in the local constant and the local linear estimators are the same, except that the equation
for the bias of the local constant estimator has the following extra term at chronological time $t$ and age $x$:

$$b_1^2 \int v^2 K_1(v) dv \left( \frac{\partial \alpha}{\partial x} (x,t) \frac{\partial \varphi}{\partial x} (x,t) \right) (\varphi(x,t))^{-1}$$

$$+ b_2^2 \int v^2 K_2(v) dv \left( \frac{\partial \alpha}{\partial t} (x,t) \frac{\partial \varphi}{\partial t} (x,t) \right) (\varphi(x,t))^{-1},$$

where $\varphi(x,t) = f_t(x)\gamma(t)$ corresponds to an exposure density describing the quantity of exposure.\(^1\) The term $\gamma(t)$ describes exposure and is approximately equal to $n^{-1} \sum_{i=1}^{n} Y_i(t)$, while $f_t(x)$ is the density of age $X_i(t)$ at time $t$.

The local constant estimator's dependency on $\varphi$ makes it unfit for estimating mortality at advanced ages because the rapid increase in mortality rates and the rapid decrease in exposure at these ages imply that the extra bias component of the local constant estimator becomes large. As a consequence, the local constant estimator significantly underestimates the true mortality.

Local constant estimates and the local linear estimates were calculated for the ages 90 and above using the Swedish data set from 1988 to 1997. Figures 1 and 2 show the smooth two-dimensional local constant estimates and the smooth two-dimensional local linear estimates, respectively, of the force of mortality. The smooth two-dimensional kernel hazard estimation was derived using a bandwidth-vector of $(7,4)$. The choice of the bandwidth-vector was based on previous studies.

From Figures 1 and 2, one can see that both surfaces are increasing with respect to age, but the rate of increase is larger for the local linear estimates. It is hard to see from these figures whether there has been a change in hazard rates over time. The estimates of the force of mortality in the years 1989, 1992, and 1996 have been outlined. The estimates in these years are presented in Figure 3 for the local constant estimator on the left and for the local linear estimator on the right.

The mortality curves in Figure 3 show that there has been a slight decline in female mortality over the decade, which is similar to that observed by Fledelius et al. (2004) for Danish data. The time trend is not so clear for ages 105 and above when comparing the years 1992

\(^1\)In the field of regression, a similar type of difference exists between the bias of the local constant estimator and the local linear estimator. The local linear regression estimator has an advantage because its bias does not depend on the design; see Fan and Gijbels (1996). The dependency of the bias of the local constant marker dependent hazard estimator on $\varphi$ corresponds to the dependency of the local constant regression estimator on the design.
Figure 1: Local Constant Estimates for Swedish Women

Figure 2: Local Linear Estimates for Swedish Women
Local Constant Estimates

Local Linear Estimates

Figure 3: Estimated Force of Mortality for Women Age 90-111

and 1996. One reason for this could be the scarcity of data in this age range. The difference between the local constant and the local linear estimates will now be explored.

Figures 4 and 5 show the observed mortality ratios and the estimated modified Gompertz-Makeham curve\(^2\) due to Lindbergson (2001) for women age 90 to 111 and men age 90 to 110, respectively, for the period 1988-1992 on the left and for the period 1993-1997 on the right. Figures 4 and 5 also depict the local constant estimates and the local linear estimates for women of age 90 to 111 for the year 1990 and 1995.

It is clear from Figures 4 and 5 that the local constant estimator underestimates the mortality rates for both men and women of age 90 and above. It also appears that the local linear estimator gives a better estimate of the mortality rates than the modified Gompertz-Makeham curve.

The smooth two-dimensional kernel hazard estimates used to obtain the 1990 and 1995 curves in Figures 4 and 5 have used a bandwidth-vector of (7, 4).

---

\(^2\) The standard Gompertz-Makeham curve for the force of mortality at age \(x\) is \(\mu_x = a + be^{cx}\), and Lindbergson's (2001) modified Gompertz-Makeham curve specifies the force of mortality as \(\mu_x = a + be^{cx} + I\{x \geq w\}k(x - w)\), where \(I(A)\) is an indicator of the event \(A\), and \(w\) is an age in excess of 90.
Figure 4: Various Estimates of Death Rates for Swedish Women Age 90–111

Figure 5: Various Estimates of Death Rates for Swedish Men Age 90–110
3 Fitting a Gompertz-Makeham Curve for Old Ages

Another way of estimating the mortality rates for age 90 and above is to fit a Gompertz-Makeham function with parameters estimated only by the observations for these ages. The standard method for fitting is to use maximum likelihood techniques with a model formulated as a generalized linear model. (See, for example, Renshaw, 1991.) Following the approach of Lindbergson (2001), we fitted the parameters using the minimum chi-square method, i.e., by minimizing the expression

\[ Q_1^2 = \sum_{x=90} (\hat{\mu}_x - g(x, a, b, c))^2 \],

where \( \hat{\mu}_x \) is the observed rate of mortality at age \( x \), \( R_x \) is the exposure at age \( x \), \( g(x, a, b, c) \) is the Gompertz-Makeham function for ages above 90, which is defined as

\[ g(x, a, b, c) = a + b \exp(cx), \quad x \geq 90, \]

and \( a, b, \) and \( c \) are the parameters to be estimated. Lindbergson (2001) derives the asymptotic normality of the estimators; more details on efficiency of these estimators can be found in Hoem (1976). The minimum chi-square method puts more weight than the maximum likelihood method on the ages where exposure is smaller, so it will generally provide a better fit in terms of the \( Q_1^2 \) measure for the very old ages. \(^3\)

A multivariate Newton-Raphson's method (see, for example, Acton, 1990) is used to estimate the Gompertz-Makeham parameters \( a, b, \) and \( c \) that minimize \( Q_1^2 \) separately for men and women using the aggregated data for each of the periods 1988-1992 and 1993-1997. Parameter estimates are listed in Table 1.

In Lindbergson (2001) the minimum chi-square method is used to fit a Gompertz-Makeham function with parameters estimated by the observations for age 28-90, i.e., by minimizing the expression

\[ Q_2^2 = \sum_{x=28}^{90} (\hat{\mu}_x - g(x, a, b, c))^2 \hat{\mu}_x / R_x \],

This is done separately for men and women for each of the periods 1988-1992 and 1993-1997. We have been able to re-create the estimates of the parameters for each of Lindbergson's four Gompertz-

\(^3\)Another possible approach, which is not addressed here, is the one based on Perks-type models for old age mortality, which have the property that the rates have a point of inflection and exhibit a plateau effect at the oldest ages; see Wang and Brown (1998).
Table 1

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>-0.47056603</td>
<td>-0.57298188</td>
</tr>
<tr>
<td>$b$</td>
<td>0.02884032</td>
<td>0.04724583</td>
</tr>
<tr>
<td>$c$</td>
<td>0.03451867</td>
<td>0.03052308</td>
</tr>
</tbody>
</table>

Makeham functions. For the modified Gompertz-Makeham function using a linear hazard for ages above 95, Lindbergson (2001) also uses the minimum chi-square method, i.e., by minimizing the expression

$$Q_3^2 = \sum_{x \geq 28} \frac{(\hat{\mu}_x - g(x, a, b, c, k))^2}{\hat{\mu}_x/R_x}$$

(17)

where

$$g(x, a, b, c, k) = \begin{cases} a + be^{cx} & x \leq 95 \\ a + bc^{x} + k(x - 95) & x > 95 \end{cases}$$

(18)

This is done separately for men and women for each of the periods 1988-1992 and 1993-1997 with parameters estimated by the observations for ages 28 and above. We have also been able to re-create the estimates of the parameters of each of Lindbergson's four modified Gompertz-Makeham curves.

Lindbergson's standard and modified Gompertz-Makeham curves are shown in Figure 6 for women and for men age 90 and above in the periods 1988-1992 and 1993-1997. It also shows the local linear estimates for women age 90-111 and men of age 90-110 in the years 1990 and 1995. These local linear estimates are also based on a bandwidth-vector of (7,4). Notice that the local linear and the fitted Gompertz-Makeham curves are close for the first half of the age range. The modified Gompertz-Makeham curve seems to overestimate the mortality around the age where the linear hazard begins, while for older ages it is usually below the other estimates. The local linear estimates for men in 1990 show a rapid decline at very old ages. This is due to the rapid decline of exposure. Nevertheless, except for ages above 108, the local linear estimates show an increase in the mortality for men from 1990 to 1995. The local linear estimates also indicate a
Figure 6: Death Rates for Swedish Women and Men in 1990 and 1995

decrease in mortality for women from 1990 to 1995 for all ages above 90.

4 Multiplicative Bias Correction

A multiplicative bias correction can be used to improve the smooth two-dimensional kernel hazard estimation of the local linear estimator of the force of mortality, as follows:

- First, an initial estimate of the force of mortality is needed.

- Next, the exposure is multiplied by this initial estimator, and the smooth two-dimensional kernel hazard estimators then are applied. The result is an estimator of the multiplicative error.

- Finally the initial estimator is multiplied by the estimated multiplicative error.

Multiplicative bias correction with a nonparametric start was introduced in nonparametric regression in Linton and Nielsen (1994), in density estimation in Jones, Linton, and Nielsen (1995), and in one-dimensional kernel hazard estimation in Nielsen (1998b), who took
Ramlau-Hansen’s (1983) estimator as the starting point. Nielsen and Tanggaard (2001) suggest the concept of local linear kernel hazard estimation in the one-dimensional case and show that the multiplicative bias correction can be understood as a procedure for minimizing a least squares criterion.\(^4\)

Nielsen and Tanggaard (2001, page 681) point out that one version of the multiplicative bias reduction method simply can be understood in the following way: in the local linear estimation procedure, the exposure term \(Y_i(t)\) is replaced by \(\tilde{\alpha}(X_i(t), t)Y_i(t)\), where \(\tilde{\alpha}\) is our preliminary estimator of the hazard rate. This estimation procedure results in an estimator \(\hat{\gamma}\) of the multiplicative correction \(\hat{\gamma} = \alpha/\tilde{\alpha}\). We consider both the fully nonparametric approach where the pilot estimated hazard \(\tilde{\alpha}\) equals our local linear estimator and the hazard analogue to the semiparametric approach of Hjort and Glad (1995), where \(\hat{\alpha}\) is estimated using the parametric Gompertz-Makeham shape.

The final multiplicatively bias corrected estimator, \(\hat{\alpha}_M(z)\), is obtained by multiplying the initial estimator \(\tilde{\alpha}\) by the estimate of the multiplicative error

\[
\hat{\alpha}_M(z) = \tilde{\alpha}(z)\hat{\gamma}(z),
\]

where the estimated multiplicative error \(\hat{\gamma}(z)\) is calculated by applying the local linear estimator with the original occurrence and an exposure equal to the original exposure multiplied by the initial estimator.

While we have not derived the theoretical properties of the multiplicatively bias corrected kernel hazard estimator, then we can conclude from Nielsen (1998b) and Nielsen and Tanggaard (2001) that in one dimension this estimator almost has zero bias for old ages when a Gompertz-Makeham shape is assumed. We expect a similar result to hold in our case.

Figure 7 shows the local linear estimates with and without a multiplicative correction and the ratios of these estimates for women of age 90–111 in the period 1988–1992. The multiplicative correction is obtained with four different initial estimates of the force of mortality:

- The first multiplicative correction of the local linear estimator is obtained by multiplying the exposure at age \(x\) and year \(t\) by the local linear estimator of the force of mortality in age \(x\) and year \(t\);

\(^4\)Hjort and Glad (1995) apply multiplicative bias correction with a parametric start to density estimation. This so-called semiparametric approach to multiplicative bias correction has a great advantage while the underlying true density is close to the parametric start.
• The second multiplicative correction is obtained by multiplying the exposure at age $x$ and year $t$ by the Gompertz-Makeham estimate of the force of mortality at age $x$ calculated in Section 3 for 1988-1992;

• The third multiplicative correction is obtained the same way as the second except that the period used is 1993-1997; and

• The fourth multiplicative correction is obtained the same way as the second multiplicative correction except that the period used is 1988-1997.

All of the local linear estimators illustrated in Figure 7 have been estimated using a bandwidth-vector of $(7, 4)$. Notice that three of the five curves are almost equal because the local linear estimators obtained by multiplying the exposure by the Gompertz-Makeham estimate of the force of mortality calculated in Section 3 are almost the same (whether we use the Gompertz-Makeham estimated for the period 1988-1992, for the period 1993-1997, or for both periods).

The ratios displayed are those of the estimates with multiplicative correction based on the Gompertz-Makeham for the period 1993-1997 divided by the baseline and the ratio of the multiplicative correction results based on the Gompertz-Makeham for the period 1988-1992 in the first five years and for the period 1993-1997 in the last five years divided by the baseline. The baseline used is the multiplicative correction of the local linear estimator based on the Gompertz-Makeham estimated for the period 1988-1992.

Because the multiplicative correction of the local linear estimator obtained by multiplying the exposure by our Gompertz-Makeham estimate is almost equal (the difference is less than one percent) no matter in which period the Gompertz-Makeham function is estimated, we have only considered multiplication with the Gompertz-Makeham estimated for the period 1988-1992 in the rest of the analysis.

Figures 8 and 9 show Lindbergson's modified Gompertz-Makeham and Gompertz-Makeham curves estimated earlier for women and men, respectively, age 90 and above for the periods 1988-1992 and 1993-1997. They also depict the local linear estimator without the multiplicative correction and the local linear estimator with multiplicative correction based on: (i) the local linear estimator, and (ii) the Gompertz-Makeham estimator calculated in Section 3. All local linear estimators illustrated in Figures 8 and 9 have been estimated using a bandwidth-vector of $(7, 4)$. 
Figure 7: Estimates and Ratios Based on Four Multiplicative Corrections

Figure 8: Death Rates for Swedish Women Age 90-111
Figures 8 and 9 illustrate that the local linear estimator with the multiplicative correction obtained by multiplying the exposure by the local linear estimator of the force of mortality gives a different estimate from that produced by the local linear estimator without the multiplicative correction. These figures also illustrate that the local linear estimator with the multiplicative correction obtained by multiplying the exposure by the Gompertz-Makeham estimate calculated in Section 3 for the period 1988-1992 is similar to the local linear estimator without the multiplicative correction.

5 Minimum Chi-Square

As a supplement to the graphical comparison of the estimates of the force of mortality we have also calculated the value of expression \(Q^2\), which is minimized when using the minimum chi-square method to estimate the force of mortality. This expression is defined above where it is used to fit a Gompertz-Makeham curve for old ages only. The only difference is that instead of the Gompertz-Makeham function \(g(x, a, b, c)\) at age \(x\) for ages above 90, we now insert \(g(x)\), which is
our nonparametric or semi-parametric estimated force of mortality at age \( x \) for ages above 90.

We have done the calculation separately for men and women for each of the periods 1988–1992 and 1993–1997. The value of \( Q^2 \) has been calculated corresponding to six ways of estimating the force of mortality \( g(x) \) and the results are presented in Table 2. The six estimates of \( g(x) \) are as follows:

(I) The standard Gompertz-Makeham curve for age 90 and above;

(II) The modified Gompertz-Makeham curve in Lindbergson (2001) that specifies a linear hazard at old ages;

(III) The local constant estimator;

(IV) The local linear estimator without the multiplicative correction;

(V) The local linear estimator with the multiplicative correction obtained by multiplying the exposure by the local linear estimator; and

(VI) The local linear estimator with multiplicative correction obtained by multiplying the exposure by the Gompertz-Makeham estimate.

The value of \( Q^2 \) is shown for the nonparametric estimates when a bandwidth-vector of (7, 4) is used. Values obtained with bandwidth-vectors of (5, 3) and (11, 6) are also displayed.

From Table 2 we conclude that Model V coupled with a bandwidth-vector of (7, 4) is best because it produces the lowest \( Q^2 \) for both men and women. In the rest of the analysis this local linear estimator is referred to as the local linear estimator with multiplicative correction. Table 2 also indicates that in terms in the chi-square measure the modified Gompertz-Makeham curve (II) in Lindbergson (2001) is not as good an estimate of the force of mortality as any of the three local linear estimates. This is true for several different choices of bandwidth-vector.

Figure 10 shows the ratio obtained when dividing the results of the local linear estimator with multiplicative correction by the modified Gompertz-Makeham function in Lindbergson (2001). So here the purely nonparametric approach is compared to the existing parametric approach. Figure 10 (left) shows the ratios for women, and Figure 10 (right) shows the ratio for men. Note that the nonparametric results refer to the single years 1990 and 1995, while the results in Lindbergson are based on the periods 1988–1992 and 1993–1997. This is not unreasonable, because, due to the choice of kernel and bandwidth-vector, all observations contribute to the nonparametric estimator in the years
Table 2
Values of $Q^2$ for Women and Men Age 90 and Above

<table>
<thead>
<tr>
<th>Model</th>
<th>(7, 4)</th>
<th>(5, 3)</th>
<th>(11, 6)</th>
<th>(7, 4)</th>
<th>(5, 3)</th>
<th>(11, 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women Age 90 and Above</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>1461.979</td>
<td>504.123</td>
<td>5194.357</td>
<td>1808.850</td>
<td>646.851</td>
<td>5968.277</td>
</tr>
<tr>
<td>IV</td>
<td>40.250</td>
<td>23.463</td>
<td>38.931</td>
<td>37.484</td>
<td>23.487</td>
<td>53.277</td>
</tr>
<tr>
<td>VI</td>
<td>45.817</td>
<td>23.975</td>
<td>73.949</td>
<td>36.560</td>
<td>22.793</td>
<td>66.872</td>
</tr>
<tr>
<td>Men Age 90 and Above</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>(7, 4)</td>
<td>(5, 3)</td>
<td>(11, 6)</td>
<td>(7, 4)</td>
<td>(5, 3)</td>
<td>(11, 6)</td>
</tr>
<tr>
<td>II</td>
<td>121.667</td>
<td>121.667</td>
<td>121.667</td>
<td>61.522</td>
<td>61.522</td>
<td>61.522</td>
</tr>
<tr>
<td>III</td>
<td>591.326</td>
<td>213.201</td>
<td>2113.362</td>
<td>809.230</td>
<td>291.742</td>
<td>2602.122</td>
</tr>
</tbody>
</table>

1990 and 1995. Note that except for men in the period 1988-1992, the nonparametric methods provide a larger force of mortality than the parametric methods for the very old age brackets.

We conclude that the nonparametric results produce lower estimates of the force of mortality around the age where the linear modification starts, as suggested by Lindbergson (2001). This result tells us that the parametric approximation may be overestimating mortality around this region while underestimating it in older ages. The fitted Gompertz-Makeham curve in Section 3 supports these conclusions.

6 Remaining Life Expectancy

The future life expectancy at ages 90 and over for the Swedish elderly are calculated and compared using our two-dimensional kernel estimates to those found using the parametric approach based on the hypothesis of a linear hazard for old ages. The estimator used is the local linear estimator with the multiplicative correction based on the local linear estimator of the force of mortality using a bandwidth-vector of (7, 4). The future life expectancy at age $x$, $\hat{e}_x$, is calculated as follows:
The results for women and men of age 90 and above are presented in Tables 3 and 4, respectively. The estimators used are the local linear estimator of the force of mortality with multiplicative correction (LLwMC) in the years 1990, 1995, and 1997 and Lindbergson's modified Gompertz-Makeham (MGM) for periods 1988-1992 and 1993-1997.

Figure 11 shows the ratios of the future life expectancies for women and men based on the local linear estimator with multiplicative correction to those based on the modified Gompertz-Makeham function. We see the LLwMC-based \( \hat{e}_x \)'s are lower than those based on MGM. The difference increases with age due to the underestimation of the hazard rate at very old ages for the modified Gompertz-Makeham method. Notice that \( \hat{e}_x \) increases for women of all ages from 1990 to 1995 under both methods. For men, \( \hat{e}_x \) increases in this period only for men ages 90 to 94. For the rest of the older ages male life expectancy seems to be decreasing under the LLwMC approach, while under the MGM approach, \( \hat{e}_x \) increases from ages 106 and above in that time period. When looking at the differences between the estimated life expectancy in 1995

\[
\hat{e}_x = \int_0^\infty \exp(-\int_0^s \alpha(x+u,t)\,du)\,ds. \tag{20}
\]
### Table 3

**Future Life Expectancy ($\hat{e}_x$) for Women**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>3.991</td>
<td>3.782</td>
<td>4.170</td>
<td>3.995</td>
<td>3.975</td>
</tr>
<tr>
<td>91</td>
<td>3.672</td>
<td>3.468</td>
<td>3.838</td>
<td>3.623</td>
<td>3.630</td>
</tr>
<tr>
<td>92</td>
<td>3.375</td>
<td>3.180</td>
<td>3.529</td>
<td>3.318</td>
<td>3.312</td>
</tr>
<tr>
<td>93</td>
<td>3.101</td>
<td>2.918</td>
<td>3.244</td>
<td>3.039</td>
<td>3.020</td>
</tr>
<tr>
<td>94</td>
<td>2.851</td>
<td>2.679</td>
<td>2.984</td>
<td>2.785</td>
<td>2.754</td>
</tr>
<tr>
<td>95</td>
<td>2.627</td>
<td>2.461</td>
<td>2.750</td>
<td>2.555</td>
<td>2.514</td>
</tr>
<tr>
<td>96</td>
<td>2.432</td>
<td>2.263</td>
<td>2.546</td>
<td>2.347</td>
<td>2.298</td>
</tr>
<tr>
<td>97</td>
<td>2.273</td>
<td>2.083</td>
<td>2.379</td>
<td>2.162</td>
<td>2.104</td>
</tr>
<tr>
<td>98</td>
<td>2.164</td>
<td>1.920</td>
<td>2.262</td>
<td>1.997</td>
<td>1.934</td>
</tr>
<tr>
<td>99</td>
<td>2.090</td>
<td>1.772</td>
<td>2.179</td>
<td>1.852</td>
<td>1.787</td>
</tr>
<tr>
<td>100</td>
<td>2.020</td>
<td>1.641</td>
<td>2.101</td>
<td>1.729</td>
<td>1.663</td>
</tr>
<tr>
<td>101</td>
<td>1.954</td>
<td>1.524</td>
<td>2.029</td>
<td>1.625</td>
<td>1.562</td>
</tr>
<tr>
<td>102</td>
<td>1.893</td>
<td>1.420</td>
<td>1.961</td>
<td>1.535</td>
<td>1.479</td>
</tr>
<tr>
<td>103</td>
<td>1.835</td>
<td>1.329</td>
<td>1.898</td>
<td>1.460</td>
<td>1.412</td>
</tr>
<tr>
<td>104</td>
<td>1.781</td>
<td>1.253</td>
<td>1.838</td>
<td>1.400</td>
<td>1.356</td>
</tr>
<tr>
<td>105</td>
<td>1.729</td>
<td>1.198</td>
<td>1.782</td>
<td>1.353</td>
<td>1.304</td>
</tr>
<tr>
<td>106</td>
<td>1.680</td>
<td>1.151</td>
<td>1.729</td>
<td>1.295</td>
<td>1.234</td>
</tr>
<tr>
<td>107</td>
<td>1.634</td>
<td>1.056</td>
<td>1.679</td>
<td>1.188</td>
<td>1.115</td>
</tr>
<tr>
<td>108</td>
<td>1.590</td>
<td>0.889</td>
<td>1.632</td>
<td>1.025</td>
<td>0.953</td>
</tr>
<tr>
<td>109</td>
<td>1.549</td>
<td>0.658</td>
<td>1.587</td>
<td>0.800</td>
<td>0.745</td>
</tr>
<tr>
<td>110</td>
<td>1.509</td>
<td>0.372</td>
<td>1.544</td>
<td>0.473</td>
<td>0.448</td>
</tr>
</tbody>
</table>
Table 4

Future Life Expectancy \( (\hat{\mu}_x) \) for Men

<table>
<thead>
<tr>
<th>Age ( x )</th>
<th>MGM 1990</th>
<th>LLwMC 1990</th>
<th>MGM 1995</th>
<th>LLwMC 1995</th>
<th>LLwMC 1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>3.181</td>
<td>3.103</td>
<td>3.289</td>
<td>3.159</td>
<td>3.150</td>
</tr>
<tr>
<td>91</td>
<td>2.946</td>
<td>2.872</td>
<td>3.034</td>
<td>2.905</td>
<td>2.891</td>
</tr>
<tr>
<td>92</td>
<td>2.733</td>
<td>2.658</td>
<td>2.798</td>
<td>2.668</td>
<td>2.649</td>
</tr>
<tr>
<td>93</td>
<td>2.542</td>
<td>2.461</td>
<td>2.581</td>
<td>2.450</td>
<td>2.427</td>
</tr>
<tr>
<td>94</td>
<td>2.379</td>
<td>2.281</td>
<td>2.383</td>
<td>2.250</td>
<td>2.225</td>
</tr>
<tr>
<td>95</td>
<td>2.251</td>
<td>2.115</td>
<td>2.209</td>
<td>2.070</td>
<td>2.042</td>
</tr>
<tr>
<td>96</td>
<td>2.175</td>
<td>1.964</td>
<td>2.061</td>
<td>1.908</td>
<td>1.875</td>
</tr>
<tr>
<td>97</td>
<td>2.133</td>
<td>1.826</td>
<td>1.952</td>
<td>1.764</td>
<td>1.726</td>
</tr>
<tr>
<td>98</td>
<td>2.093</td>
<td>1.701</td>
<td>1.901</td>
<td>1.636</td>
<td>1.592</td>
</tr>
<tr>
<td>99</td>
<td>2.054</td>
<td>1.594</td>
<td>1.891</td>
<td>1.521</td>
<td>1.471</td>
</tr>
<tr>
<td>100</td>
<td>2.016</td>
<td>1.511</td>
<td>1.881</td>
<td>1.420</td>
<td>1.367</td>
</tr>
<tr>
<td>101</td>
<td>1.980</td>
<td>1.457</td>
<td>1.871</td>
<td>1.333</td>
<td>1.286</td>
</tr>
<tr>
<td>102</td>
<td>1.945</td>
<td>1.423</td>
<td>1.862</td>
<td>1.254</td>
<td>1.227</td>
</tr>
<tr>
<td>103</td>
<td>1.912</td>
<td>1.401</td>
<td>1.852</td>
<td>1.167</td>
<td>1.185</td>
</tr>
<tr>
<td>104</td>
<td>1.879</td>
<td>1.391</td>
<td>1.843</td>
<td>1.065</td>
<td>1.163</td>
</tr>
<tr>
<td>105</td>
<td>1.848</td>
<td>1.366</td>
<td>1.833</td>
<td>0.954</td>
<td>1.161</td>
</tr>
<tr>
<td>106</td>
<td>1.817</td>
<td>1.333</td>
<td>1.824</td>
<td>0.857</td>
<td>1.212</td>
</tr>
<tr>
<td>107</td>
<td>1.787</td>
<td>1.286</td>
<td>1.815</td>
<td>0.764</td>
<td>1.258</td>
</tr>
<tr>
<td>108</td>
<td>1.759</td>
<td>1.206</td>
<td>1.806</td>
<td>0.633</td>
<td>1.130</td>
</tr>
<tr>
<td>109</td>
<td>1.731</td>
<td>1.075</td>
<td>1.797</td>
<td>0.453</td>
<td>0.775</td>
</tr>
<tr>
<td>110</td>
<td>1.704</td>
<td>0.739</td>
<td>1.788</td>
<td>0.291</td>
<td>0.421</td>
</tr>
</tbody>
</table>
Figure 11: Ratios of Future Life Expectancies for Swedish Women and Men

and 1997 (i.e., the last two columns of Tables 3 and 4) our estimates indicate that women's life expectancy has slightly decreased. For men, there seems to be little difference in life expectancy for the ages between 90 and 102, while it has increased above age 102.

7 Conclusions

In our study of the mortality experience for the Swedish population age 90 and above in the period from 1988 to 1997, the following are our major findings:

- The local constant estimator tends to underestimate the force of mortality;
- The results obtained by the modified Gompertz-Makeham method seem to be influenced by the choice of the age where the linear hazard starts. In general, the modified Gompertz-Makeham estimates of the force of mortality seem to be too crude;
• The local linear method has shown to be a suitable alternative for estimating the force of mortality and the multiplicative correction seems to improve it; and

• It seems that the exponential increase of mortality continues even in old age, because the local linear methods results in an estimated force of mortality increasing more than the linear hazard of the modified Gompertz-Makeham. This is also indicated by the fact that the remaining life expectancies based on the local linear estimates are lower than the ones based on the modified Gompertz-Makeham results.

An advantage of the local linear approximation is that no iterative method for estimation is needed. It is a completely nonparametric approach that does not require a fixed parametric shape. On the other hand, it also has some disadvantages: extrapolation to higher ages is not possible, and the nonparametric method has to cope with bias at the end points. When looking at our data set, our methods suggest that life expectancy is lower than the one obtained with the parametric approach, especially for ages above 100.

References


Estimation of Complete Period Life Tables for Singaporeans

Siu-Hang Li* and Wai-Sum Chan†

Abstract‡

Complete period life tables, with death rates for every year of age, are not available in Singapore. This study constructs such tables for Singaporeans from the limited mortality information contained in the abridged life tables provided by the Singapore Department of Statistics. We find that linear interpolation, Whittaker graduation, and the Coale-Kisker method together can generate complete life tables that are smooth and continuous. The validity of the complete life tables generated by our method is further confirmed by (1) comparing the life expectancies calculated from our estimated life tables with those provided by the Singapore Department of Statistics, and (2) comparing the shapes of the mortality functions derived from our life tables with those derived from the Commissioner's Valuation Tables for assured lives in Singapore.

Key words and phrases: graduation, data-disaggregation, Whittaker-Henderson method, Coale-Kisker method

*Siuh-Hang Li, M.Phil. Candidate, is a graduate student in statistics and actuarial science in the University of Hong Kong. His research area include retirement income security analysis, mortality studies, and the applications of actuarial techniques in law courts.

Mr. Li's address is: Department of Statistics and Actuarial Science, University of Hong Kong, HONG KONG. E-mail: h9902614@hkusua.hku.hk

†Wai-Sum Chan, Ph.D., F.S.A., is an associate professor of statistics and actuarial science in the University of Hong Kong. He holds a Ph.D. in Statistics from Temple University. His research interests include the modeling of actuarial assumptions, actuarial evidence in courts, health care financing, and long-term care insurance modeling.

Dr. Chan's address is: Department of Statistics and Actuarial Science, University of Hong Kong, HONG KONG. E-mail: chanws@hku.hk

‡The authors are indebted to the anonymous referees for their helpful comments on an earlier version of the article. The authors also wish to acknowledge and express their appreciation for the support of a research grant from the Research Grants Council of the Hong Kong Special Administrative Region (Competitive Earmarked Research Grant Project No. HKU 7111/01H). The abridged life table data used in this paper were kindly provided by the Singapore Department of Statistics and were used with their permission.
1 Introduction

Period life tables (also known as current life tables) represent the mortality profile of a community at a specific point of time. These tables show the conditional probabilities of death, denoted as $q_x$, for every single year of age from 0 to a maximum age such as 100 or 110 and can be varied by such factors as gender and race. In the United States, complete period life tables, by age, sex, and race are published from time to time by the National Center of Health Statistics (NCHS). In the United Kingdom, similar complete period life tables, known as English Life Tables, are prepared by the Government Actuary, based on the mortality experience of the general population in England and Wales. Complete period life tables, however, are not necessarily available in some developed countries, such as Singapore.

To glimpse the mortality pattern in Singapore, one can refer to two series of available tables. The first series is known as the Commissioner’s Valuation Table (CVT), developed by the Monetary Authority of Singapore (Singapore’s central bank), based on the mortality experience of the insured population. The CVT has been widely used in the insurance industry for the purpose of valuation. Regulation 26(5) of the Insurance Regulations 1992 in Singapore specifies that in respect to a policy other than an annuity, the minimum reserves for the valuation shall be made by using the 1992 Commissioner’s Valuation Table for male lives and the 1992 Commissioner’s Valuation Table with a three-year age setback for female lives. Similar to the English Life Tables and the U.S. life tables, the CVTs are complete in the sense that the conditional probabilities of death are shown for every single year of age. Despite their completeness, the use of CVTs is limited to the purpose of insurance valuation. At a specific point of time, the insured and general populations could share a similar shape of mortality profile, but the level of mortality for the insured is typically lower than that for the general population, mainly due to the underwriting procedures employed prior to issuing insurance policies. Using the CVT for mortality analyses of the general population is, therefore, inappropriate.

The second available series is the set of official life tables provided by the Singapore Department of Statistics. These official life tables are constructed based on the mortality experience of the entire Singaporean population and, therefore, applicable in mortality analyses of the general population. Unfortunately, for the purpose of easing workload and expenses, the mortality data in the official life tables are presented in an abridged form. Instead of every single year of age, values of the central death rates are shown at age 0, age group 1–4, quinquennial
age groups 5–9, 10–14, and so on up to 65–69, and the open age group of 70 and over. Abridged life tables can allow the comparison of the mortality experiences of different territories and the longitudinal study of mortality levels of a population. In many actuarial and demographic applications, however, abridged life tables do not suffice.

In actuarial practice, complete life tables are always required for the computation of monetary functions involving life contingencies, and the absence of such tables often leads to problems. For example, Chen and Wong (1997) examined the adequacy of the benefits from the Central Provident Fund (CPF) savings in Singapore using actuarial simulations. They computed the monthly benefit to be received by a retiree by assuming that the accumulated value of the retiree's CPF savings was to be progressively liquidated by a life annuity. As no complete life table for the general Singaporean population existed at that time, they used an annuity certain with term equal to the expected future lifetime at the retirement age as a proxy for the value of the whole life annuity. It can be easily proved that the value of such an annuity certain is always less than that of its corresponding life annuity. In other words, all of the estimated monthly benefits in the study were overstated.

Another example occurs in the area of the assessment of future financial loss in personal injury litigation. In such cases courts often adopt the multiplicand and multiplier approach to determine the lump sum amount of compensation for the plaintiff for future loss of earnings and to cover consequential expenses. In the first stage of this approach, a multiplicand, which represents the future annual loss of income and consequential expenses, is decided. In the second stage, the multiplicand is multiplied by an appropriate multiplier, decided by the judge, to get the lump sum amount of compensation for the plaintiff. In the United Kingdom, tables of actuarially calculated multipliers, known as 'Ogden Tables', have been available since 1984. Sarony et al. (2003) also developed tables of multipliers for Hong Kong. The calculation of these actuarial multipliers, similar to the case for life annuities, requires the probabilities of death for all ages. Thus, tables of these multipliers are not available in Singapore at present.

The absence of appropriate complete period life tables for the general population inevitably handicaps demographers in Singapore and partly explains why mortality studies in Singapore have been scanty. In the implementation of the well-known Lee-Carter method of mortality forecasting, Lee and Carter (1992) pointed out that under the circumstances of population aging and demographic shifts, failing to consider the mortality distribution of the open age group would be uninformative and might even lead to serious distortion. This calls for life tables
with more information in the open age group, preferably with death
rates shown for every single year of age.

In this study, we develop a method to produce complete period life
tables with $q_x$'s for every single year of age from 0 to 99 from the lim-
ited mortality information shown in the abridged life tables published
by the Singapore Department of Statistics using the ideas of interpola-
tion, extrapolation, and graduation. To ensure that the estimated life
tables are of high quality, the estimation method must consistently ful-
fill several criteria:

- First, the complete life expectancy at birth, denoted as $e_0$, cal-
culated from our estimated complete period life table should be
close to that officially publicized by the Singapore Department of
Statistics. An absolute difference greater than one is regarded as
unsatisfactory. In a statistical sense, this is to match the first mo-
moment derived from our estimated probabilities of death with the
true first moment.

- Second, the mortality profiles described by our estimated life ta-
bles should share the properties of the typical laws of mortality—
the probability of death should decrease from birth to a minimum
value at around age 10 and then monotonically increase gradually.

- Third, the shapes of the plots of the mortality functions derived
from our estimated life tables should be similar to those derived
from the CVTs. In particular, the curves of the logarithms of the
probabilities of death are compared.

2 The Estimation Procedure

The Singapore Department of Statistics' abridged life tables con-
tain age-specific death rates for age 0, age group 1–4, quinquennial age
groups 5–9, 10–14, etc., up to 65–69, and the open age group of 70 and
over (Table 1). The so-called age-specific death rates here actually refer
to the number of resident deaths per thousand residents of a specific
age group during a given year.

Our estimation procedure is divided into three stages: (i) deriving
the initial (raw) estimates using interpolative methods; (ii) graduation of
these raw rates to produce a smooth and more reliable representation
of the underlying pattern of mortality; and (iii) extrapolation of the rates
to individual ages 70 and over.
Table 1
Abridged Life Table
For Singaporeans (2001)
Age-Specific Death Rates

<table>
<thead>
<tr>
<th>Age</th>
<th>1000 x $nM_x$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
</tr>
<tr>
<td>0</td>
<td>2.4</td>
</tr>
<tr>
<td>1-4</td>
<td>0.3</td>
</tr>
<tr>
<td>5-9</td>
<td>0.1</td>
</tr>
<tr>
<td>10-14</td>
<td>0.1</td>
</tr>
<tr>
<td>15-19</td>
<td>0.4</td>
</tr>
<tr>
<td>20-24</td>
<td>0.7</td>
</tr>
<tr>
<td>25-29</td>
<td>0.7</td>
</tr>
<tr>
<td>30-34</td>
<td>0.7</td>
</tr>
<tr>
<td>35-39</td>
<td>1.0</td>
</tr>
<tr>
<td>40-44</td>
<td>1.6</td>
</tr>
<tr>
<td>45-49</td>
<td>2.5</td>
</tr>
<tr>
<td>50-54</td>
<td>4.6</td>
</tr>
<tr>
<td>55-59</td>
<td>8.1</td>
</tr>
<tr>
<td>60-64</td>
<td>13.2</td>
</tr>
<tr>
<td>65-69</td>
<td>23.2</td>
</tr>
<tr>
<td>70+</td>
<td>58.3</td>
</tr>
</tbody>
</table>

2.1 Stage One: Initial Estimation

To obtain an initial estimate of the $q_x$'s in the complete life tables, the central death rate for every single year of age ($m_x$) is used through an interpolation method. Let $nM_x$ be the crude central death rate for persons in the integer age group $x$ to $x + n - 1$ shown in the abridged life tables, and let $m_x$ be the unknown single year central death rate for persons age $x$ last birthday. The quantity $m_x$ is to be obtained in this stage of the estimation.

Before performing the interpolation exercise, we treat $nM_x$ as the single year central death rate at the mid-point of its corresponding age group. That is, we set $m_7$ as $sM_5$, $m_{12}$ as $sM_{10}$, and so on. For the open age group, we assume that it closes at age 100 and, hence, $m_{83}$ is taken as the central death rate for the open age group 70 and over shown on the abridged life tables. The infant mortality rate, $m_0$, is given exactly
as \(_{1}M_{0}\) in the abridged life tables; and \(_{4}m_{1}\) is taken as \(_{4}M_{1}\) without any interpolation (to avoid unwanted inflation due to the domination of \(_{0}m_{0}\)). The remaining values of the single year central death rates are determined using linear interpolation. Mathematically, we have

\[
m_{x} = \begin{cases} 
\frac{1}{2} (5M_5 - 4M_1)(x - 2) + 4M_1 & \text{for } x = 2, 3, 4, 5, 6; \\
\frac{1}{2} (5M_{k+3} - 5M_{k-2})(x - k) + 5M_{k-2} & \text{for } x = k, \ldots, k + 4 \text{ and } k = 7, 12, 17, 22, \ldots, 62, \\
\frac{1}{16} (30M_{70} - 5M_{65})(x - 67) + 5M_{65} & \text{for } x = 67, 68, \ldots, 82, \\
\frac{1}{17} (1 - 30M_{70})(x - 83) + 30M_{70} & \text{for } x = 83, 84, \ldots, 100.
\end{cases}
\]

The initial estimates of \(_{x}m_{x}\) for males and females are given in Tables 2 and 3.

2.2 Stage Two: Graduation

The next step is to graduate the initial estimates to ensure they conform to the generally accepted beliefs about the underlying mortality rates: they are smooth, regular, and continuous.\(^1\)

Graduation techniques have been extensively employed in dealing with the problem of raggedness in the initial estimates derived from mortality data. (See, for example, Renshaw et al., 1996; Renshaw, 1995; Renshaw and Hatzopoulos, 1995.) Though there are several methods of graduation, the two basic ones are:

1. Parametric gradation, which is performed by fitting a parametric formula (law of mortality) over the entire age range. An example is the Heligman and Pollard model (Heligman and Pollard, 1980); and

2. Non-parametric graduation—examples include moving weighted average graduation (Ramsay, 1991) and Whittaker graduation (Whittaker, 1923).

London (1985) provides a thorough discussion of the various types of graduation methods used by actuaries.

Among the various graduation techniques, Whittaker gradation (also known as Whittaker-Henderson graduation) performs well in terms of

\(^1\)Miller (1946) stated that capricious irregularities in the tables from age to age, disturbing the orderly progression of premiums, etc., would be inconsistent with the common sense view that such figures should be reasonably regular and would tend to arouse an entirely justifiable skepticism.
our described estimation criteria. Whittaker graduation allows practitioners to customize the tradeoff between goodness of fit and smoothness and has been extensively employed. (See Guerrero et al., 2001 and Lowrie, 1993.) The objective of Whittaker graduation is to choose the \( \{m_x^\theta\} \) sequence that minimizes the equation

\[
J = F + hS \equiv \sum_{x=0}^{n} w_x (m_x^\theta - m_x)^2 + h \sum_{x=0}^{n-z} (\Delta^z m_x^\theta)^2
\]

where \( n \) is the highest age used, \( m_x \) is the initial estimate of the central death rate at age \( x \), \( m_x^\theta \) denotes the graduated value of the central death rate at age \( x \), \( w_x \) is a positive weight imposed at age \( x \), \( h \) is a nonnegative parameter, \( \Delta \) is the forward difference operator such that \( \Delta m_x^\theta = m_x^\theta - m_{x-1}^\theta \), and \( z \) is the number of times the difference operator is applied. The quantity \( F \equiv \sum_{x=1}^{n} w_x (m_x^\theta - m_x)^2 \) is a measure of fitness, \( S \equiv \sum_{x=1}^{n-z} (\Delta^z m_x^\theta)^2 \) is a measure of smoothness, and \( h \) reflects the relative importance of smoothness over fit. By minimizing \( F \), \( \{m_x^\theta\} \) will tend to be close to \( \{m_x\} \), resulting in a better goodness of fit. By minimizing \( S \), \( \{m_x^\theta\} \) will tend to lie on an adaptive polynomial of degree \( z \), resulting in a higher degree of smoothness. There is no optimal solution for \( z \) and \( h \). They are determined subjectively by the practitioner.

In our estimation procedure, we use the special case of the Whittaker graduation, where \( w_x = 1 \) for all \( x \). By trial and error, \( z \) is fixed to be 2, and the smoothing parameter, \( h \), is chosen to be 0.5 for balancing goodness of fit and smoothness. Small changes in \( z \) or \( h \) have only a negligible effect on the overall result. The graduation is applied up to \( x = 69 \) (i.e., \( n = 69 \)). For \( x > 69 \) (the open age group), central death rates are computed using an extrapolative procedure to be described in the next stage. Rewriting \( J \) in matrix form with the pre-determined values of \( w_x, z \) and \( h \) yields

\[
J = (m^\theta - m)^T (m^\theta - m) + \frac{1}{2} m^\theta' k_2 k_2' m^\theta
\]

where \( m^\theta = (m_1^\theta, \ldots, m_n^\theta)^T, m = (m_1, \ldots, m_n)^T \) and \( k_2 \) is an \( (n-2) \times n \) matrix such that

\[
k_2(i, j) = (-1)^{2+i-j} \left[ \frac{2!}{(j-i)! (2-j+i)!} \right]
\]

for \( i = 1, \ldots, n-2, j = 1, \ldots, n \) with \( k_2(i, j) = 0 \) for \( j < i \) or \( j > 2 + i \).

Note that \( J \) is minimized when
Figure 1: Initial Estimates and Graduated Values of $-\ln(m_x)$, Male 2001

\[ m^\theta = (I + 0.5k_2'k_2)^{-1}m, \]

where $I$ is an $n \times n$ identity matrix. A proof can be found in London (1985, page 57).

Tables 2 and 3 show the initial and graduated estimates of the central death rates obtained in this stage for males and females, respectively. Figure 1 displays the plot of $-\ln(m_x)$ for both the initial and graduated estimates. The notice that the graduated estimates show an improvement in smoothness without much deviation from the initial values. The shape of the plot agrees reasonably with our prior assumption of the underlying survival model.

### 2.3 Stage Three: Extrapolation

Table 4 shows the preliminary values of the single year central death rates obtained in the initial stage. Notice that at advanced ages, especially at ages over 90, the values are unexpectedly high. The disappointing performance of the interpolation at such advanced ages is in large part due to the lack of mortality information: only a single rate is provided for the open age group of 70 and over in the abridged life
<table>
<thead>
<tr>
<th>Age</th>
<th>Initial Estimate</th>
<th>Graduated</th>
<th>Age</th>
<th>Initial Estimate</th>
<th>Graduated</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.002400</td>
<td>0.002400</td>
<td>35</td>
<td>0.000880</td>
<td>0.000875</td>
</tr>
<tr>
<td>1</td>
<td>0.000300</td>
<td>0.000307</td>
<td>36</td>
<td>0.000940</td>
<td>0.000938</td>
</tr>
<tr>
<td>2</td>
<td>0.000300</td>
<td>0.000290</td>
<td>37</td>
<td>0.001000</td>
<td>0.001014</td>
</tr>
<tr>
<td>3</td>
<td>0.000260</td>
<td>0.000260</td>
<td>38</td>
<td>0.001120</td>
<td>0.001118</td>
</tr>
<tr>
<td>4</td>
<td>0.000220</td>
<td>0.000221</td>
<td>39</td>
<td>0.001240</td>
<td>0.001235</td>
</tr>
<tr>
<td>5</td>
<td>0.000180</td>
<td>0.000179</td>
<td>40</td>
<td>0.001360</td>
<td>0.001355</td>
</tr>
<tr>
<td>6</td>
<td>0.000140</td>
<td>0.000139</td>
<td>41</td>
<td>0.001480</td>
<td>0.001478</td>
</tr>
<tr>
<td>7</td>
<td>0.000100</td>
<td>0.000109</td>
<td>42</td>
<td>0.001600</td>
<td>0.001614</td>
</tr>
<tr>
<td>8</td>
<td>0.000100</td>
<td>0.000098</td>
<td>43</td>
<td>0.001780</td>
<td>0.001777</td>
</tr>
<tr>
<td>9</td>
<td>0.000100</td>
<td>0.000096</td>
<td>44</td>
<td>0.001960</td>
<td>0.001950</td>
</tr>
<tr>
<td>10</td>
<td>0.000100</td>
<td>0.000095</td>
<td>45</td>
<td>0.002140</td>
<td>0.002125</td>
</tr>
<tr>
<td>11</td>
<td>0.000100</td>
<td>0.000098</td>
<td>46</td>
<td>0.002320</td>
<td>0.002313</td>
</tr>
<tr>
<td>12</td>
<td>0.000100</td>
<td>0.000114</td>
<td>47</td>
<td>0.002500</td>
<td>0.002557</td>
</tr>
<tr>
<td>13</td>
<td>0.000160</td>
<td>0.000158</td>
<td>48</td>
<td>0.002920</td>
<td>0.002911</td>
</tr>
<tr>
<td>14</td>
<td>0.000220</td>
<td>0.000217</td>
<td>49</td>
<td>0.003340</td>
<td>0.003318</td>
</tr>
<tr>
<td>15</td>
<td>0.000280</td>
<td>0.000278</td>
<td>50</td>
<td>0.003760</td>
<td>0.003737</td>
</tr>
<tr>
<td>16</td>
<td>0.000340</td>
<td>0.000339</td>
<td>51</td>
<td>0.004180</td>
<td>0.004171</td>
</tr>
<tr>
<td>17</td>
<td>0.000400</td>
<td>0.000400</td>
<td>52</td>
<td>0.004600</td>
<td>0.004666</td>
</tr>
<tr>
<td>18</td>
<td>0.000460</td>
<td>0.000461</td>
<td>53</td>
<td>0.005300</td>
<td>0.005290</td>
</tr>
<tr>
<td>19</td>
<td>0.000520</td>
<td>0.000522</td>
<td>54</td>
<td>0.006000</td>
<td>0.005975</td>
</tr>
<tr>
<td>20</td>
<td>0.000580</td>
<td>0.000583</td>
<td>55</td>
<td>0.006700</td>
<td>0.006675</td>
</tr>
<tr>
<td>21</td>
<td>0.000640</td>
<td>0.000642</td>
<td>56</td>
<td>0.007400</td>
<td>0.007391</td>
</tr>
<tr>
<td>22</td>
<td>0.000700</td>
<td>0.000686</td>
<td>57</td>
<td>0.008100</td>
<td>0.008176</td>
</tr>
<tr>
<td>23</td>
<td>0.000700</td>
<td>0.000702</td>
<td>58</td>
<td>0.009120</td>
<td>0.009105</td>
</tr>
<tr>
<td>24</td>
<td>0.000700</td>
<td>0.000703</td>
<td>59</td>
<td>0.010140</td>
<td>0.010094</td>
</tr>
<tr>
<td>25</td>
<td>0.000700</td>
<td>0.000702</td>
<td>60</td>
<td>0.011160</td>
<td>0.011095</td>
</tr>
<tr>
<td>26</td>
<td>0.000700</td>
<td>0.000701</td>
<td>61</td>
<td>0.012180</td>
<td>0.012150</td>
</tr>
<tr>
<td>27</td>
<td>0.000700</td>
<td>0.000700</td>
<td>62</td>
<td>0.013200</td>
<td>0.013431</td>
</tr>
<tr>
<td>28</td>
<td>0.000700</td>
<td>0.000699</td>
<td>63</td>
<td>0.015200</td>
<td>0.015171</td>
</tr>
<tr>
<td>29</td>
<td>0.000700</td>
<td>0.000698</td>
<td>64</td>
<td>0.017200</td>
<td>0.017139</td>
</tr>
<tr>
<td>30</td>
<td>0.000700</td>
<td>0.000697</td>
<td>65</td>
<td>0.019200</td>
<td>0.019161</td>
</tr>
<tr>
<td>31</td>
<td>0.000700</td>
<td>0.000698</td>
<td>66</td>
<td>0.021200</td>
<td>0.021188</td>
</tr>
<tr>
<td>32</td>
<td>0.000700</td>
<td>0.000714</td>
<td>67</td>
<td>0.023200</td>
<td>0.023248</td>
</tr>
<tr>
<td>33</td>
<td>0.000760</td>
<td>0.000758</td>
<td>68</td>
<td>0.025394</td>
<td>0.025390</td>
</tr>
<tr>
<td>34</td>
<td>0.000820</td>
<td>0.000815</td>
<td>69</td>
<td>0.027588</td>
<td>0.027569</td>
</tr>
</tbody>
</table>
# Table 3
Initial Estimates and Graduated Values of $m_x$, Female 2001

<table>
<thead>
<tr>
<th>Age</th>
<th>Initial Estimate</th>
<th>Graduated</th>
<th>Age</th>
<th>Initial Estimate</th>
<th>Graduated</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.002100</td>
<td>0.002100</td>
<td>35</td>
<td>0.000560</td>
<td>0.000559</td>
</tr>
<tr>
<td>1</td>
<td>0.000300</td>
<td>0.000307</td>
<td>36</td>
<td>0.000580</td>
<td>0.000579</td>
</tr>
<tr>
<td>2</td>
<td>0.000300</td>
<td>0.000290</td>
<td>37</td>
<td>0.000600</td>
<td>0.000609</td>
</tr>
<tr>
<td>3</td>
<td>0.000260</td>
<td>0.000260</td>
<td>38</td>
<td>0.000660</td>
<td>0.000658</td>
</tr>
<tr>
<td>4</td>
<td>0.000220</td>
<td>0.000221</td>
<td>39</td>
<td>0.000720</td>
<td>0.000716</td>
</tr>
<tr>
<td>5</td>
<td>0.000180</td>
<td>0.000179</td>
<td>40</td>
<td>0.000780</td>
<td>0.000775</td>
</tr>
<tr>
<td>6</td>
<td>0.000140</td>
<td>0.000139</td>
<td>41</td>
<td>0.000840</td>
<td>0.000838</td>
</tr>
<tr>
<td>7</td>
<td>0.000100</td>
<td>0.000109</td>
<td>42</td>
<td>0.000900</td>
<td>0.000914</td>
</tr>
<tr>
<td>8</td>
<td>0.000100</td>
<td>0.000099</td>
<td>43</td>
<td>0.001020</td>
<td>0.001018</td>
</tr>
<tr>
<td>9</td>
<td>0.000100</td>
<td>0.000097</td>
<td>44</td>
<td>0.001140</td>
<td>0.001134</td>
</tr>
<tr>
<td>10</td>
<td>0.000100</td>
<td>0.000097</td>
<td>45</td>
<td>0.001260</td>
<td>0.001253</td>
</tr>
<tr>
<td>11</td>
<td>0.000100</td>
<td>0.000098</td>
<td>46</td>
<td>0.001380</td>
<td>0.001377</td>
</tr>
<tr>
<td>12</td>
<td>0.000100</td>
<td>0.000109</td>
<td>47</td>
<td>0.001500</td>
<td>0.001524</td>
</tr>
<tr>
<td>13</td>
<td>0.000140</td>
<td>0.000139</td>
<td>48</td>
<td>0.001720</td>
<td>0.001716</td>
</tr>
<tr>
<td>14</td>
<td>0.000180</td>
<td>0.000179</td>
<td>49</td>
<td>0.001940</td>
<td>0.001929</td>
</tr>
<tr>
<td>15</td>
<td>0.000220</td>
<td>0.000222</td>
<td>50</td>
<td>0.002160</td>
<td>0.002147</td>
</tr>
<tr>
<td>16</td>
<td>0.000260</td>
<td>0.000261</td>
<td>51</td>
<td>0.002380</td>
<td>0.002374</td>
</tr>
<tr>
<td>17</td>
<td>0.000300</td>
<td>0.000286</td>
<td>52</td>
<td>0.002600</td>
<td>0.002643</td>
</tr>
<tr>
<td>18</td>
<td>0.000280</td>
<td>0.000282</td>
<td>53</td>
<td>0.003000</td>
<td>0.002994</td>
</tr>
<tr>
<td>19</td>
<td>0.000260</td>
<td>0.000263</td>
<td>54</td>
<td>0.003400</td>
<td>0.003386</td>
</tr>
<tr>
<td>20</td>
<td>0.000240</td>
<td>0.000241</td>
<td>55</td>
<td>0.003800</td>
<td>0.003789</td>
</tr>
<tr>
<td>21</td>
<td>0.000220</td>
<td>0.000220</td>
<td>56</td>
<td>0.004200</td>
<td>0.004196</td>
</tr>
<tr>
<td>22</td>
<td>0.000200</td>
<td>0.000205</td>
<td>57</td>
<td>0.004600</td>
<td>0.004629</td>
</tr>
<tr>
<td>23</td>
<td>0.000200</td>
<td>0.000199</td>
<td>58</td>
<td>0.005120</td>
<td>0.005113</td>
</tr>
<tr>
<td>24</td>
<td>0.000200</td>
<td>0.000197</td>
<td>59</td>
<td>0.005640</td>
<td>0.005616</td>
</tr>
<tr>
<td>25</td>
<td>0.000200</td>
<td>0.000196</td>
<td>60</td>
<td>0.006160</td>
<td>0.006123</td>
</tr>
<tr>
<td>26</td>
<td>0.000200</td>
<td>0.000198</td>
<td>61</td>
<td>0.006680</td>
<td>0.006664</td>
</tr>
<tr>
<td>27</td>
<td>0.000200</td>
<td>0.000214</td>
<td>62</td>
<td>0.007200</td>
<td>0.007342</td>
</tr>
<tr>
<td>28</td>
<td>0.000260</td>
<td>0.000259</td>
<td>63</td>
<td>0.008320</td>
<td>0.008296</td>
</tr>
<tr>
<td>29</td>
<td>0.000320</td>
<td>0.000318</td>
<td>64</td>
<td>0.009440</td>
<td>0.009376</td>
</tr>
<tr>
<td>30</td>
<td>0.000380</td>
<td>0.000381</td>
<td>65</td>
<td>0.010560</td>
<td>0.010484</td>
</tr>
<tr>
<td>31</td>
<td>0.000440</td>
<td>0.000441</td>
<td>66</td>
<td>0.011680</td>
<td>0.011650</td>
</tr>
<tr>
<td>32</td>
<td>0.000500</td>
<td>0.000491</td>
<td>67</td>
<td>0.012800</td>
<td>0.013055</td>
</tr>
<tr>
<td>33</td>
<td>0.000520</td>
<td>0.000521</td>
<td>68</td>
<td>0.014969</td>
<td>0.014939</td>
</tr>
<tr>
<td>34</td>
<td>0.000540</td>
<td>0.000541</td>
<td>69</td>
<td>0.017138</td>
<td>0.017033</td>
</tr>
</tbody>
</table>
tables. Thus, we must re-estimate the single year central death rate for ages over 69 through an extrapolative model that realistically reflects the underlying law of mortality at very advanced ages. The classical Gompertz curve is not an ideal choice, as empirical evidence indicates that mortality at older ages is not Gompertzian (Olshansky and Carnes, 1997). Coale and Kisker (1990) showed that, for a number of developed countries, mortality rates increase at a linearly decreasing rate, rather than at a constant rate as the Gompertz curve assumes. We apply the method suggested by Coale and Guo (1989), namely the Coale-Kisker method, which assumes a linearly decreasing rate, to extend the death rates up to age 99, based on the graduated values of $m_x$ around age 70 obtained in the second stage of our estimation.

For $x \geq 70$, define

$$k(x) = k(x - 1) - R,$$

where $k(x) = \ln(m_x/m_{x-1})$. Extending the formula up to $x = 110$ and summing, we get

$$k(70) + \cdots + k(110) = \ln(m_{110}/m_{69}) = 41k(69) - 861R.$$

Solving for $R$, we obtain

$$R = \frac{41k(69) + \ln(m_{69}) - \ln(m_{110})}{861}. \quad (4)$$

To minimize the effects of random fluctuations, $k(69)$ is replaced by $k^*(69)$, which is a moving average of $k(67)$ to $k(71)$, i.e.,

\begin{table}
\centering
\caption{Initial Estimate of $m_x$ at Advanced Ages, 2001}
\begin{tabular}{lll}
\hline
Age & Male & Female \\
\hline
70 & 0.01920 & 0.01056 \\
75 & 0.02978 & 0.01931 \\
80 & 0.04075 & 0.03015 \\
85 & 0.05172 & 0.04099 \\
90 & 0.16909 & 0.15956 \\
95 & 0.44606 & 0.43971 \\
\hline
\end{tabular}
\end{table}
\[
k^*(69) = \frac{k(67) + k(68) + k(69) + k(70) + k(71)}{5} = \frac{k(67) + k(68) + 3k(69) - 3R}{5}.
\]

Similarly, \(\ln(m_{69})\) is replaced by \(\ln(m^*_{69})\), which is defined as
\[
\ln(m^*_{69}) = k^*(69) + \ln(m^*_{68}),
\]
where
\[
\ln(m^*_{68}) = \ln\left(\frac{m_{67} + m_{68} + m_{69}}{3}\right).
\]
Substituting \(k^*(69)\) and \(\ln(m^*_{69})\) in \(R\), we obtain
\[
R = \frac{42}{5} \left[ k(67) + k(68) + k(69) \right] + \ln\left(\frac{m_{67} + m_{68} + m_{69}}{3m_{110}}\right). \tag{5}
\]

According to the original Coale-Kisker method, \(m_{110}\) is assumed to be 1.0 for males. Such an assumption is based on the fact that there are almost no deaths occurring at ages greater than 110. For females, \(m_{110}\) is chosen as 0.8 to avoid imposing a crossover of male and female mortality at age 110. Coale and Kisker (1990) demonstrated that the choice of 1.0 or 0.8 for \(m_{110}\) has only a minor effect on the constructed life table for white males; we found that it is also true for the Singaporean life table. The values of \(m_{67}, m_{68},\) and \(m_{69}\) used are the graduated values of the central death rates obtained in the second stage of our estimation.

3 Results

Finally we estimate the probability of death, \(q_x\), using the assumption of uniform distribution of death, as (Bowers et al., 1997, page 90):
\[
q_x = \frac{2m_x}{2 + m_x}. \tag{6}
\]
Tables 5 and 6 show an example of a complete period life table estimated for males and females by our suggested procedures. The mortality profiles that are described by our estimated life tables share the properties of the typical laws of mortality: mortality rates drop from a relatively large value at birth to a minimum value at around age 10 and then rises in a monotonic, continuous, and smooth manner. Table 7 presents a longitudinal comparison between the life expectancies at birth derived from our complete life table and those officially provided by the Singapore Department of Statistics. We could observe no systematic discrepancies for both sexes. The average difference between the two values is 0.26 and the maximum absolute difference is strictly less than one, showing that the criterion of first moment matching is satisfied.

Figures 2 compares the plots of \(-\log(q_x)\) arising from our estimated complete life table with those from the Commissioner's Valuation Table (CVT) provided by the Monetary Authority of Singapore. Before age 50, their shapes are close to each other, but a parallel shift exists: the profile for the CVT is higher than that for our life table, reflecting the fact that the mortality level for assured lives is generally lower. After age 50, the curves coincide. The consistency further ensures that our complete life tables are representative of the mortality pattern of the general Singaporean population.

4 Concluding Remarks and Future Research

This paper proposes a simple method for estimating complete period life tables for Singaporeans from the abridged life tables that are provided by the Singapore Department of Statistics. We found that linear interpolation, supplemented with Whittaker graduation and Coale-Kisker extrapolation, could generate complete period life tables that are smooth and continuous and reasonably agree with the presumed underlying law of mortality. The reliability of our complete life tables is further substantiated by their consistency with the Commissioner's Valuation Tables and their agreement with the actual life expectancies.

\(^2\)Complete period life tables from 1981 to 2001 for both sexes are available on request.
<table>
<thead>
<tr>
<th>Age</th>
<th>$q_x$</th>
<th>Age</th>
<th>$q_x$</th>
<th>Age</th>
<th>$q_x$</th>
<th>Age</th>
<th>$q_x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0024</td>
<td>26</td>
<td>0.0007</td>
<td>51</td>
<td>0.0042</td>
<td>76</td>
<td>0.0480</td>
</tr>
<tr>
<td>1</td>
<td>0.0003</td>
<td>27</td>
<td>0.0007</td>
<td>52</td>
<td>0.0047</td>
<td>77</td>
<td>0.0521</td>
</tr>
<tr>
<td>2</td>
<td>0.0003</td>
<td>28</td>
<td>0.0007</td>
<td>53</td>
<td>0.0053</td>
<td>78</td>
<td>0.0564</td>
</tr>
<tr>
<td>3</td>
<td>0.0002</td>
<td>29</td>
<td>0.0007</td>
<td>54</td>
<td>0.0060</td>
<td>79</td>
<td>0.0612</td>
</tr>
<tr>
<td>4</td>
<td>0.0002</td>
<td>30</td>
<td>0.0007</td>
<td>55</td>
<td>0.0067</td>
<td>80</td>
<td>0.0663</td>
</tr>
<tr>
<td>5</td>
<td>0.0001</td>
<td>31</td>
<td>0.0007</td>
<td>56</td>
<td>0.0074</td>
<td>81</td>
<td>0.0719</td>
</tr>
<tr>
<td>6</td>
<td>0.0001</td>
<td>32</td>
<td>0.0007</td>
<td>57</td>
<td>0.0081</td>
<td>82</td>
<td>0.0779</td>
</tr>
<tr>
<td>7</td>
<td>0.0001</td>
<td>33</td>
<td>0.0008</td>
<td>58</td>
<td>0.0091</td>
<td>83</td>
<td>0.0844</td>
</tr>
<tr>
<td>8</td>
<td>0.0001</td>
<td>34</td>
<td>0.0008</td>
<td>59</td>
<td>0.0100</td>
<td>84</td>
<td>0.0914</td>
</tr>
<tr>
<td>9</td>
<td>0.0001</td>
<td>35</td>
<td>0.0009</td>
<td>60</td>
<td>0.0110</td>
<td>85</td>
<td>0.0990</td>
</tr>
<tr>
<td>10</td>
<td>0.0001</td>
<td>36</td>
<td>0.0009</td>
<td>61</td>
<td>0.0121</td>
<td>86</td>
<td>0.1072</td>
</tr>
<tr>
<td>11</td>
<td>0.0001</td>
<td>37</td>
<td>0.0010</td>
<td>62</td>
<td>0.0133</td>
<td>87</td>
<td>0.1161</td>
</tr>
<tr>
<td>12</td>
<td>0.0001</td>
<td>38</td>
<td>0.0011</td>
<td>63</td>
<td>0.0151</td>
<td>88</td>
<td>0.1256</td>
</tr>
<tr>
<td>13</td>
<td>0.0002</td>
<td>39</td>
<td>0.0012</td>
<td>64</td>
<td>0.0170</td>
<td>89</td>
<td>0.1359</td>
</tr>
<tr>
<td>14</td>
<td>0.0002</td>
<td>40</td>
<td>0.0014</td>
<td>65</td>
<td>0.0190</td>
<td>90</td>
<td>0.1469</td>
</tr>
<tr>
<td>15</td>
<td>0.0003</td>
<td>41</td>
<td>0.0015</td>
<td>66</td>
<td>0.0210</td>
<td>91</td>
<td>0.1588</td>
</tr>
<tr>
<td>16</td>
<td>0.0003</td>
<td>42</td>
<td>0.0016</td>
<td>67</td>
<td>0.0230</td>
<td>92</td>
<td>0.1716</td>
</tr>
<tr>
<td>17</td>
<td>0.0004</td>
<td>43</td>
<td>0.0018</td>
<td>68</td>
<td>0.0251</td>
<td>93</td>
<td>0.1853</td>
</tr>
<tr>
<td>18</td>
<td>0.0005</td>
<td>44</td>
<td>0.0019</td>
<td>69</td>
<td>0.0272</td>
<td>94</td>
<td>0.2001</td>
</tr>
<tr>
<td>19</td>
<td>0.0005</td>
<td>45</td>
<td>0.0021</td>
<td>70</td>
<td>0.0295</td>
<td>95</td>
<td>0.2158</td>
</tr>
<tr>
<td>20</td>
<td>0.0006</td>
<td>46</td>
<td>0.0023</td>
<td>71</td>
<td>0.0320</td>
<td>96</td>
<td>0.2327</td>
</tr>
<tr>
<td>21</td>
<td>0.0006</td>
<td>47</td>
<td>0.0026</td>
<td>72</td>
<td>0.0347</td>
<td>97</td>
<td>0.2507</td>
</tr>
<tr>
<td>22</td>
<td>0.0007</td>
<td>48</td>
<td>0.0029</td>
<td>73</td>
<td>0.0376</td>
<td>98</td>
<td>0.2699</td>
</tr>
<tr>
<td>23</td>
<td>0.0007</td>
<td>49</td>
<td>0.0033</td>
<td>74</td>
<td>0.0408</td>
<td>99</td>
<td>0.2904</td>
</tr>
<tr>
<td>24</td>
<td>0.0007</td>
<td>50</td>
<td>0.0037</td>
<td>75</td>
<td>0.0443</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6
Complete Period Life Table (Female, 2001)

<table>
<thead>
<tr>
<th>Age</th>
<th>$q_x$</th>
<th>Age</th>
<th>$q_x$</th>
<th>Age</th>
<th>$q_x$</th>
<th>Age</th>
<th>$q_x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0021</td>
<td>1</td>
<td>0.0003</td>
<td>26</td>
<td>0.0002</td>
<td>51</td>
<td>0.0024</td>
</tr>
<tr>
<td>1</td>
<td>0.0003</td>
<td>2</td>
<td>0.0003</td>
<td>27</td>
<td>0.0002</td>
<td>52</td>
<td>0.0026</td>
</tr>
<tr>
<td>2</td>
<td>0.0003</td>
<td>3</td>
<td>0.0003</td>
<td>28</td>
<td>0.0003</td>
<td>53</td>
<td>0.0030</td>
</tr>
<tr>
<td>3</td>
<td>0.0002</td>
<td>4</td>
<td>0.0003</td>
<td>29</td>
<td>0.0003</td>
<td>54</td>
<td>0.0034</td>
</tr>
<tr>
<td>4</td>
<td>0.0002</td>
<td>5</td>
<td>0.0004</td>
<td>30</td>
<td>0.0004</td>
<td>55</td>
<td>0.0038</td>
</tr>
<tr>
<td>5</td>
<td>0.0001</td>
<td>6</td>
<td>0.0004</td>
<td>31</td>
<td>0.0004</td>
<td>56</td>
<td>0.0042</td>
</tr>
<tr>
<td>6</td>
<td>0.0001</td>
<td>7</td>
<td>0.0005</td>
<td>32</td>
<td>0.0005</td>
<td>57</td>
<td>0.0046</td>
</tr>
<tr>
<td>7</td>
<td>0.0001</td>
<td>8</td>
<td>0.0005</td>
<td>33</td>
<td>0.0005</td>
<td>58</td>
<td>0.0051</td>
</tr>
<tr>
<td>8</td>
<td>0.0001</td>
<td>9</td>
<td>0.0005</td>
<td>34</td>
<td>0.0006</td>
<td>59</td>
<td>0.0056</td>
</tr>
<tr>
<td>9</td>
<td>0.0001</td>
<td>10</td>
<td>0.0006</td>
<td>35</td>
<td>0.0006</td>
<td>60</td>
<td>0.0061</td>
</tr>
<tr>
<td>10</td>
<td>0.0001</td>
<td>11</td>
<td>0.0006</td>
<td>36</td>
<td>0.0006</td>
<td>61</td>
<td>0.0066</td>
</tr>
<tr>
<td>11</td>
<td>0.0001</td>
<td>12</td>
<td>0.0006</td>
<td>37</td>
<td>0.0006</td>
<td>62</td>
<td>0.0073</td>
</tr>
<tr>
<td>12</td>
<td>0.0001</td>
<td>13</td>
<td>0.0007</td>
<td>38</td>
<td>0.0007</td>
<td>63</td>
<td>0.0083</td>
</tr>
<tr>
<td>13</td>
<td>0.0001</td>
<td>14</td>
<td>0.0007</td>
<td>39</td>
<td>0.0008</td>
<td>64</td>
<td>0.0093</td>
</tr>
<tr>
<td>14</td>
<td>0.0002</td>
<td>15</td>
<td>0.0008</td>
<td>40</td>
<td>0.0007</td>
<td>65</td>
<td>0.0104</td>
</tr>
<tr>
<td>15</td>
<td>0.0002</td>
<td>16</td>
<td>0.0008</td>
<td>41</td>
<td>0.0009</td>
<td>66</td>
<td>0.0116</td>
</tr>
<tr>
<td>16</td>
<td>0.0003</td>
<td>17</td>
<td>0.0009</td>
<td>42</td>
<td>0.0009</td>
<td>67</td>
<td>0.0130</td>
</tr>
<tr>
<td>17</td>
<td>0.0003</td>
<td>18</td>
<td>0.0010</td>
<td>43</td>
<td>0.0010</td>
<td>68</td>
<td>0.0148</td>
</tr>
<tr>
<td>18</td>
<td>0.0003</td>
<td>19</td>
<td>0.0011</td>
<td>44</td>
<td>0.0011</td>
<td>69</td>
<td>0.0169</td>
</tr>
<tr>
<td>19</td>
<td>0.0003</td>
<td>20</td>
<td>0.0013</td>
<td>45</td>
<td>0.0013</td>
<td>70</td>
<td>0.0192</td>
</tr>
<tr>
<td>20</td>
<td>0.0003</td>
<td>21</td>
<td>0.0014</td>
<td>46</td>
<td>0.0014</td>
<td>71</td>
<td>0.0218</td>
</tr>
<tr>
<td>21</td>
<td>0.0002</td>
<td>22</td>
<td>0.0015</td>
<td>47</td>
<td>0.0015</td>
<td>72</td>
<td>0.0247</td>
</tr>
<tr>
<td>22</td>
<td>0.0002</td>
<td>23</td>
<td>0.0017</td>
<td>48</td>
<td>0.0017</td>
<td>73</td>
<td>0.0279</td>
</tr>
<tr>
<td>23</td>
<td>0.0002</td>
<td>24</td>
<td>0.0019</td>
<td>49</td>
<td>0.0019</td>
<td>74</td>
<td>0.0315</td>
</tr>
<tr>
<td>24</td>
<td>0.0002</td>
<td>25</td>
<td>0.0021</td>
<td>50</td>
<td>0.0021</td>
<td>75</td>
<td>0.0355</td>
</tr>
</tbody>
</table>
Table 7
Complete Life Expectancies at Birth for 1988–2001

<table>
<thead>
<tr>
<th>Year</th>
<th>Males</th>
<th></th>
<th></th>
<th>Females</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est. $\hat{e}_0$</td>
<td>$\hat{e}_0$</td>
<td>DIFF</td>
<td>Est. $\hat{e}_0$</td>
<td>$\hat{e}_0$</td>
<td>DIFF</td>
</tr>
<tr>
<td>1988</td>
<td>73.35</td>
<td>72.60</td>
<td>0.75</td>
<td>77.14</td>
<td>76.90</td>
<td>0.24</td>
</tr>
<tr>
<td>1989</td>
<td>73.10</td>
<td>72.90</td>
<td>0.20</td>
<td>77.51</td>
<td>77.20</td>
<td>0.31</td>
</tr>
<tr>
<td>1990</td>
<td>73.83</td>
<td>73.10</td>
<td>0.73</td>
<td>77.90</td>
<td>77.60</td>
<td>0.30</td>
</tr>
<tr>
<td>1991</td>
<td>74.10</td>
<td>73.50</td>
<td>0.60</td>
<td>78.34</td>
<td>77.90</td>
<td>0.44</td>
</tr>
<tr>
<td>1992</td>
<td>74.06</td>
<td>73.80</td>
<td>0.26</td>
<td>78.41</td>
<td>78.20</td>
<td>0.21</td>
</tr>
<tr>
<td>1993</td>
<td>74.76</td>
<td>74.00</td>
<td>0.76</td>
<td>78.31</td>
<td>78.30</td>
<td>0.01</td>
</tr>
<tr>
<td>1994</td>
<td>74.55</td>
<td>74.10</td>
<td>0.45</td>
<td>78.46</td>
<td>78.40</td>
<td>0.06</td>
</tr>
<tr>
<td>1995</td>
<td>74.35</td>
<td>74.20</td>
<td>0.15</td>
<td>78.46</td>
<td>78.60</td>
<td>-0.14</td>
</tr>
<tr>
<td>1996</td>
<td>74.84</td>
<td>74.50</td>
<td>0.34</td>
<td>78.65</td>
<td>78.90</td>
<td>-0.25</td>
</tr>
<tr>
<td>1997</td>
<td>75.60</td>
<td>74.90</td>
<td>0.70</td>
<td>79.14</td>
<td>79.10</td>
<td>0.04</td>
</tr>
<tr>
<td>1998</td>
<td>75.72</td>
<td>75.30</td>
<td>0.42</td>
<td>79.10</td>
<td>79.40</td>
<td>-0.30</td>
</tr>
<tr>
<td>1999</td>
<td>76.24</td>
<td>75.60</td>
<td>0.64</td>
<td>79.49</td>
<td>79.70</td>
<td>-0.21</td>
</tr>
<tr>
<td>2000</td>
<td>76.49</td>
<td>76.10</td>
<td>0.39</td>
<td>79.85</td>
<td>80.10</td>
<td>-0.25</td>
</tr>
<tr>
<td>2001</td>
<td>77.10</td>
<td>76.40</td>
<td>0.70</td>
<td>80.22</td>
<td>80.40</td>
<td>-0.18</td>
</tr>
</tbody>
</table>

Our complete life tables can be used to compute values of life annuities, perform a more accurate analysis of retirement security, and derive tables of multipliers for the purpose of assessing the amount of pecuniary loss in personal injury litigations for the general Singaporean population. The proposed method can also be applied, with certain modifications, in countries other than Singapore where complete period life tables are not available.

Further work is needed to develop a better understanding of the demography of Singapore. As mortality has been improving continuously over the years in Singapore, it will be interesting to explore the way in which projected mortality improvements can be built into specialized complete life tables used for annuities.

References

Figure 2: Male and Female $-\log(q_x)$ from Our Complete Life Table and CVT (1992)


Approximating the Bias and Variance of Chain Ladder Estimates Under a Compound Poisson Model

Janagan Yogaranpan,* Sue Clarke,+ Shauna Ferris,‡ and John Pollard§

Abstract

We consider the problem of estimating the outstanding claims produced by a homogeneous general insurance portfolio. The specific model considered in this paper is one where the number of claims in any loss period follows a Poisson distribution, settlement delays follow the same multinomial distribution, and settlements are single lump sums that are independent identically distributed random variables. Simulations using this model reveal that the development ratios and the outstanding claims estimates produced using the chain ladder method are positively biased. We obtain approximate formulas

*Janagan Yogaranpan, B.Com. (Hons), A.I.A.A., graduated with first class honors from Macquarie University in 2000. He is an actuarial analyst at ING Australia. Mr. Yogaranpan's address is: ING Australia, 347 Kent St, Sydney, NSW 2000, AUSTRALIA. E-mail: janagan.yogaranpan@ing.com.au
+Sue Clarke, B.Tech., F.LA.A., is a lecturer at Macquarie University. She was a statistician in the finance industry then joined Skandia Life before joining this university. Ms. Clarke's address is: Actuarial Studies Department, EFS, Macquarie University, NSW 2109, AUSTRALIA. E-mail: sue.clarke@efs.mq.edu.au
‡Shauna Ferris, M.Comp., F.LA.A., is a lecturer at Macquarie University. She spent several years working in life insurance, superannuation, and banking before joining this university. Ms. Ferris' address is: Actuarial Studies Department, EFS, Macquarie University, NSW 2109, AUSTRALIA. E-mail: shauna.ferris@efs.mq.edu.au
§John Pollard, Ph.D., F.I.A., F.I.A.A., F.S.S., F.A.S.S.A., is an emeritus professor at Macquarie University. 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. Dr. Pollard's address is: Actuarial Studies Department, EFS, Macquarie University, NSW 2109, AUSTRALIA. E-mail: john.pollard@efs.mq.edu.au

The authors wish to thank the two anonymous referees and the editor for their helpful comments and suggestions that have improved this paper.
for the biases using Taylor series expansions of the random variables about their means. The same methods are used to obtain approximations for the variances and covariances of the projection ratios and the outstanding claims estimates. A simulation study reveals that our formulas are highly accurate.

Key words and phrases: outstanding claims, reserving, stochastic run-off triangles, chain ladder moments

1 Introduction

Suppose there are data available for \( n \) calendar accident years, with the calendar years labeled 0, 1, ..., \( n-1 \). We define the total claims paid in development year \( j \) of accident year \( i \) as \( S_{ij} \), where \( i, j = 0, 1, 2, \ldots, n-1 \). Our aim is to estimate the outstanding claims at the end of calendar year \( n-1 \). The claim payments that are known to date form the upper triangle of the claims run-off as shown below.

<table>
<thead>
<tr>
<th>Year of Origin (( i ))</th>
<th>Development Year (( j ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_{00} )</td>
<td>( S_{01} )</td>
</tr>
<tr>
<td>( S_{10} )</td>
<td>( S_{11} )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( S_{n-2,0} )</td>
<td>( S_{n-2,1} )</td>
</tr>
<tr>
<td>( S_{n-1,0} )</td>
<td></td>
</tr>
</tbody>
</table>

For notational convenience we define \( X_{ab} \) as the sum of all run-off entries in the rectangle from cell (0,0) to cell (a,b) inclusive:

\[
X_{ab} = \sum_{i=0}^{a} \sum_{j=0}^{b} S_{ij}
\]

In the standard application of the chain ladder method, the development ratios

\[
m_{r/r-1} = \frac{X_{n-r-1,r}}{X_{n-r-1,r-1}}
\]
are calculated for \( r = 1, 2, \ldots, n - 1 \). If we define \( Y_i \) as the total claim payments observed to date for accident year \( i \), and \( M_i \) as the product of all development ratios employed in the development of \( Y_i \), then

\[
Y_i = \sum_{j=0}^{n-i-1} S_{ij} \tag{2}
\]

\[
M_i = \prod_{k=n-i}^{n-1} m_{k/k-1} \tag{3}
\]

and the estimated outstanding claims for accident year \( i \), \( OS_i \), is

\[
OS_i = Y_i \times (M_i - 1). \tag{4}
\]

The total of \( OS_i \)'s over accident years with incomplete run-off (i.e., \( i = 1, 2, \ldots, n - 1 \)) gives the chain ladder's overall outstanding claims estimate.

Mack and Venter (2000), Renshaw and Verrall (1998), and other authors have noted that the chain ladder method was originally developed as a deterministic algorithm with no stochastic model underlying it. To estimate the bias and the prediction uncertainty (variance of outstanding claims) a stochastic model is essential. The conclusions reached depend on the model selected.

For example, Murphy (1994) adopted a regression approach to the chain ladder development process and concluded that the simple average development factor method and the weighted average development factor method are unbiased. Gogol (1995), however, has pointed out that "it is only because Murphy's models have unrealistic properties that it is possible to prove that the estimators are unbiased." Gogol demonstrated mathematically why there is positive bias. Mack (1993), on the other hand, assumed

\[
\mathbb{E} [C_{i,r+1} | C_{i,k}, \ k = 0, 1, \ldots, r] = \rho_r C_{i,r} \tag{5}
\]

where \( C_{i,j} = \sum_{k=0}^{j} S_{i,k} \), and \( \rho_r \) is a constant independent of \( i \).

The assumption implied by equation (5) does not hold for our simple compound Poisson model. In our model the number of claims for a particular accident year is Poisson, the number of these claims settled in the various development years follows the multinomial distribution, and lump sum claim amounts are independent and identically distributed. Nor does it hold for the run-off of numbers of claims under the above Poisson/multinomial assumptions, for which the usual
The inapplicability of the Mack model to the collective model has been noted by Schiegl (2002), who points out that the expected cumulative claims by the end of development year \( r+1 \) should be proportional to the expected (as opposed to actual) cumulative claims at the end of the previous development year, i.e.,

\[
\mathbb{E} [C_{i,r+1} | C_{i,k}, \ k = 0, 1, \ldots, r] = \xi_r \mathbb{E} [C_{i,r}]
\]

where \( \xi_r \) is a constant independent of \( i \).

Stanard (1985) performed a number of simulations to investigate the bias and variance in outstanding claims estimates for various loss reserving methods including the chain ladder. His simulations assumed a relatively small random number of claims for each development year, a uniform distribution of accidents over the year, and an exponential distribution for claim reporting and for claim settlement. Stanard found that the chain ladder produced a substantial positive bias and, by considering the ratio of two random variables, proved there must be a bias.

Schiegl (2002) used simulation to investigate the safety loading required in conjunction with a chain ladder estimate of outstanding claims. For normalization purposes, she defined the relative bias as the expected value of the difference between the chain ladder and the simulated estimates of outstanding claims divided by the square root of the mean square error of the chain ladder estimate. She pointed out that because of correlations between the numerator and denominator, the sign of her relative bias may differ from the un-normalized expected difference between the chain ladder estimate and actual outstanding claims value. That is, if she had not divided by the square root of the estimated mean square error, she may have found a positive bias instead of a negative bias.

In practice, actuaries wish to know whether the chain ladder estimate tends to be higher or lower than the underlying value, as well as the variance of the various chain ladder estimates. These are the problems addressed in this paper. We note that Taylor (2002) has confirmed the positive bias that we demonstrate.
2 The Model

Consider a homogeneous general insurance portfolio described below:

1. The individual claim amounts are independent identically distributed random variables with first- and second-order moments about the origin of $\tau_1$ and $\tau_2$, respectively;

2. Each claim is settled as a single lump sum amount with no partial payment before settlement;

3. The total number of claims occurring in accident year $i$ (including IBNR claims) is a Poisson random variable with mean $\lambda_i$; and

4. The probability that a claim is settled in development year $j$ is $p_j$ with $n$ being sufficiently large so that $\sum_{j=0}^{n-1} p_j = 1$.

Let $N_{ij}$ denote the number of claims settled for accident year $i$ in development year $j$, then $N_{ij}$ is a multinomial variable conditional on a Poisson variable. It follows, therefore, that $N_{ij}$ is a Poisson random variable with expectation $\lambda_ip_j$, and the $N_{ij}$s are mutually independent for $i, j = 0, 1, \ldots, n-1$. Thus, the total claim payments for accident year $i$ made in development year $j$, $S_{ij}$, has a compound Poisson distribution with mean $\lambda_ip_j\tau_1$ and variance $\lambda_ip_j\tau_2$. Furthermore, the run-off entries $S_{ij}$ are mutually independent for $i, j = 0, 1, \ldots, n-1$.

For convenience, we define $\gamma$ to be the ratio of the variance to the mean for each of the $S_{ij}$s. Under our model this ratio is

$$\gamma = \frac{\text{Var}[S_{ij}]}{\mathbb{E}[S_{ij}]} = \frac{\lambda_ip_j\tau_2}{\lambda_ip_j\tau_1} = \frac{\tau_2}{\tau_1},$$

which is independent of the accident and development years.

We further define the expected ultimate total claims cost for accident year $i$, $\alpha_i$, as

$$\alpha_i = \mathbb{E}\left[\sum_{j=0}^{n-1} S_{ij}\right] = \sum_{j=0}^{n-1} \lambda_ip_j\tau_1 = \lambda_i\tau_1.$$

The following notation is used for convenience:
and

$$P_j = \sum_{u=0}^{j} P_u.$$  \hspace{1cm} (9)

The source of the bias of chain ladder estimates under the compound Poisson model lies in the definition of the development ratios. Consider development ratios defined as the ratio of the expectations (as opposed to the ratio of observed actual values), i.e., as

$$m_{r/r-1} = \frac{E[X_{n-r-1,r}]}{E[X_{n-r-1,r-1}]}.$$  \hspace{1cm} (10)

Noting that $E[X_{ij}] = \theta_i P_j$ and $P_{n-1} = 1$, then

$$\frac{\sum_{i=1}^{n-1} OS_i}{\sum_{i=1}^{n-1} Y_i \times \left( \prod_{k=n-i}^{n-1} \frac{\theta_{n-k-1} P_k}{\theta_{n-k-1} P_{k-1}} - 1 \right)}$$

$$= \sum_{i=1}^{n-1} \alpha_i P_{n-i-1} \times \left( \prod_{k=n-i}^{n-1} \frac{P_k}{P_{k-1}} - 1 \right)$$

$$= \sum_{i=1}^{n-1} \alpha_i (1 - P_{n-i-1}),$$

which is exactly the expected amount of claims in the unobservable part of the run-off. Therefore using equation (10) for development ratios leads to unbiased estimates under our model. The chain ladder method, however, corresponds to equation (1) instead. As small as this distinction may seem, it introduces biases under our model.

### 3 The Bias and Variance of Development Ratios

For any rectangle of cells $A$ of the run-off, define $T_A$ to be the total claim payments observed in $A$, and define $\mu_A = E[T_A]$. Consider a general development ratio $m$ with rectangles $A$ and $B$ in the run-off defined such that $m = T_A/T_B$. As noted in Section 2, the unbiased ratio required to project the outstanding claims of the portfolio is $\mu_A/\mu_B$. The ratio $m$ on the other hand, has expectation

$$E[m] = E\left[\frac{T_A}{T_B}\right] = \frac{\mu_A}{\mu_B} E\left[\left(1 + \frac{T_A - \mu_A}{\mu_A}\right)\left(1 + \frac{T_B - \mu_B}{\mu_B}\right)^{-1}\right].$$  \hspace{1cm} (11)
Assuming the expression within the square brackets in equation (11) can be expanded as a series of the form \((1 + x)^{-1} = 1 - x + x^2 - x^3 + \ldots\), and assuming \(\mu_A\) and \(\mu_B\) are sufficiently large so that all third and higher order terms are negligible leads to

\[
E[m] \approx \frac{\mu_A}{\mu_B} (1 + V_B A),
\]  

(12)

where for convenience we define the terms

\[
K_{AB} = \frac{\text{Cov}[T_A, T_B]}{\mu_A \mu_B}
\]  

(13)

and

\[
V_{BA} = K_{BB} - K_{AB}.
\]  

(14)

The quantity \(V_{BA}\) is termed the approximate proportional bias. From equations (12), (13), and (14) it is apparent that a stochastic run-off model yields a bias in chain ladder estimates under the compound Poisson model.

The following theorem is needed to assist in the development of our approximations. An illustration and proof of Theorem 1 is given in Appendix A.

**Theorem 1.** Let \(G\) and \(H\) be rectangles of cells in the run-off, and let \(R\) be the smallest rectangle that includes all the cells of \(G\) and \(H\). If the rows of \(R\) coincide with the rows of either \(G\) or \(H\) and the columns of \(R\) coincide with the columns of either \(G\) or \(H\), then

\[
K_{GH} = \frac{Y}{\text{Total Payments Expected in } R}
\]

where \(Y\) is given in equation (6).

Because of the manner in which development ratios are calculated, \(B \subseteq A\), so the smallest rectangle including all the cells of \(A\) and \(B\) is \(A\) itself; while, trivially, the smallest rectangle for \(B\) and \(B\) is \(B\). Therefore applying Theorem 1 to equation (14), we deduce that the approximate proportional bias in \(m\) is

\[
V_{BA} = \frac{Y}{\mu_B} - \frac{Y}{\mu_A}
\]  

(15)

The direction of the bias depends on the relationship between \(\mu_A\) and \(\mu_B\). For example, if negative incremental claims are allowed, then it is
possible that $\mu_A < \mu_B$, and the bias in the chain ladder development ratios may be positive or negative. Under our compound Poisson model with nonnegative incremental claims, however, $\mu_A > \mu_B$; therefore, the development ratios used to project the cumulative sums of claim payments are positively biased, and the chain ladder approach will tend to overestimate outstanding claims liabilities.

The variance of the development ratio can be found as follows:

$$\mathbb{E}[m^2] = \left(\frac{\mu_A}{\mu_B}\right)^2 \mathbb{E}\left[\left(1 + \frac{T_A - \mu_A}{\mu_A}\right)^2 \left(1 + \frac{T_B - \mu_B}{\mu_B}\right)^{-2}\right].$$

Again, using a binomial expansion and neglecting the appropriate terms yields the approximation

$$\mathbb{E}[m^2] \approx \left(\frac{\mu_A}{\mu_B}\right)^2 (1 + K_{AA} + 3K_{BB} - 4K_{AB}).$$

As $B \subset A$, it follows from Theorem 1 that $K_{AA} = K_{AB}$. Subtracting $(\mathbb{E}[m])^2$ as approximated using equation (12) and ignoring third order terms, we conclude that

$$\text{Var}[m] \approx \left(\frac{\mu_A}{\mu_B}\right)^2 V_{BA}. \quad (16)$$

4 Bias and Variance of Outstanding Claims

Two of the concerns in outstanding claims estimation are the bias and the variance of the overall estimate. The definition of the variance of the outstanding claims estimate requires some clarification. In our opinion there are three main variance measures that practitioners might consider:

1. The variance of the actual outstanding claims amount;

2. The variance of the outstanding claims estimate based upon particular estimates of the model parameters; and

3. The variance of the chain ladder outstanding claims estimate.

Many authors concentrate on the second measure. Given the subject of this paper, however, it is the third measure that is relevant and is the one used in this paper. Some preliminary results are now given.
4.1 Preliminary Results

4.1.1 Development Ratios Are Effectively Uncorrelated

Consider a $5 \times 5$ run-off triangle with rows and columns numbered from 0 to 4. The development ratios $m_{1/0}$ and $m_{3/2}$ are based on rectangles of cells $A$, $B$, $C$, and $D$ defined such that $m_{1/0} = T_A/T_B$ and $m_{3/2} = T_C/T_D$. This implies $T_A = X_{3,1}$, $T_B = X_{3,0}$, $T_C = X_{1,3}$, and $T_D = X_{1,2}$. Working through the expansions as before and taking expectations, we find that

$$
\mathbb{E} \left[ m_{1/0} m_{3/2} \right] \approx \frac{\mu_A \mu_C}{\mu_B \mu_D} (1 + V_{BA} + V_{DC} + K_{AC} + K_{BD} - K_{CB} - K_{AD}).
$$

From Theorem 1, $K_{AC} = K_{CB}$ and $K_{BD} = K_{AD}$, so that the subscripted $K$ terms in equation (17) sum to zero. By inspection, the right side of equation (17) is equal to $\mathbb{E} [m_{1/0}] \mathbb{E} [m_{3/2}]$ as found using equation (12) and ignoring all third order and higher terms. We conclude that the covariance of the two development ratios is approximately zero. The same approach can be used to show that any two arbitrary development ratios are effectively uncorrelated.

4.1.2 Uncorrelated Accident Year Payments and Development Ratios

Within the same $5 \times 5$ run-off triangle as before, define $C$ as the rectangle of cells relating to claim payments observed to date in respect of accident year 1, i.e., $C$ contains $\{S_{ij}\}$ where $i = 1$ and $j = 0, 1, 2, 3$. Consider the development ratio $m_{2/1}$, with rectangles of cells $A$ and $B$ defined such that $m_{2/1} = T_A/T_B$. By inspection, $T_A$, $T_B$, and $T_C$ share common run-off entries, so $T_C$ is not independent of $m_{2/1}$. Adopting the same approach as before, we discover that

$$
\mathbb{E} \left[ T_C m_{2/1} \right] = \mathbb{E} \left[ \frac{T_C T_A}{T_B} \right] \approx \frac{\mu_C \mu_A}{\mu_B} (1 + V_{BA} + K_{AC} - K_{BC}).
$$

From Theorem 1, $K_{AC} = K_{BC}$. The right side of equation (18) is therefore equal to $\mathbb{E} [T_C] \mathbb{E} [m_{2/1}]$ as approximated by equation (12). So with covariance approximately zero, $T_C$ and $m_{2/1}$ are effectively uncorrelated. The same conclusion is reached irrespective of the accident year chosen and the development ratio involving common run-off entries.
4.2 Bias, Variances, and Covariances

4.2.1 Product of Development Ratios

Consider three different development ratios, $m_1$, $m_2$, and $m_3$, defined as $T_A/T_B$, $T_C/T_D$, and $T_E/T_F$, respectively. Using the same techniques as before, we can show that

$$\mathbb{E}[m_1 m_2 m_3] \approx \frac{\mu_A \mu_C \mu_E}{\mu_B \mu_D \mu_F} (1 + V_{BA} + V_{DC} + V_{FE})$$

Proportional Bias $\approx V_{BA} + V_{DC} + V_{FE}$

$$\text{Var}[m_1 m_2 m_3] \approx \left(\frac{\mu_A \mu_C \mu_E}{\mu_B \mu_D \mu_F}\right)^2 (V_{BA} + V_{DC} + V_{FE})^2.$$ 

In general, the proportional bias in the product of a set of development ratios is approximately the sum of the relevant $V$ terms, and the variance is approximately the sum of the relevant $V$ terms multiplied by the square of the product of the relevant unbiased ratios.

4.2.2 Accident Year Outstanding Claims Estimates

Let us use $Y_i$ and $M_i$ as defined in equations (2) and (3). $Y_i$ is an unbiased estimator of the expected total claim payments to date, and as $Y_i$ and $M_i$ do not depend on any common run-off entries, they must be independent. According to equations (4) and (19), therefore, the actual bias (not the proportional bias) in $O_S_i$ is approximately the total claims expected in accident year $i$ (that is, $\alpha_i$) multiplied by the sum of the $V$ terms relating to $M_i$.

Given the independence of $Y_i$ and $M_i$, the variance of $O_S_i$ can be found as follows:

$$\text{Var}[(M_i - 1)Y_i] = \text{Var}[M_i] \mathbb{E}[Y_i^2] + \text{Var}[Y_i] (\mathbb{E}[M_i] - 1)^2.$$ 

(21)

4.2.3 Covariances Between Outstanding Claims Estimates for Different Accident Years

Consider $Y_i$, $M_i$, and $O_S_i$ as defined in equations (2), (3), and (4), and similarly define $O_S_q$, $Y_q$, and $M_q$ for a later accident year $q$. Under our model, $Y_i$ and $Y_q$ are independent. Because $Y_q$ relates to a later accident year than $Y_i$, $M_i$ is a factor of $M_q$. So let us write $M_q$ as $M^*M_i$, where $M^*$ and $M_i$ do not contain any common development ratios and are
effectively uncorrelated. The covariance of the two outstanding claims estimates is therefore:

$$\text{Cov}[OS_t, OS_q] = E[(M_i - 1)Y_i (M^*_i - 1)Y_q]$$
$$- E[(M_i - 1)Y_i] E[(M^*_i - 1)Y_q]. \quad (22)$$

Taking account of the independence of $Y_i$ and $Y_q$, and the fact that all the other $M$ and $Y$ terms in equation (22) are effectively uncorrelated, we deduce that

$$\text{Cov}[OS_t, OS_q] \approx E[Y_i] E[Y_q] E[M^*_i] \text{Var}[M_i] \quad (23)$$

for $i = 1, \ldots, q - 1$.

4.2.4 Variance of Overall Outstanding Claims Estimate

The overall outstanding claims estimate is the sum of the estimates for the individual accident years. Its variance is readily approximated from the variances of individual accident year estimates and the covariances of these estimates.

4.2.5 Non-Homogeneous Model of Claim Settlements

Thus far we have considered claim size patterns that are independent of notification delays. Given the strong assumption of independence of run-off entries, the above results will hold if it can be further assumed that claims at differing levels of severity are mutually independent.

For example, separate run-off triangles and sets of parameters for small, medium, and large claim sizes can be investigated, with the results of this paper applicable to each of these triangles. The items of interest can then be aggregated.

4.3 Practical Formulas

It is possible to simplify the results obtained so far in Section 4 for practical application by noting that the $V$ term for the development ratio $m_{r/r-1}$ can be expressed as

$$V_r = \frac{Y}{\theta_{n-r-1}} \left( \frac{1}{P_{r-1}} - \frac{1}{P_r} \right). \quad (24)$$
If we then define the (backwards) cumulative sum of the $V$ terms as

$$V^C_r = \sum_{u=r}^{n-1} V_u,$$  \hspace{1cm} (25)

we discover from equations (19) and (20) that

$$\mathbb{E} \left[ m_{r/r-1} \times m_{r+1/r} \times \cdots \times m_{n-1/n-2} \right] \approx \frac{1}{P_{r-1}} \left( 1 + V^C_r \right)$$  \hspace{1cm} (26)

and

$$\text{Var} \left[ m_{r/r-1} \times m_{r+1/r} \times \cdots \times m_{n-1/n-2} \right] \approx \left( \frac{1}{P_{r-1}} \right)^2 V^C_r.$$  \hspace{1cm} (27)

For $i = 1, \ldots, n-1$, the approximate bias of accident year $i$'s outstanding claims estimate is

$$\text{Bias} (\text{OS}_i) \approx \alpha_i V^C_{n-i}.$$  \hspace{1cm} (28)

Furthermore, using equations (26), (27), (21), and (23) and simplifying (details are given in Appendix B),

$$\text{Var} [\text{OS}_i] \approx \alpha_i^2 V^C_{n-i} + \alpha_i P_{n-i-1} \gamma \left( \frac{1 - P_{n-i-1}}{P_{n-i-1}} \right)^2$$

$$+ \alpha_i \gamma \left( \frac{3}{P_{n-i-1}} - 2 \right) V^C_{n-i}$$  \hspace{1cm} (29)

and

$$\text{Cov} [\text{OS}_i, \text{OS}_q] \approx \alpha_i \alpha_q V^C_{n-i}$$  \hspace{1cm} (30)

for $i = 1, \ldots, q-1$. Larger $\alpha$ parameters correspond to higher expected total claims. As the portfolios considered become larger, the $V_r$ and $V^C_r$ terms approach zero, as do the biases in individual development ratios and their products. This is due to the sums of $\alpha$ parameters that appear in the denominators of $V$ terms.

Furthermore the approximate biases (equation (28)) are not linear in the total expected amounts of claims, as the $V$ terms of equation (28) are multiplied by other $\alpha$ parameters. Nevertheless, we see that if the portfolio changes with all $\alpha$'s increasing by a common factor, then the approximate biases of the outstanding claims estimates do
not change. This is not an intuitive result. We also see from equations (24) and (28) that the approximate biases are linear in $\gamma$. Taylor (2002) also details this result and finds numerical proof in a comparison of Exhibits I and II of Stanard (1985), whose simulation models roughly follow the restrictions of Section 2.

The covariances between different accident year claims estimates (equation (30)) grow in proportion to the portfolio size, as do the first two terms of the variance result for the accident year estimate (equation (29)). The last term in this variance approximation is similar to the bias approximations, as it does not grow in proportion to portfolio size.

5 Simulation Study

5.1 Simulation Results for Individual Development Ratios

One million simulations of a $5 \times 5$ run-off were used to test the results for individual development ratios. The assumed underlying claim number parameters and settlement proportions are shown in Table 2. Individual claim sizes were assumed to be exponentially distributed with a mean of 500. Therefore, $\tau_1 = 500$, $\tau_2 = 500,000$, and $\gamma = \tau_2 / \tau_1 = 1,000$. This highly skewed distribution was used to stress test the results, as with lower skewness we might expect our formulas to produce better approximations.

<table>
<thead>
<tr>
<th>Accident Year $(i)$</th>
<th>Claim Frequency $(\lambda_i)$</th>
<th>Development Year $(j)$</th>
<th>Proportion Settled $(p_j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>200</td>
<td>0</td>
<td>40%</td>
</tr>
<tr>
<td>1</td>
<td>300</td>
<td>1</td>
<td>30%</td>
</tr>
<tr>
<td>2</td>
<td>240</td>
<td>2</td>
<td>20%</td>
</tr>
<tr>
<td>3</td>
<td>360</td>
<td>3</td>
<td>5%</td>
</tr>
<tr>
<td>4</td>
<td>220</td>
<td>4</td>
<td>5%</td>
</tr>
</tbody>
</table>

The observed proportional biases and variances are compared in Table 3 with the approximate theoretical values. The proportional bias shown in Table 3 is the average simulated ratio less the unbiased ra-
tio, expressed as a proportion of the unbiased ratio. It is clear that equations (15) and (16) produce reliable approximations.

### Table 3
Proportional Biases and Variances Estimated by Simulation

<table>
<thead>
<tr>
<th>RATIO</th>
<th>BIAS</th>
<th>VAR</th>
<th>NUM</th>
<th>DENOM</th>
<th>APBIAS</th>
<th>APVAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{1/0}$</td>
<td>0.0020</td>
<td>0.0061</td>
<td>385,000</td>
<td>220,000</td>
<td>0.0019</td>
<td>0.0060</td>
</tr>
<tr>
<td>$m_{2/1}$</td>
<td>0.0009</td>
<td>0.0014</td>
<td>333,000</td>
<td>259,000</td>
<td>0.0009</td>
<td>0.0014</td>
</tr>
<tr>
<td>$m_{3/2}$</td>
<td>0.0002</td>
<td>0.0003</td>
<td>237,500</td>
<td>225,000</td>
<td>0.0002</td>
<td>0.0003</td>
</tr>
<tr>
<td>$m_{4/3}$</td>
<td>0.0005</td>
<td>0.0006</td>
<td>100,000</td>
<td>95,000</td>
<td>0.0005</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

Notes: RATIO = Development ratio; BIAS = Proportional bias of simulated ratios; VAR = Variance of simulated ratios; NUM = Expected numerator; DENOM = Expected denominator; APBIAS = Approximate proportional bias based on equation (15); and APVAR = Approximate variance based on equation (16).

While the approximations in Table 3 are consistent with the estimates obtained by simulation, the number of claims assumed was large. The simulations were therefore repeated with a tenfold decrease in the Poisson claim frequencies. Given the tiny size of the run-off, the approximations are surprisingly good (Table 4).

### Table 4
Simulated Biases and Variances with Reduced Claim Frequencies

<table>
<thead>
<tr>
<th>RATIO</th>
<th>BIAS</th>
<th>VARIANCE</th>
<th>APROXBIA</th>
<th>APROXVAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{1/0}$</td>
<td>0.0206</td>
<td>0.0720</td>
<td>0.0195</td>
<td>0.0597</td>
</tr>
<tr>
<td>$m_{2/1}$</td>
<td>0.0093</td>
<td>0.0164</td>
<td>0.0086</td>
<td>0.0142</td>
</tr>
<tr>
<td>$m_{3/2}$</td>
<td>0.0026</td>
<td>0.0030</td>
<td>0.0023</td>
<td>0.0026</td>
</tr>
<tr>
<td>$m_{4/3}$</td>
<td>0.0064</td>
<td>0.0088</td>
<td>0.0053</td>
<td>0.0058</td>
</tr>
</tbody>
</table>

Notes: RATIO = Development ratio; BIAS = Proportional bias of simulated ratios; VARIANCE = Variance of simulated ratios; APROXBIA = Approximate proportional bias based on equation (15); and APROXVAR = Approximate variance based on equation (16).

### 5.2 Simulation Results for Outstanding Claims Estimates

The same assumptions and simulations were used to determine the bias (Table 5) and second-order moments (Table 6) of the chain ladder
outstanding claims estimates for each of the accident years. The comparisons with the approximate theoretical bias (Table 5) and second order moments (Table 7) are good. Calculation details for the theoretical formulas are given in Appendix C.

### Table 5

<table>
<thead>
<tr>
<th>Accident Year</th>
<th>Bias of Simulated Estimates</th>
<th>Approximate Bias from Equation (28)</th>
<th>Expected Outstanding Claims</th>
<th>Relative Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81.84</td>
<td>78.95</td>
<td>7,500</td>
<td>1.09%</td>
</tr>
<tr>
<td>2</td>
<td>94.70</td>
<td>91.23</td>
<td>12,000</td>
<td>0.79%</td>
</tr>
<tr>
<td>3</td>
<td>301.78</td>
<td>291.28</td>
<td>54,000</td>
<td>0.56%</td>
</tr>
<tr>
<td>4</td>
<td>391.81</td>
<td>392.29</td>
<td>66,000</td>
<td>0.59%</td>
</tr>
<tr>
<td>Overall Result</td>
<td>870.13</td>
<td>853.75</td>
<td>139,500</td>
<td>0.62%</td>
</tr>
</tbody>
</table>

The variance of the overall outstanding claims estimate based on the simulation study is simply the sum of all the moments in Table 6, namely $4.362 \times 10^8$. This agrees closely with the approximate value obtained from Table 7 of $4.270 \times 10^8$. The discrepancy is about 2.1%.

As we know the parameters underlying the model, the variance of the true outstanding claims for all accident years (that is, the first variance measure mentioned in Section 4) can be evaluated quickly and easily: $1.395 \times 10^8$. The variance of the chain ladder estimate is around three times as great, reflecting the uncertainty introduced by the need to use parameters estimated by the chain ladder method.

### Table 6

| Simulated Outstanding Claims Estimates by Accident Year $\times 10^{-6}$ |
|-------------------------------|-------------------------------|-----------------------------|-----------------------------|
| 1                             | 2                             | 3                           | 4                           |
| 1                             | 12.75                         | 9.80                        | 14.78                       | 9.02           |
| 2                             | 9.80                          | 12.73                       | 16.95                       | 10.37          |
| 3                             | 14.78                         | 16.95                       | 77.63                       | 32.91          |
| 4                             | 9.02                          | 10.37                       | 32.91                       | 145.47         |
Table 7
Approximate Covariance Matrix of Outstanding Claims
Estimates by Accident Year $\times 10^{-6}$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.33</td>
<td>9.47</td>
<td>14.21</td>
<td>8.68</td>
</tr>
<tr>
<td>3</td>
<td>14.21</td>
<td>16.42</td>
<td>76.24</td>
<td>32.04</td>
</tr>
<tr>
<td>4</td>
<td>8.68</td>
<td>10.04</td>
<td>32.04</td>
<td>144.31</td>
</tr>
</tbody>
</table>

The high accuracy with which the approximate theoretical variance estimates the true variance is due to some extent to the relatively high (but nevertheless realistic) assumed claim frequency. With lower claim frequencies, the errors in the approximate variances and covariances become more significant. With a tenfold decrease in claim frequencies, the error in the approximate variance of the overall outstanding claims estimate rises to 27.1% ($4.543 \times 10^7$ compared with a simulated value of $5.819 \times 10^7$)—or a 13% error in the standard deviation. Even in this situation, with relatively few expected claims and a highly skewed claim size distribution, the approximations still provide reasonable indications of the degree of uncertainty in the chain ladder outstanding claims estimates.

In practice the underlying parameter values will be unknown. Estimating parameters from actual insurance data will introduce uncertainty and possibly biases in the parameter estimates, which in turn will affect the approximations of this paper. Such impacts are beyond the scope of this paper—it must be emphasized that the approximations are valid only if the parameter values are known in advance. For this reason we have performed a simulation study rather than applying the approximations to actual insurance data.

6 Concluding Remarks

The formulas we have derived allow accurate approximations to the biases introduced when the traditional chain ladder method is used to estimate the outstanding claims under a compound Poisson run-off, and accurate approximations to the variances and covariances of these estimates. Our analysis also reveals that under our simple stochastic
model the development ratios of the chain ladder method are essentially uncorrelated, but are biased. Even in the ideal situation of a large portfolio with independent entries, outstanding claims estimates for different accident years are significantly correlated.

References


 Appendix A. Illustration and Proof of Theorem 1

Illustration

In the context of a $5 \times 5$ run-off with rows and columns numbered 0 to 4, consider rectangles $G$ containing elements $\{S_{ij}\} (i = 0, 1; j = 0, 1, 2, 3)$, and $H$ containing elements $\{S_{ij}\} (i = 0, 1, 2, 3; j = 1, 2)$.

The smallest rectangle $R$ incorporating all the elements of $G$ and $H$ is made up of the elements $\{S_{ij}\} (i = 0, 1, 2, 3; j = 0, 1, 2, 3)$. We note that the rows of $R$ coincide with the rows of $H$ and that the columns of $R$ coincide with those of $G$.

Recall the definition (given in Section 3) of $T_A$ and $\mu_A$ for any rectangle of cells $A$. Because of the mutual independence of all cells of the run-off under our model, $\text{Cov}[T_G, T_H] = \text{Var}[T_W]$, which by equation (6) and the assumptions of Section 2, yields

$$\text{Cov}[T_G, T_H] = \text{Var}[T_W] = \sum_{i=0}^{1} \sum_{j=1}^{2} \alpha_i p_j y.$$

The expectations of $T_G$ and $T_H$ are, respectively

$$\mu_G = \sum_{i=0}^{1} \sum_{j=0}^{3} \alpha_i p_j \quad \text{and} \quad \mu_H = \sum_{i=0}^{4} \sum_{j=1}^{2} \alpha_i p_j.$$

Therefore, $K_{GH}$ (defined in equation (13)) is given by

$$K_{GH} = \frac{y}{\sum_{i=0}^{4} \alpha_i \sum_{j=0}^{3} p_j},$$

where the denominator is the total claim payments expected in $R$.

Proof of Theorem 1

For any rectangle $G$, define $\theta_G$ as the sum of all the $\{\alpha_i\}$ values that relate to the cells of $G$ within the run-off, and $P_G$ as the sum of all the $\{p_j\}$ values that relate to to the cells of $G$. Then, with (i) $G$, $H$, and $R$ defined as in the statement of the theorem; (ii) $W = G \cap H$; and (iii) using the fact that under the model all elements are independent,

$$\mu_G = \theta_G P_G;$$
\[ \mu_H = \theta_HP_H; \]
\[ \mu_R = \max(\theta_G, \theta_H) \times \max(P_G, P_H); \]
\[ \text{Cov}[T_G, T_H] = \text{Var}[T_W] = y \min(\theta_G, \theta_H) \min(P_G, P_H) \]
\[ K_{GH} = \frac{y \min(\theta_G, \theta_H) \min(P_G, P_H)}{(\theta_G P_G) (\theta_H P_H)}. \]

But
\[ \theta_G \theta_H = \min(\theta_G, \theta_H) \max(\theta_G, \theta_H); \]
\[ P_G P_H = \min(P_G, P_H) \max(P_G, P_H). \]

It follows that
\[ K_{GH} = \frac{y}{\mu_R}, \]

and the proof of the theorem is complete.

**Appendix B. Derivation of Approximations**

**Variance Approximations**

Recall the definition of \( M_i \) in equation (3) and the variance equation of equation (21). Substituting equations (26) and (27) in equation (21) yields

\[ \text{Var}[OS_i] \approx \frac{1}{p_{n-i-1}^2} V_{n-i}^C \left[ (\alpha_i P_{n-i-1})^2 + (\alpha_i P_{n-i-1} Y) \right] \]
\[ + (\alpha_i P_{n-i-1} Y) \left[ \frac{1}{P_{n-i-1}} \left( 1 + V_{n-i}^C \right) - 1 \right]^2 \]
\[ = \alpha_i^2 V_{n-i}^C + \frac{1}{P_{n-i-1}^2} V_{n-i}^C \left( \alpha_i P_{n-i-1} Y \right) \]
\[ + (\alpha_i P_{n-i-1} Y) \left[ \left( \frac{1 - P_{n-i-1}}{P_{n-i-1}} \right) + \frac{V_{n-i}^C}{P_{n-i-1}} \right]^2. \]

Ignoring third order and higher terms in the expansion of the above expression and simplifying, we find that

\[ \text{Var}[OS_i] \approx \alpha_i^2 V_{n-i}^C + \alpha_i P_{n-i-1} Y \left( \frac{1 - P_{n-i-1}}{P_{n-i-1}} \right)^2 + \alpha_i Y \left( \frac{3}{P_{n-i-1}} - 2 \right) V_{n-i}^C. \]
Covariance Approximations

In the context of equation (23),

\[ M^* = m_{n-q/n-q-1} \times \ldots \times m_{n-i-1/n-i-2} \]

for \( i = 1, 2, \ldots, q - 1 \). Let \( V^* \) be the bias term corresponding to \( M^* \). Now substituting equations (26) and (27) in equation (23) yields

\[
\text{Cov} [OS_i, OS_q] \approx \alpha_i p_{n-i-1} \alpha_q p_{n-q-1} \left( \frac{p_{n-i-1}}{p_{n-q-1}} (1 + V^*) \right) \times \frac{V_{n-i}^C}{p_{n-i-1}^2}
\]

for \( i = 1, 2, \ldots, q - 1 \). Ignoring third order terms in the expansion of the above expression, we arrive at the simplified result:

\[
\text{Cov} [OS_i, OS_q] \approx \alpha_i \alpha_q V_{n-i}^C
\]

for \( i = 1, 2, \ldots, q - 1 \).

Appendix C. Bias and Variance Approximations

Table C1 shows some of the details of the calculations behind the approximations displayed in Tables 5 and 7. The assumptions are the same as those used earlier in Table 2, and the claim size moments are \( \tau_1 = 500 \) and \( \tau_2 = 500,000 \). In this example, \( \gamma = \tau_2/\tau_1 = 1,000 \), and \( n = 5 \).

<table>
<thead>
<tr>
<th>( i )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_i )</td>
<td>200</td>
<td>300</td>
<td>240</td>
<td>360</td>
<td>220</td>
<td>Table 2</td>
</tr>
<tr>
<td>( \alpha_i )</td>
<td>100,000</td>
<td>150,000</td>
<td>120,000</td>
<td>180,000</td>
<td>110,000</td>
<td>Eqn (7)</td>
</tr>
<tr>
<td>( \theta_i )</td>
<td>100,000</td>
<td>250,000</td>
<td>370,000</td>
<td>550,000</td>
<td>660,000</td>
<td>Eqn (8)</td>
</tr>
<tr>
<td>( p_i )</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.05</td>
<td>0.05</td>
<td>Table 2</td>
</tr>
<tr>
<td>( P_i )</td>
<td>0.4</td>
<td>0.7</td>
<td>0.9</td>
<td>0.95</td>
<td>1.00</td>
<td>Eqn (9)</td>
</tr>
<tr>
<td>( V_i )</td>
<td>0.0019</td>
<td>0.0009</td>
<td>0.0002</td>
<td>0.0005</td>
<td></td>
<td>Eqn (24)</td>
</tr>
<tr>
<td>( V_{5-i}^C )</td>
<td>0.0005</td>
<td>0.0008</td>
<td>0.0016</td>
<td>0.0036</td>
<td></td>
<td>Eqn (25)</td>
</tr>
<tr>
<td>OS_i bias</td>
<td>78.95</td>
<td>91.23</td>
<td>291.28</td>
<td>392.29</td>
<td></td>
<td>Eqn (28)</td>
</tr>
</tbody>
</table>
The approximate covariance matrix of the outstanding claims estimates by accident year \( \times 10^{-6} \) (using equations (29) and (30)) is

\[
\begin{pmatrix}
12.33 & 9.47 & 14.21 & 8.68 \\
9.47 & 12.40 & 16.42 & 10.04 \\
14.21 & 16.42 & 76.24 & 32.04 \\
8.68 & 10.04 & 32.04 & 144.31
\end{pmatrix}
\]

E.g.: \( \text{Cov}[OS_1, OS_2] = 150,000 \times 120,000 \times 0.0005263 = 9.47 \times 10^6 \).
Rapid Calculation of the Price of Guaranteed Minimum Death Benefit Ratchet Options Embedded in Annuities

Eric R. Ulm*

Abstract†

This paper presents a new method of obtaining quick and accurate values and deltas for discrete lookback options using Taylor series expansions. This method is applied to the case of ratchet guaranteed minimum death benefits attached to annuity contracts, and the method is extended to include annuities where a fixed fund is attached to the variable account. Finally, both the speed and the accuracy of the method are compared to Monte Carlo simulation and the exact analytic solution. The Taylor expansion method is shown to be faster and, in most cases, more accurate than the alternative methods.

Key words and phrases: Taylor series, multivariate normal, lookback option, Monte Carlo simulation, risk, lognormal distribution, Black-Scholes formula, geometric Brownian motion

1 Introduction

One of the biggest developments in the life insurance industry in the last decade or so has been the invention and growth of various equity options embedded in variable annuity contracts. These options range

*Eric Ulm, Ph.D., F.S.A., is a visiting assistant professor of actuarial science at the University of Central Florida. He has previously worked for several years as an actuary in the corporate actuarial department of a U.S. insurance company.

Dr. Ulm's address is: Department of Statistics and Actuarial Science, University of Central Florida, Orlando FL 32816, U.S.A. E-mail: eulm@mail.ucf.edu

†The author would like to thank Steven Craighead for useful discussions and assistance in implementing the necessary programs and Lijia Guo and Morgan Wang for helpful discussions. In addition, the author thanks several anonymous referees for comments that have improved the presentation of the paper.
from simple (e.g., guaranteeing the return of principal invested should the annuitant die), to complex options such as a minimum guaranteed fund amount equal to some function of the past history of the fund should the annuitant choose to annuitize at guaranteed purchase rates (Milevsky and Posner, 2001).

The market for customers of variable annuities is competitive, and one of the ways producers attempt to distinguish themselves both from mutual fund providers and other variable annuity providers is by including a death benefit option with the contract. The cheapest and simplest case is one where the insurance company promises to pay out at least a return of the premium paid into the contract on the death of the contract owner, regardless of the actual performance of the underlying funds. The death benefit is commonly made more complicated, as well as more valuable and more expensive in several ways. Many companies offer not only a return of the premiums paid but an accumulation of the premiums at a minimal interest rate on death of the contract owner.

Even more generous provisions exist including reset and ratchet benefits. Death benefits on contracts with a reset provision will be reset to the fund value at various times during the life of the contract and can move up or down, but usually not below the return of premium. Death benefits on ratchet contracts ratchet to the value of the fund at various times during the life of the contract, but only if the resulting benefit is higher than the one in force before the ratchet. Otherwise, the death benefit remains where it was before the ratchet date. A good overview of the state of the market is contained in Milevsky and Posner (2001).

Many companies market and sell annuity products without recognizing the need to hedge the underlying risk and thereby expose themselves to unnecessary levels of equity risk. In addition, financial papers frequently present results for more traditional traded options that are not easily transferred to an insurance environment where the options are embedded in other contracts. Also, analytic results are not obtainable in all cases. When they are, they frequently require the calculation of the multivariate cumulative normal distribution function. This function is not directly computable. Although approximations do exist, they are slow in practice. Therefore, insurance companies usually resort to Monte Carlo simulations to value these options. It can be time-consuming to obtain even a passable value for all the options in all inforce contracts by this method. The situation becomes even worse when simulating the option value along many paths for cash flow testing or ALM purposes. The value of the option must be obtained by
simulation at each time period along each testing path that could lead to billions of required simulations.

Much work has been done to attempt to value these options. An exact analytic solution was obtained for an individual discrete lookback put by Collin-Dufresne, Keirstad, and Ross (1997) by using a change of numeraire, which is theoretically valuable, but involves the cumulative multivariate normal distribution—a distribution that is difficult and time-consuming to evaluate in practice. In addition, the result has been obtained for a variable fund only, while this paper addresses the addition of a fixed fund to the account as well. Tiong (2000) uses the method of Esscher transforms pioneered by Gerber and Shiu (1994) to obtain analytic solutions for cliquet options in equity indexed annuities. Also, Milevsky and Posner (2001) have recently valued lookback guaranteed minimum death benefit (GMDB) options analytically in the specific case of an at-the-money continuous lookback option on a variable fund only, when mortality follows some simple analytic forms.

Why is there a need for a new paper on this subject? The approach of this paper addresses some of the major shortcomings in the practical implementation of the methods described above. The analytic solution is theoretically valuable but there are two major drawbacks involved with using it. First, it is time-consuming. There is no easy way to evaluate the function quickly and accurately because of the large number of multivariate normal functions that must be evaluated. The evaluation of the cumulative multivariate normal function has been addressed by many authors, including Gupta (1963), Wang and Kennedy (1990), Wang (1991), Terza and Welland (1991), Genz (1992), and Genz (2004). The method used in the actual comparisons is that of Somerville (1998a, 1998b). Second, the result has been obtained for a variable fund only, while this paper addresses the addition of a fixed fund to the account as well.

This paper describes a method to compute the Taylor coefficients of the value of a discrete lookback put option, a method that can easily be extended to the case of a ratchet GMDB where the value of the contract at death is the maximum value of the contract at any policy anniversary or the account value at death if larger. It can also be extended to fit the general case where an insurance company faces the variable account is attached to various other accounts that earn a fixed rate independent of equity performance.

While a Taylor series expansion is not as theoretically appealing as a closed form solution, it is just as valuable in practice, especially if it produces relatively quick and accurate results. The major problem,
however, is determining the coefficients of the Taylor series expansion. The next section addresses this issue.

2 Taylor Expansion of Discrete Lookback Put

2.1 Black-Scholes Case

This paper primarily addresses the ratchet GMDB where the death benefit is the maximum value the contract attains on any policy anniversary or the contract value at death, whichever is greater. This is analogous to a series of discrete lookback puts where the notional amount of the put is equal to the original fund multiplied by the probability that the annuitant dies at that point without having previously lapsed his or her policy. This requires a model of surrender and mortality that also affects the GMDB value. We begin by valuing an arbitrary discrete lookback put and then show how it can be extended to cases more relevant to insurance company annuities.

Consider a put issued at time $t_0$ and coming due at some known and fixed time $t_N$. Its value at time $t_N$ is the maximum of the underlying fund values at times $t_1 < t_2 < t_3 \cdots < t_{N-1} < t_N$ and the initial strike $X_0$. We want to determine $V_0$ (the ratchet GMDB value at time $t_0$) given $F_0$ (the total fund value at time $t_0$), $X_0$ (the strike at time $t_0$), and $p_f$ and $p_v$ (the initial percentages invested in fixed and variable accounts). In addition, the risk free rate ($r$), stock volatility ($\sigma$), asset charges ($q$), which are analogous to dividends in the analysis, and fixed growth rate ($g$) are assumed known. For $n = 1, \ldots, N$, we assume $V_n$ is of the form:

$$V_n = F_n \sum_{j,k} f_{j,k} \left( \ln \frac{S_n}{J_n} \right) \phi_n \left( \frac{S_n}{J_n} \right)$$

where $J_n$ is the adjusted strike price given by

$$J_n = \left( \frac{X_n / F_n - p_f}{p_v} \right) S_n,$$

$S_n$ is the stock index at time $t_n$, $F_n$ is the account value at time $t_n$ that behaves as

$$F_n = \left( p_{f_{n-1}} (1 + i_{n-1}) + p_{v_{n-1}} e^{\gamma_n} \right) F_{n-1}$$
\[ \phi_n = i_n \frac{p_{f_n}}{p_{v_n}} \]  

(2)

where \( p_{f_n} \) and \( p_{v_n} \) are the percentages of the fund in the fixed and variable account at time \( t_n \) with \( p_{f_n} + p_{v_n} = 1 \), and

\[ i_n = \exp \left( g \left( t_{n+1} - t_n \right) \right) - 1. \]

Note that \( i_n \) is equivalent to the strike on the variable fund only, taking out the effect of the fixed account. Assume the Taylor series coefficients of \( f_{jkn} \) are known, we show that if \( V_n \) is of the form given in equation (1) then so is \( V_{n-1} \). As \( V_{N-1} \) is simply the Black-Scholes result and can be shown to be of the form given in equation (1), then \( V_0 \) can be shown to be of the same form by induction. The values of the Taylor coefficients are derived automatically during the induction step.

First, we derive the Taylor expansion for a Black-Scholes put at time \( t_{N-1} \) with payoff at time \( t_N \). While this expansion could be obtained by repeated differentiation of the Black-Scholes formula, it will be derived in a more complicated manner similar to algorithmic differentiation [see Wengert (1964)] to illustrate some of the general concepts used in the inductive step. Let \( S_N \) be the stock level and \( F_N \) be the fund level at time \( t_N \). Assume risk-neutral valuation, \( \mu = r - q - \sigma^2 / 2 \), and define

\[ r_n = (t_{n+1} - t_n) r, \]
\[ \mu_n = (t_{n+1} - t_n) \mu, \]
\[ \sigma_n = \sigma \sqrt{(t_{n+1} - t_n)}, \]

and

\[ y_n = \ln \left( S_n / S_{n-1} \right) \]

with \( i_n \) and \( \phi_n \) defined as above. Then \( X_{N-1} \) is the strike on the fund at time \( t_{N-1} \), and the put obeys the equation:

\[ V_{N-1} = e^{-r_{N-1}} E_{S_{N-1}} \left[ (X_{N-1} - F_N)_+ \right] \]  

(3)

where \( x_+ = \max(0, x) \), and \( E_{S_{N-1}} \) denotes expectation condition on (i.e., given) \( S_{N-1} \). We will assume the stock index follows a geometric Brownian motion so that \( y_n \) is normally distributed with mean \( \mu_{n-1} \) and variance \( \sigma^2_{n-1} \), i.e., \( y_n \sim \mathcal{N}(\mu_{n-1}, \sigma^2_{n-1}) \). The integral for the value of this put option is given by:
Equation (4) can be explained by assuming the option is on the variable fund only (hence the \( p \) term in front of the integral), the stock market is normalized to 1 at time \( t_{N-1} \), and the adjusted strike on the variable fund drops by \( \phi_{N-1} \) due to an increase in the relative size of the fixed fund between \( t_{N-1} \) and \( t_N \). Equation (4) can be divided into two integrals, one involving \( e^{-\xi_{N-1}} \) and the other involving \( e^{\gamma_N} \). The terms independent of \( \gamma_N \) can be pulled outside the integral. The remaining terms can be expanded into their individual Taylor expansions and multiplied term by term. This creates two integrals to evaluate:

\[
F_{N-1} \left[ \frac{p_{\gamma_N} e^{-\gamma_N} e^{-\mu_{N-1}^2/2\sigma_{N-1}^2}}{\sqrt{2\pi\sigma_{N-1}^2}} \right] \times \left[ e^{-\xi_{N-1}} \int_{-\infty}^{-\xi_{N-1}} \left[ 1 - \frac{\gamma_N^2}{2\sigma_{N-1}^2} + \frac{\gamma_N^4}{2!(2\sigma_{N-1}^2)} + \cdots \right] d\gamma_N \right. \\
\left. \times [1 + \left( \frac{\mu_{N-1}}{\sigma_{N-1}^2} \right) \gamma_N + \frac{1}{2!} \left( \frac{\mu_{N-1}}{\sigma_{N-1}^2} \right)^2 \gamma_N^2 + \cdots ] d\gamma_N \right] 
\]

and

\[
F_{N-1} \left[ \frac{p_{\gamma_N} e^{-\gamma_N} e^{-\mu_{N-1}^2/2\sigma_{N-1}^2}}{\sqrt{2\pi\sigma_{N-1}^2}} \right] \times \left[ e^{-\xi_{N-1}} \int_{-\infty}^{-\xi_{N-1}} \left[ 1 - \frac{\gamma_N^2}{2\sigma_{N-1}^2} + \frac{\gamma_N^4}{2!(2\sigma_{N-1}^2)} + \cdots \right] d\gamma_N \right. \\
\left. \times [1 + \left( \frac{\mu_{N-1}}{\sigma_{N-1}^2} + 1 \right) \gamma_N + \frac{1}{2!} \left( \frac{\mu_{N-1}}{\sigma_{N-1}^2} + 1 \right)^2 \gamma_N^2 + \cdots ] d\gamma_N \right].
\]
The integrals in expressions (6) and (7) can be split into two integrals, the first one from $-\infty$ to 0 and the second one from 0 to $-\xi_{N-1}$. A Taylor expansion in $\xi_{N-1}$ can be obtained for the integral from 0 to $-\xi_{N-1}$ by multiplying the internal expansions in $y_N$ term by term, integrating term by term and substituting $-\xi_{N-1}$ at the upper bound. Evaluation at the lower bound gives zero. The integrals from $-\infty$ to 0 can be done exactly from equation (4) without any expansions:

$$\sqrt{2\pi \sigma_{N-1}^2} \ e^{\frac{1}{2} \sigma_{N-1}^2 / 2 \sigma_{N-1}^2} \ N \left( -\frac{\mu_{N-1}}{\sigma_{N-1}} \right)$$

for expression (6) and

$$\sqrt{2\pi \sigma_{N-1}^2} \ e^{(\mu_{N-1}+\sigma_{N-1})^2 / 2 \sigma_{N-1}^2} \ N \left( -\frac{\mu_{N-1}}{\sigma_{N-1}} + \sigma_{N-1} \right)$$

for expression (7).

Finally, the Taylor expansion for the integral in expression (6) can be multiplied by the expansion of $e^{-\xi_{N-1}}$ term by term, then added to the expansion for the integral in expression (7). Define a function $\alpha_{N-1}(\xi_{N-1})$ such that

$$V_{N-1} = F_{N-1} p_{v_{N-1}} \alpha_{N-1}(\xi_{N-1})$$

$$= F_{N-1} p_{v_{N-1}} \alpha_{N-1} (-\ln [J_{N-1}/S_{N-1} - \phi_{N-1}]) . \quad (8)$$

To put this into the form of equation (1), we will assume the parameter $\phi_{N-1}/(J_{N-1}/S_{N-1})$ is sufficiently small so that all expansions are valid. (This parameter is small if the percentage of funds in the fixed fund is small.) In Section 3 we show how to expand the radius of convergence if this parameter is not small.

First, we expand the logarithm:

$$V_{N-1} = F_{N-1} p_{v_{N-1}} \alpha_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) - \ln \left( 1 - \frac{\phi_{N-1}}{J_{N-1}/S_{N-1}} \right) \right)$$

$$= F_{N-1} p_{v_{N-1}} \alpha_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) + \psi \right)$$

where

$$\psi = \sum_{k=1}^{\infty} \frac{1}{k} \left( \frac{\phi_{N-1}}{J_{N-1}/S_{N-1}} \right)^k . \quad \text{(9)}$$
As \( \phi_{N-1}/(J_{N-1}/S_{N-1}) \) is small so is \( \psi \). A Taylor series expansion of \( \alpha_{N-1} \) about \( \ln (S_{n-1}/J_{n-1}) \) yields:

\[
V_{N-1} = F_{N-1} p_{v_{N-1}} \left[ \alpha_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) \right) + \sum_{k=1}^{\infty} \alpha^{(k)}_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) \right) \frac{\psi^k}{k} \right],
\]

where \( \alpha^{(k)}_{N-1} \) denotes the \( k \)th derivative of \( \alpha_{N-1} \). Rearranging this expression for \( V_{N-1} \) as a power series in \( \phi_{N-1}/(J_{N-1}/S_{N-1}) \) gives

\[
V_{N-1} = F_{N-1} p_{v_{N-1}} \left\{ \alpha_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) \right) + \alpha^{(1)}_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) \right) \frac{\phi_{N-1}}{J_{N-1}/S_{N-1}} \right. \\
+ \left[ \frac{1}{2} \alpha^{(1)}_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) \right) + \frac{1}{2} \alpha^{(2)}_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) \right) \right] \left( \frac{\phi_{N-1}}{J_{N-1}/S_{N-1}} \right)^2 \\
+ \left[ \frac{1}{3} \alpha^{(1)}_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) \right) + \frac{1}{2} \alpha^{(2)}_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) \right) + \frac{1}{6} \alpha^{(3)}_{N-1} \left( \ln \left( \frac{S_{N-1}}{J_{N-1}} \right) \right) \right] \\
\times \left( \frac{\phi_{N-1}}{J_{N-1}/S_{N-1}} \right)^3 + \cdots \right\} \quad (9)
\]

As the Taylor coefficients of \( \alpha_{N-1} \) are known from the term by term integrations of expressions (6) and (7), so are those of the \( \alpha^{(k)}_{N-1} \)'s. From the definition of \( \phi_{N-1} \) the quantity \( p_{v_{N-1}} \) can be expanded as:

\[
p_{v_{N-1}} = \frac{1}{1 + \frac{\phi_{N-1}}{i_{N-1}}} = 1 - \frac{\phi_{N-1}}{i_{N-1}} - \left( \frac{\phi_{N-1}}{i_{N-1}} \right)^2 - \left( \frac{\phi_{N-1}}{i_{N-1}} \right)^3 + \cdots \quad (10)
\]

assuming \( |\phi_{N-1}/i_{N-1}| < 1 \). Substituting equation (10) into (9) then multiplying term by term, and summing gives:

\[
V_{N-1} = F_{N-1} \sum_{j,k} f_{j,k} \left( \frac{S_{N-1}}{J_{N-1}} \right) \phi_{N-1}^{j} \left( \frac{S_{N-1}}{J_{N-1}} \right)^k \quad (11)
\]

as hoped. Equation (10) dominates the convergence properties, and, therefore, equation (11) converges only for \( p_{f_{N-1}}/p_{v_{N-1}} < 1 \).

### 2.2 The Induction Step

For the induction step, it is important to know how variables at time \( t_n \) depend on those same variables at \( t_{n-1} \). For instance:
\[ F_n = \left( p_{f_{n-1}} (1 + i_{n-1}) + p_{v_{n-1}} e^{\gamma_n} \right) F_{n-1} \]  
(12)

and

\[ \phi_n = \frac{1 + i_{n-1}}{e^{\gamma_n}} \left( \frac{i_n}{i_{n-1}} \right) \phi_{n-1}. \]  
(13)

Writing the integral for the ratchet put at time \( t_{n-1} \) gives:

\[
V_{n-1} = e^{-\gamma_{n-1}} \mathbb{E}_{S_{n-1}} \left[ V_n (\gamma_n) \right] \\
= e^{-\gamma_{n-1}} \mathbb{E}_{S_{n-1}} \left[ F_n \sum_{j,k}^{k \leq j} f_{jkn} \left( \ln \left( \frac{S_n}{J_n} \right) \right) \phi_j^k \left( \frac{S_n}{J_n} \right)^k \right] \\
= F_{n-1} e^{-\gamma_{n-1}} \int_{-\infty}^{\infty} [p_{f_{n-1}} (1 + i_{n-1}) + p_{v_{n-1}} e^{\gamma_n}] \\
\times \sum_{j,k}^{k \leq j} f_{jkn} \left( \ln \left( \frac{S_n}{J_n} \right) \right) \phi_j^k \left( \frac{S_n}{J_n} \right)^k \left( \frac{1 + i_{n-1}}{i_{n-1}} \right) j \left( \frac{i_n}{i_{n-1}} \right)^j \\
\times e^{-\left( \gamma_{n-1} - \mu_{n-1} \right)^2 / 2\sigma_{n-1}^2} dy_n. \]  
(14)

During the time interval \((t_{n-1}, t_n)\), the strike \( X_{n-1} \) remains the same. The adjusted strike, however, falls from \( J_{n-1} \) to \( J_{n-1} - \phi_{n-1} S_{n-1} \) because the fixed fund has risen the time interval \((t_{n-1}, t_n)\) and takes up a larger percentage of the strike \( X_{n-1} \). If \( X_{n-1} > F_{n-1} \) then the strike does not ratchet to a higher value and \( J_n = J_{n-1} - \phi_{n-1} S_{n-1} \) if \( S_n < J_{n-1} - \phi_{n-1} S_{n-1} \). If \( F_{n-1} > X_{n-1} \), then the strike does ratchet, and \( J_n = S_n \) when \( S_n > J_{n-1} - \phi_{n-1} S_{n-1} \). So:

\[
V_{n-1} = F_{n-1} e^{-\gamma_{n-1}} \sqrt{2\pi \sigma_{n-1}^2} \\
\times \left[ p_{f_{n-1}} (1 + i_{n-1}) \int_{-\infty}^{\ln(J_{n-1}/S_{n-1} - \phi_{n-1})} \sum_{j,k}^{k \leq j} f_{jkn} \left( \ln \left( \frac{e^{\gamma_n}}{J_{n-1}/S_{n-1} - \phi_{n-1}} \right) \right) \right] \\
\times \phi_{n-1}^j \left( \frac{1 + i_{n-1}}{i_{n-1}} \right)^j \left( J_{n-1}/S_{n-1} - \phi_{n-1} \right)^k e^{-k \gamma_n \mu_{n-1} \gamma_n \left( \frac{i_n}{i_{n-1}} \right)^j} e^{-\left( \gamma_{n-1} - \mu_{n-1} \right)^2 / 2\sigma_{n-1}^2} dy_n.
\]
Let us introduce a new dummy variable \( \delta \) such that \( \delta = 1 \) if an integral in equation (15) is a variable integral (i.e., the third and fourth integrals), and \( \delta = 0 \) if an integral in equation (15) is a fixed integral (i.e., the first and second integrals). For a given arbitrary \( j \) and \( k \), each of the integrals in equation (15) is proportional to:

\[
\begin{align*}
&\exp\left(-\mu_{n-1}/2\sigma_n^2\right) \phi_{n-1}^j \left(1 + i_{n-1}\right)^j \left(\frac{i_n}{i_{n-1}}\right)^j \\
&\times \int_{-\infty}^{e^{-\mu_{n-1}/2\sigma_n^2}} e^{\left(\frac{\mu_{n-1}}{\sigma_n^2} + \delta + k - j\right)\nu} e^{-\nu^2/2\sigma_n^2} f_{jkn}(\nu + \xi_{n-1}) e^{k\xi_{n-1}} d\nu
\end{align*}
\]

where \( \xi \) is defined in equation (5). Changing variables to \( \nu = \gamma_n + \xi_{n-1} \) in the integral in expression (16) and ignoring external constants gives:

\[
\begin{align*}
&\exp\left[-f_n(\delta + \mu_{n-1}/\sigma_n^2)\xi_{n-1} - \frac{\xi_{n-1}^2}{2}\right] e^{-\xi_{n-1}^2/2\sigma_n^2} \\
&\times \int_{-\infty}^{0} \exp\left[\frac{\xi_{n-1}^2}{\sigma_n^2} + \left(\frac{\mu_{n-1}}{\sigma_n^2} + \delta + k - j\right)\nu\right] f_{jkn}(\nu) e^{-\nu^2/2\sigma_n^2} d\nu.
\end{align*}
\]

Using the expansions:
\[
\exp \left[ \frac{\xi_{n-1} v}{\sigma^2_{n-1}} \right] = \sum_{k=0}^{\infty} a_k \left( \xi_{n-1} v \right)^k
\]

and

\[
\exp \left[ \left( \frac{\mu_{n-1}}{\sigma^2_{n-1}} + \delta + k - j \right) v \right] f_{jkn} (v) = \sum_{k=0}^{\infty} b_k v^k
\]

and multiplying these expansions term by term yields the following integral:

\[
\int_{-\infty}^{0} \left[ (a_0 b_0) + (a_1 b_0 \xi_{n-1} + a_0 b_1) v + (a_2 b_0 \xi^2_{n-1} + a_1 b_1 \xi_{n-1} + a_0 b_2) v^2 + \cdots \right] e^{-v^2/2\sigma^2_{n-1}} \, dv. \quad (18)
\]

The definite integral for each independent power of \(v\) can be found by integration by parts and the normal integral. This turns this integral into a power series in \(\xi_{n-1}\):

\[
\begin{align*}
a_0 & \left[ b_0 \left( \sqrt{\frac{\pi}{2}} \right) \sigma_{n-1} + b_1 \left( -\sigma^2_{n-1} \right) + b_2 \left( \sqrt{\frac{\pi}{2}} \right) \sigma^3_{n-1} + \cdots \right] \\
+ a_1 & \left[ b_0 \left( -\sigma^2_{n-1} \right) + b_1 \left( \sqrt{\frac{\pi}{2}} \right) \sigma^3_{n-1} + b_2 \left( -2\sigma^4_{n-1} \right) + \cdots \right] \xi_{n-1} \\
+ a_2 & \left[ b_0 \left( \sqrt{\frac{\pi}{2}} \right) \sigma^3_{n-1} + b_1 \left( -2\sigma^4_{n-1} \right) + b_2 \left( 3\sqrt{\frac{\pi}{2}} \right) \sigma^5_{n-1} + \cdots \right] \xi^2_{n-1} \\
+ \cdots \quad (19)
\end{align*}
\]

Unfortunately, the interior coefficients are sequences in \(b_n\). Several methods are available for accelerating the convergence of these terms, but it is still necessary to compute a substantial number of them. Expression (19) can then be multiplied term by term with the Taylor expansions of the external terms in expression (17). This gives:

\[
V_{n-1} = F_{n-1} \left[ p_{f_{n-1}} \left( 1 + i_{n-1} \right) \sum_{j,k}^{k \leq j} \phi^j_{n-1} g_{j,k,n-1,\delta=0} \left( \xi_{n-1} \right) \right]
\]
\[ + p_{\nu_{n-1}} \sum_{j,k}^{k \leq j} \phi_{n-1}^{j} g_{j,k,n-1,0,0} \left( \frac{1}{J_{n-1}/S_{n-1} - \phi_{n-1}} \right) \]

Some of the external coefficients have been pulled into the function \( g \) in equation (20).

The second and fourth integrals in equation (15) can be treated similarly:

\[ (1 + \delta_{n-1}) \left( \frac{i_{n}}{i_{n-1}} \right)^{j} \phi_{n-1}^{j} e^{-\mu_{n-1}^{2}/2\sigma_{n-1}^{2}} f_{j,kn}(0) \]

\[ \times \int_{-\xi_{n-1}}^{\infty} e^{(\delta_{-j} + \frac{\mu_{n-1}}{\sigma_{n-1}^{2}})_{n}^{2}} e^{-\frac{\epsilon_{n-1}^{2}}{2\sigma_{n-1}^{2}}} \, d\gamma_{n}. \]  

The integral in expression (21) can be separated into an integral from 0 to \( \infty \) and an integral from \( -\xi_{n-1} \) to 0. The exponential functions inside the integral from \( -\xi_{n-1} \) to 0 can be expanded, multiplied term by term, and finally integrated term by term analogous to the procedure introduced above. The constant of integration is obtained by calculating the integral from 0 to \( \infty \) exactly:

\[ \sqrt{2\pi\sigma_{n-1}^{2}} e^{(\mu_{n-1}^{2} + \sigma_{n-1}^{2}(\delta - j))/2\sigma_{n-1}^{2}} \mathcal{N} \left( \frac{\mu_{n-1}}{\sigma_{n-1}} + \delta - j \right), \]

finally giving:

\[ V_{n-1} = F_{n-1} \left\{ p_{f_{n-1}} \left( 1 + i_{n-1} \right) \right\} \]

\[ \times \sum_{j,k}^{k \leq j} \phi_{n-1}^{j} \left[ g_{j,k,n-1,0,0} \left( \ln \left( \frac{1}{J_{n-1} - \phi_{n-1}} \right) \right) \right] \]

\[ + h_{j,k,n-1,0,0} \left( \ln \left( \frac{1}{J_{n-1}/S_{n-1} - \phi_{n-1}} \right) \right) \]

\[ + p_{\nu_{n-1}} \sum_{j,k}^{k \leq j} \phi_{n-1}^{j} \left[ g_{j,k,n-1,1,1} \left( \ln \left( \frac{1}{J_{n-1}/S_{n-1} - \phi_{n-1}} \right) \right) \right] \]

\[ + h_{j,k,n-1,1,1} \left( \ln \left( \frac{1}{J_{n-1}/S_{n-1} - \phi_{n-1}} \right) \right) \].
We now define the function \( G = g + h \), which can be expanded similarly to the method in equations (8) to (9) yielding:

\[
V_{n-1} = F_{n-1} \left\{ p_{f_{n-1}} \left( 1 + i_{n-1} \right) \sum_{j,k}^{k \leq j} \phi_{n-1}^{j} \left[ G_{j,k,n-1,0} \left( \ln \left( \frac{S_{n-1}}{J_{n-1}} \right) \right) + \frac{1}{2} \left( G_{j,k,n-1,0}^{(1)} + G_{j,k,n-1,0}^{(2)} \right) \left( \frac{\phi_{n-1}}{J_{n-1}/S_{n-1}} \right)^2 + \cdots \right] \\
+ \sum_{j,k}^{k \leq j} \phi_{n-1}^{j} \left[ G_{j,k,n-1,1,1}^{(1)} \left( \ln \left( \frac{S_{n-1}}{J_{n-1}} \right) \right) + \frac{1}{2} \left( G_{j,k,n-1,1,1}^{(1)} + G_{j,k,n-1,1,1}^{(2)} \right) \left( \frac{\phi_{n-1}}{J_{n-1}/S_{n-1}} \right)^2 + \cdots \right] \right\}
\]

where the superscript \((k)\) denotes \( k \)th order differentiation. We can then group like terms as:

\[
V_{n-1} = F_{n-1} \left\{ p_{f_{n-1}} \left( 1 + i_{n-1} \right) \right. \\
\left. + \sum_{j,k}^{k \leq j} Q_{j,k,n-1,0} \left( \ln \left( \frac{S_{n-1}}{J_{n-1}} \right) \right) \left( \frac{(\phi_{n-1})^j}{(J_{n-1}/S_{n-1})^k} \right) \right. \\
+ \sum_{j,k}^{k \leq j} Q_{j,k,n-1,1,1} \left( \ln \left( \frac{S_{n-1}}{J_{n-1}} \right) \right) \left( \frac{(\phi_{n-1})^j}{(J_{n-1}/S_{n-1})^k} \right) \right\}
\]

where
\[ Q_{j,0,n-1,\delta} = \sum_{k=0}^{j} G_{j,k,n-1,\delta} \]
\[ Q_{j,1,n-1,\delta} = Q_{j-1,0,n-1,\delta}^{(1)} \]
\[ Q_{j,2,n-1,\delta} = \frac{1}{2} \left( Q_{j-2,0,n-1,\delta}^{(1)} + Q_{j-2,0,n-1,\delta}^{(2)} \right) \]
\[ Q_{j,3,n-1,\delta} = \frac{1}{3} Q_{j-3,0,n-1,\delta}^{(1)} + \frac{1}{2} Q_{j-3,0,n-1,\delta}^{(2)} + \frac{1}{6} Q_{j-3,0,n-1,\delta}^{(3)} \]

etc. Finally, we substitute equation (10) into (24), which leads to:

\[
V_{n-1} = F_{n-1} \left[ \left( \frac{\phi_{n-1}}{i_{n-1}} - \frac{\phi_{n-1}^2}{i_{n-1}^2} + \frac{\phi_{n-1}^3}{i_{n-1}^3} - \cdots \right) (1 + i_{n-1}) \right. \\
\times \left( \sum_{k \leq j} \sum_{j,k} Q_{j,k,n-1,0} \left( \ln \left( \frac{S_{n-1}}{J_{n-1}} \right) \right) \frac{(\phi_{n-1})^j}{(J_{n-1}/S_{n-1})^k} \right) \\
+ \left( 1 - \frac{\phi_{n-1}}{i_{n-1}} + \frac{\phi_{n-1}^2}{i_{n-1}^2} - \frac{\phi_{n-1}^3}{i_{n-1}^3} + \cdots \right) \\
\times \left( \sum_{k \leq j} \sum_{j,k} Q_{j,k,n-1,1} \left( \ln \left( \frac{S_{n-1}}{J_{n-1}} \right) \right) \frac{(\phi_{n-1})^j}{(J_{n-1}/S_{n-1})^k} \right). \tag{25} \]

This yields as a final result through interchange of summation indices:

\[
V_{n-1} = F_{n-1} \sum_{k \leq j} f_{j,k,n-1} \left( \ln \left( \frac{S_{n-1}}{J_{n-1}} \right) \right) \phi_{n-1} \left( \frac{S_{n-1}}{J_{n-1}} \right)^k \tag{26} \]

where

\[
f_{j,k,n-1} = \sum_{l=k}^{j} (-1)^{(l-j)} \frac{Q_{l,k,n-1,1}}{(i_{n-1})^{(j-l)}} \\
+ \sum_{l=k;k<j}^{j-1} (-1)^{(l-j+1)} \frac{Q_{l,k,n-1,0}}{(i_{n-1})^{(j-l)}} (1 + i_{n-1}). \]

Once the function can be shown to be of the form:

\[
V_n = F_n \sum_{j,k}^{k \leq j} f_{j,k,n} \left( \ln \left( \frac{S_{n-1}}{J_{n-1}} \right) \right) \phi_{n-1} \left( \frac{S_{n-1}}{J_{n-1}} \right)^k. \]
the final term exponential term can be expanded, multiplied term by term, and then summed over \( k \) to give:

\[
V_n (\xi_n, \phi_n) = F_n \sum_{j} f_{j,k} (\xi_n) \phi_n^j \tag{28}
\]

where the functions \( f_{j,k} \) result from the Taylor series expansions and from the functions \( f_{j,n} \) and, therefore, are known.

The biggest practical issue in the determination of the coefficients is the summation of the series in equation (19). We found that keeping 128 terms and using several iterations of Euler's method followed by one iteration of Levin's method from Fessler, Ford, and Smith (1983) worked best. Accuracy of the coefficients, and, therefore, of the final result, might be improved further by the use of infinite precision arithmetic.

3 Comparison to Monte Carlo Results

The first issue that needs to be addressed when comparing the results of the Taylor expansion to the results of Monte Carlo simulation is the issue of convergence of the series. As each step involves only functions that are analytic everywhere, it seems reasonable that the final function for the lookback value should also be analytic everywhere. In practice, however, the convergence range might be limited. In nearly all cases, the agreement is poor outside of the range \( 0.5 < S/X < 2.0 \), and, in many cases, the effective range of convergence is even smaller than this.

3.1 Increasing the Range of Convergence to \( +\infty \)

The convergence radius can be extended to infinity on both ends by considering the limits. The limit \( \xi \to +\infty \) corresponds to an extremely low strike. In this case, the strike is nearly certain to ratchet to the fund value at time \( t_1 \). This implies a value at time \( t_0 \) of:

\[
V_0 (\infty, \phi) = e^{-r t_1} \int_{-\infty}^{\infty} F_1 \sum_{j} \phi_1^j f_{j,1} (0) \frac{e^{-\left(\gamma_1 - \mu_0\right)^2/2\sigma_0^2}}{\sqrt{2\pi\sigma_0^2}} \, d\gamma_1. \tag{29}
\]

As \( F_1 \) and \( \phi_1 \) are defined in equations (12) and (13), this gives:
\[ V_0(\infty, \phi) = F_0 \left[ e^{-qt_1} \sum_j f_{j1}(0) \phi_j^j \right. \]
\[ \times \int_{-\infty}^{\infty} \left( p_f (1 + i_0) + p_v e^{\gamma_1} \right) \frac{(1 + i_0)^j}{e^{j\gamma_1}} \left( \frac{i_1}{i_0} \right)^j \frac{e^{-(\gamma_1 - \mu_0)^2/2\sigma_0^2}}{\sqrt{2\pi\sigma_0^2}} \, d\gamma_1 \right]. \]

(30)

The integral can be evaluated directly. Remembering to expand \( p_f \) and \( p_v \) as in equation (11) and multiplying term by term gives:

\[ V_0(\infty, \phi) = F_0 H(\phi) = F_0 \left( H_0 + H_1 \phi + H_2 \phi^2 + H_3 \phi^3 + \cdots \right) \]

(31)

where

\[ H_0 = e^{-qt_1} f_{01}(0) \]
\[ H_1 = \left[ e^{(g-r)t_1} \right] \frac{f_{01}(0)}{i_0} + e^{-qt_1} f_{11}(0) (1 + i_0) \left( \frac{i_1}{i_0} \right) e^{-\mu_0 e^{-\sigma_0^2/2}} \]
\[ H_2 = \left[ e^{-(g-r)t_1} \right] \frac{f_{01}(0)}{i_0} + e^{-qt_1} f_{21}(0) (1 + i_0)^2 \left( \frac{i_1}{i_0} \right)^2 e^{-2\mu_0} \]
\[ + \left[ e^{(g-r+\sigma^2)t_1} \right] f_{11}(0) (1 + i_0) \left( \frac{i_1}{i_0} \right) e^{-\mu_0 e^{-\sigma_0^2/2}} \]

and

\[ H_3 = \left[ e^{(g-r)t_1} \right] \frac{f_{01}(0)}{i_0} \]
\[ + \left[ e^{qt_1} - e^{(g-r+\sigma^2)t_1} \right] \frac{f_{11}(0)}{i_0} (1 + i_0) \left( \frac{i_1}{i_0} \right) e^{-\mu_0 e^{-\sigma_0^2/2}} \]
\[ + \left[ e^{(g-r+2\sigma^2)t_1} \right] f_{21}(0) (1 + i_0)^2 \left( \frac{i_1}{i_0} \right)^2 e^{-2\mu_0} \]
\[ + e^{-qt_1} f_{31}(0) (1 + i_0)^3 \left( \frac{i_1}{i_0} \right)^3 e^{-3\mu_0 e^{-\sigma_0^2/2}}. \]

To extend the convergence of the function to the entire range \( 0 \leq \xi < \infty \), we can create a multipoint Pade approximant [see Baker and Graves-Morris (1996, pages 335-361)] for the functions \( f_j(\xi) - H_j \). A multipoint Pade approximant is a rational function of two polynomials whose
values and derivatives agree to given, possibly different, orders at two different points. The points chosen at this stage are $\xi = 0$ and $\xi = +\infty$.

The Taylor expansions are known at $\xi = 0$, and the limiting behavior must decay to 0 as $\xi \to +\infty$. It seems likely that the asymptotic expansion near $+\infty$ should be 0 as the probabilities depend on the cumulative normal distribution that has this asymptotic expansion. The approximant that has these features is simply the reciprocal of the Taylor expansion of the reciprocal, and this is the function we will use for comparison purposes in the range $0 \leq \xi < \infty$.

It is true that in the limit $\xi \to -\infty$, $\phi / (J/S)$ is no longer small, as was assumed in the derivation of equation (9) and equation (23). This difficulty is solved in the same manner, by creating a multipoint Pade approximation in $\phi$ at $\phi = 0$ and $\phi = \infty$ from the coefficients in equation (28).

### 3.2 Increasing the Range of Convergence to $-\infty$

Now, the limit $\xi \to -\infty$ corresponds to an extremely high strike. In this case, the GMDB is unlikely to ratchet, and the value of the option approaches the value of a simple Black-Scholes put. The Taylor expansion of this put is available by the same methods used in Section 2.1 with $t_1 = t_N$, the maturity of the option. Subtracting this expansion from the full Taylor expansion, an expansion for the value of the ratchet alone can be obtained. This value should drop to 0 as $\xi \to -\infty$ for the same reasons as the $\xi \to +\infty$ limit. The solution in the range $-\infty < \xi < 0$ will, therefore, be the sum of the exact Black-Scholes calculation and the multipoint Pade approximant for the excess contribution of the ratchet. Figure 1 shows a comparison of these values to Monte Carlo simulations using 32,000 antithetic scenarios. We typically use a risk-free rate of 5%, risk fees of 115 bp consistent with the market survey in Milevsky and Posner (2001), a stock market volatility of 20%, a fixed fund return of 5%, and time between ratchets (excluding the first and last) of 12 months. We use terms up to $\xi^{64}$ and $\phi^2$ and fixed fund percentages from 0% to 90% fixed in 10% increments. A range of strikes near 1 is presented to highlight the agreement, which is good, though, over the entire range $-\infty < \xi < 0$.

The final issue is how to connect the two approximations at $\xi = 0$. The two functions are discontinuous if $\phi \neq 0$. Strikes greater than 1 tend to produce greater agreement near $\xi = 0$, so the approximation for strikes less than one are scaled to this value. Figure 2 shows the comparison at low strike prices after the results have been rescaled.
Figure 1: Comparison of Simulation and Taylor Expansion of a 24 Month Lookback Option with the First Ratchet in 12 Months at High Strike Prices

Figure 2: Comparison of Simulation and Taylor Expansion of a 24 Month Lookback Option with the First Ratchet in 12 Months at Low Strike Prices
Accuracy in the calculation of the value of the option is not the only consideration in evaluating the approximation. It is frequently necessary to calculate the value of various derivatives of the function in order to hedge the option. Figures 3 and 4 show the comparison of delta with the Monte Carlo values. For low strikes, delta is positive because the option is likely to ratchet at its next opportunity. The higher the stock market, the higher the value to which the strike will ratchet. On the other hand, for high strikes, delta is negative, because the ratchet is less relevant and the lookback delta approaches the delta of a simple put.

![Figure 3: Comparison of Simulation and Taylor Expansion of the Value of Delta for a 24 Month Lookback Option with the First Ratchet in 12 Months](image)

Finally, we need to show the results of a full calculation of the ratchet value for all ranges of strikes using a series of puts with maturities from zero to 96 months weighted by their expected exercise probability in the double decrement model. We use a mortality rate equal to that in the 2000 GAM table and lapse percentages that rise from 0% in year one to a maximum of 22.5% in year nine and settling in at a long-term rate of 19.5%. Instead of weighting the values of the individual lookback puts, a weighted average of the Taylor expansions is used as the Taylor expansion of the full GMDB value. This results in a time savings of a factor of 100 or so and is a feature of the Taylor expansion method that cannot be duplicated by the simulation or analytic methods. The value of the GMDB is then calculated from this new expansion in a manner
similar to that of the individual puts. Figure 5 shows comparisons of the approximation to the simulated values in a range of strike values near 1. Similar agreement is obtained over the entire range. Figure 6 shows comparisons of delta to the simulated values. The values of delta show minor disagreements in the range of strikes between 1.5 and 2 on the order of perhaps a few basis points. We would consider this disagreement to be minor, as it would be dwarfed by any errors in assumptions of risk-free rate, volatility, lapse rates, and mortality used in the approximation.

3.3 Time Comparisons

There are at least two advantages of using the Taylor expansion instead of Monte Carlo simulation. The first is that the expansion is more accurate in most cases because simulation contains random errors. This is by no means true in all cases, particularly those at high fixed fund percentages, as can be seen in the figures. The major advantage, however, is the increase in speed obtained by using a function that can be quickly evaluated, rather than performing a large number of simulations. To demonstrate this, we compared the results of 69,120 calculations of both the approximation and 32,000 antithetic Monte Carlo scenarios. We used ages ranging from 55 to 66, durations rang-
Figure 5: Comparison of Simulation and Taylor Expansion of a GMDB Ratchet Option on a 55 Year Old One Year After Issue

Figure 6: Comparison of Simulation and Taylor Expansion of the Value of Delta for a GMDB Ratchet Option on a 55 Year Old One Year After Issue
ing from one to 96 months, fixed percentages ranging from 0% to 50%
in steps of 10%, and ten different strike values. The computations were
performed on a Dell Computer with 2.8GHz Pentium processor. It took
40.203 seconds for the 69,120 Taylor series expansions to be computed
for an average of 0.582 milliseconds per point. It took 199,606 seconds
(2 days, 9 hours) to complete the 69,120 Monte Carlo simulations for
an average of 2.888 seconds per point. This shows the approximation
will improve runtimes by a factor of about 5,000 if 32,000 scenarios
are used. Equivalently, the approximation runs about as fast as six to
seven Monte Carlo runs.

4 Comparisons to the Analytic Solution

Now that we have shown the superiority of the Taylor expansion
method over the most commonly used method of simulation, we wish
to show its superiority over the analytic solution in Collin-Dufresne,
Keirstad, and Ross (1997), hereafter referred to as CKR, in both speed
and accuracy. First, the formulas will be reproduced here to prevent
unnecessary cross-referencing. The price of a single discrete lookback
put is given as:

\[ P = V \left( S_0, X_0, t_N, \sigma, r, q, t_1, \cdots, t_N \right) - S_0 e^{-q t_N} + X_0 e^{-r t_N} \]  (32)

where

\[ V \left( S_0, X_0, t_N, \sigma, r, q, t_1, \cdots, t_N \right) = \sum_{j=1}^{N} H_{j} I_{n-j} S_0 e^{(r-q) t_j - r t_N} \]

\[ - X_0 \left[ 1 - \mathcal{N}_N \left( -d^Q (X_0, t_1), \cdots, -d^Q (X_0, t_N); \{ C_{ik}^{(1)} \} \right) \right], \]

\[ H_j = \mathcal{N}_j \left( \frac{\mu + \sigma^2}{\sigma} \sqrt{t_j - t_1}, \cdots, \frac{\mu + \sigma^2}{\sigma} \sqrt{t_j - t_{j-1}}, d^R (X_0, t_j); \{ C_{ik}^{(2)} \} \right) \]

\[ I_{N-j} = \mathcal{N}_{N-j} \left( -\frac{\mu}{\sigma} \sqrt{t_{j+1} - t_j}, \cdots, -\frac{\mu}{\sigma} \sqrt{t_{j+N} - t_j}; \{ C_{ik}^{(1)} \} \right) \]

\[ d^Q (X_0, t_j) = \frac{\log (S_0 / X_0) + \mu t_j}{\sigma \sqrt{t_j}} \]
\[ d^R(X_0, t_j) = d^Q(X_0, t_j) + \sigma \sqrt{t_j} \]

and the covariance matrices are:

\[
C_{ik}^{(1)} = \sqrt{\frac{t_{i \wedge k}}{t_{i \vee k}}} \\
C_{ik}^{(2)} = \sqrt{\frac{t_j + (i \wedge k) - t_j}{t_j + (i \vee k) - t_j}} \\
C_{ik}^{(3)} = \sqrt{\frac{t_j - t_{i \vee k}}{t_j - t_{i \wedge k}}}, \quad i, k \neq j \\
C_{ij}^{(3)} = C_{ji}^{(3)} = \sqrt{\frac{t_j - t_i}{t_j}}, \quad i \neq j.
\]

The evaluation of one lookback put value requires the evaluation of \(2N + 1\) cumulative multivariate normal functions with an average of \(N/2\) variables. They are evaluated in the comparison by the method of Somerville (1998a, 1998b). This by itself can be time-consuming. Using 10,000 directions, a point with \(N = 2\) requires on average 92 milliseconds to compute and a point with \(N = 8\) requires 1.52 seconds on the 2.8GHz Dell. This should be compared with about 0.5 milliseconds per point for the Taylor expansions. Figure 7 shows comparisons of the values obtained from the Taylor expansion with the CKR results for a 67 month lookback put 5 months after issue. No comparisons were done for any fixed fund percentages other than 0%, as the CKR function doesn't apply to this case. The Taylor expansion agrees well with the exact solution, and the disagreements are primarily due to random errors in the evaluation of the multivariate normal functions. These errors could be reduced if more directions were used in the CKR evaluation, but this would increase the time for a method that is already much slower than the Taylor expansion.

Next, we use the CKR formula to evaluate a complete ratchet GMDB. In the case of the Taylor expansion, the coefficients could be weighted and summed prior to evaluation, which results in a substantial time savings. There is no comparable procedure for the CKR formula. The value of the lookback put must be found every month and then multiplied by the probabilities of death and summed. A comparison of results for a 55 year old one month after issue is found in Figure 8. The agreement in values is better than for the individual puts because the random errors in each put value have a tendency to cancel out. The evaluation time, however, has grown enormously. The average time to compute one point is 54.9 seconds, compared with 0.582 milliseconds for the
References


Credibility Theory and Geometry

Elias S.W. Shiu* and Fuk Yum Sing†

Abstract‡

We present a geometric approach to studying greatest accuracy credibility theory. Our main tool is the concept of orthogonal projections. We show, for example, that to determine the Bühlmann credibility premium is to find the coefficients of the minimum-norm vector in an affine space spanned by certain orthogonal random variables. Our approach is illustrated by deriving various common credibility formulas. Several equivalent forms of the credibility factor $Z$ are derived by means of similar triangles.

Key words and phrases: greatest accuracy credibility theory, Bühlmann credibility premium, credibility factor, affine space, inner product, orthogonal projection, Bühlmann-Straub model

*Elias S. W. Shiu, Ph.D., A.S.A., is Principal Financial Group Foundation Professor of actuarial science at the University of Iowa, and visiting Chair Professor of actuarial science in the Department of Applied Mathematics, Hong Kong Polytechnic University, China. He received a Ph.D. in mathematics from the California Institute of Technology in 1975. From 1976 to 1991, he was a professor of actuarial science at the University of Manitoba and a consultant for the Great-West Life Assurance Company in Winnipeg, Canada. He is an editor of Insurance: Mathematics and Economics and a co-editor of the North American Actuarial Journal.

Dr. Shiu’s address is: Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa 52242-1409, U.S.A. E-mail: eshiu@stat.uiowa.edu

†Fuk Yum Sing, M. Phil., is a visiting lecturer in the Department of Applied Mathematics, The Hong Kong Polytechnic University. He received his master’s degree in mathematics from the University of Hong Kong in 1977. He joined the then Hong Kong Polytechnic (now The Hong Kong Polytechnic University) in 1976.

Mr. Sing’s address is: Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, CHINA.

‡The authors thank the anonymous referees for their insightful comments. Elias Shiu gratefully acknowledges the generous support from the Principal Financial Group Foundation and Robert J. Myers, F.C.A., F.C.A.S., F.S.A.
1 Introduction

Credibility theory, which is called a cornerstone of actuarial science by some authors (Longley-Cook 1962, page 194; Hickman and Heacox 1999, page 1), is a required part of education syllabi of major international professional organizations including the Society of Actuaries, the Institute and Faculty of Actuaries, and the Casualty Actuarial Society. One of the texts recommended by the Society of Actuaries for studying credibility theory is Klugman, Panjer, and Willmot (1998). This text uses a traditional probability/statistics approach to derive credibility formulas. The main purpose of this paper is to present a geometric approach to derive and extend some of the results in Klugman, Panjer, and Willmot (1998, Sections 5.4.2, 5.4.3, and 5.4.4).

The main tool used in this paper is the concept of orthogonal projections. Background materials on the inner product, affine space, and inner product space of square-integrable random variables are presented in Section 2. The assumption of a risk parameter $\theta$, conditional on which the claims $\{X_j\}$ are independent, implies that the random variables $\{X_j - E[X_j|\theta]\}$ can be viewed as orthogonal vectors. Section 3 shows that to determine the credibility premium is to find the coefficients of the vector with the smallest length in an affine space containing these orthogonal vectors. With the expressions for the optimal coefficients, Section 4 derives various credibility formulas in the Klugman, Panjer, and Willmot textbook. For some readers, Section 5 may be the most intriguing section in this paper. By means of similar triangles, it derives various equivalent forms of the credibility factor $Z$. Section 6 presents several more interesting formulas.

There are many books and survey articles on credibility theory including: Buhlmann (1970), Kahn (1975), Goovaerts and Hoogstad (1987), Heilmann (1988), Straub (1988), Goovaerts et al. (1990), Venter (1990), Sundt (1993), Waters (1993), Goulet (1998), Klugman, Panjer, and Willmot (1998), Herzog (1999), Kaas et al. (2001), and Mahler and Dean (2001). These authors use probability theory and other tools to develop and explain credibility formulas and concepts. This paper's approach, which de-emphasizes probability theory, may be more appealing to some actuarial practitioners and students.
2 Some Mathematical Preliminaries

2.1 Inner Product Space and Orthogonal Projections

An inner product space is a vector space $V$ (over the real numbers) together with an inner product (also called scalar product or dot product) defined on $V \times V$. Corresponding to each pair of vectors $u$ and $v$ in $V$, the inner product $\langle u, v \rangle$ is a real number. The inner product satisfies the following axioms:

1. $\langle u, v \rangle = \langle v, u \rangle$;
2. $\langle cu, v \rangle = c \langle u, v \rangle$ for each real number $c$;
3. $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$;
4. $\langle u, u \rangle \geq 0$, and $\langle u, u \rangle = 0$ if and only if $u = 0$, the zero vector.

The norm (or length) of a vector $u$ is $\|u\| = \sqrt{\langle u, u \rangle}$. For each pair of nonzero vectors $u$ and $v$, the quantity $\langle u, v \rangle / (\|u\| \|v\|)$ can be interpreted as the cosine of the angle between $u$ and $v$. If $\langle u, v \rangle = 0$, we say that the vectors are orthogonal and we write $u \perp v$. Because

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2 + 2 \langle u, v \rangle,$$

the vectors $u$ and $v$ are orthogonal if and only if the Pythagorean equation holds:

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2.$$

Let $U$ be a subspace of an inner product space $V$ and $v$ be an arbitrary vector in $V$. We are interested in finding the vector $u$ in $U$ closest to $v$ in the sense that it minimizes the norm $\|v - u\|$. It is not difficult to show (Luenberger 1969, page 50, Theorem 1) that, if there is $u_0 \in U$ such that

$$\|v - u_0\| \leq \|v - u\| \quad \text{for all} \quad u \in U,$$

then $u_0$ is unique. Furthermore, a necessary and sufficient condition that $u_0 \in U$ is a unique minimizing vector in $U$ is the following:

$$(v - u_0) \perp u \quad \text{for all} \quad u \in U.$$  \hfill (1)

It is easy to see that two conditions, each of which is equivalent to condition (1), are
\[ \langle v, u \rangle = \langle u_0, u \rangle \quad \text{for all } u \in U \] (2)

and

\[ \|v - u\|^2 = \|v - u_0\|^2 + \|u_0 - u\|^2 \quad \text{for all } u \in U. \] (3)

The vector \( u_0 \) is called the \textit{orthogonal projection of} \( v \) onto \( U \).

Consider the special case where \( U \) is a one-dimensional subspace spanned by a nonzero vector \( u^* \). Then it follows from equation (2) that the vector \( u_0 \) is

\[ \frac{\langle v, u^* \rangle}{\langle u^*, u^* \rangle} u^* = \left( \frac{v}{\|v\|}, \frac{u^*}{\|u^*\|} \right) \|v\| \frac{u^*}{\|u^*\|}. \] (4)

With the inner product on the right side of equation (4) being interpreted as the cosine of the angle between the vectors \( v \) and \( u^* \), the geometric explanation of the left side of equation (4) is obvious.

### 2.2 Vector with Minimal Norm in an Affine Space

Let \( v_1, v_2, \ldots, v_m \) be \( m \) vectors in a vector space \( V \). The affine space (also called affine set or linear variety) spanned by these vectors is the set of vectors of the form \( \sum_{j=1}^{m} c_j v_j \) with real coefficients \( c_1, c_2, \ldots, c_m \) satisfying

\[ \sum_{j=1}^{m} c_j = 1. \] (5)

There is no restriction on the sign of the coefficients. Assuming \( V \) is an inner product space and the \( m \) vectors are nonzero and mutually orthogonal, we claim the vector

\[ w = \sum_{j=1}^{m} \hat{c}_j v_j, \] (6)

with

\[ \hat{c}_j = \frac{1}{\sum_{k=1}^{m} \frac{1}{\|v_k\|^2}} \] \( j = 1, 2, \ldots, m, \) (7)
is the vector having the minimal norm in the affine space spanned by \( \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \). To see this, we use the assumption that the vectors \( \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \) are mutually orthogonal to obtain

\[
\| \sum_{j=1}^{m} c_j \mathbf{v}_j \|^2 = \sum_{j=1}^{m} c_j^2 \| \mathbf{v}_j \|^2,
\]

(8)

which is called Parseval’s identity. The optimal coefficients \( \{c_j\} \) are then determined by minimizing the right side of equation (8) subject to the constraint of equation (5). This optimization problem can be readily solved using the method of Lagrange multipliers, and the solution is the system of equations (7).

It follows from equations (6), (7), and (8) that

\[
\| \mathbf{w} \|^2 = \frac{1}{\sum_{k=1}^{m} \| \mathbf{v}_k \|^2}.
\]

(9)

Equation (9) shows that \( \| \mathbf{w} \|^2 \) is \( 1/m \) of the harmonic mean of \( \| \mathbf{v}_1 \|^2, \| \mathbf{v}_2 \|^2, \ldots, \| \mathbf{v}_m \|^2 \).

An alternative approach to deriving the system of equations (7) is to show that \( \mathbf{w} \) is the vector of minimal norm in an affine space iff

\[
\mathbf{w} \perp (\mathbf{v} - \mathbf{w})
\]

(10)

for all vectors \( \mathbf{v} \) in the affine space. For further discussion, see Luenberger (1969, page 64).

2.3 Inner Product Space of Random Variables

For a given sample space, the set of square-integrable random variables (random variables with finite variance) forms an inner product space (Luenberger, 1969; Small and McLeish, 1994). For each pair of square-integrable random variables \( X \) and \( Y \), the inner product is defined to be \( \langle X, Y \rangle = \mathbb{E}[XY] \).

Let \( g \) be a function such that \( g(Y) \) is a square-integrable random variable. Then, by the law of iterated expectations,

\[
\langle X, g(Y) \rangle = \mathbb{E}[Xg(Y)]
\]

\[
= \mathbb{E}[\mathbb{E}[Xg(Y)|Y]]
\]

\[
= \mathbb{E}[\mathbb{E}[X|Y]g(Y)]
\]

\[
= \langle \mathbb{E}[X|Y], g(Y) \rangle.
\]
Hence, \((X - \mathbb{E}[X|Y]) \perp g(Y)\), and we have the Pythagorean equation:

\[
\|X - g(Y)\|^2 = \|X - \mathbb{E}[X|Y]\|^2 + \|\mathbb{E}[X|Y] - g(Y)\|^2.
\] (11)

The conditional expectation \(\mathbb{E}[X|Y]\) is the orthogonal projection of \(X\) onto the subspace of square-integrable functions of \(Y\). Note that, by the law of iterated expectations,

\[
\|X - \mathbb{E}[X|Y]\|^2 = \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y]] = \mathbb{E}[\text{Var}(X|Y)].
\] (12)

If \(g(Y)\) is the constant random variable that takes the value \(\mathbb{E}[X]\), i.e., if \(g(Y) = \mathbb{E}[X]\), then equation (11) is the well-known variance decomposition equation

\[
\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}[\mathbb{E}[X|Y]].
\] (13)

The above can be generalized in various ways. In particular, we have Exercise 5.83(a) in Klugman, Panjer, and Willmot (1998):

\[
\|X - g(X)\|^2 = \|X - \mathbb{E}[X|X]\|^2 + \|\mathbb{E}[X|X] - g(X)\|^2
\] (14)

where \(X\) denotes the random variables \(X_1, X_2, ..., X_n\). Also, equations (11), (12), and (13) can be generalized as

\[
\langle W - f(Y), X - g(Y) \rangle = \langle W - \mathbb{E}[W|Y], X - \mathbb{E}[X|Y] \rangle + \langle \mathbb{E}[W|Y] - f(Y), \mathbb{E}[X|Y] - g(Y) \rangle,
\]

\[
\langle W - \mathbb{E}[W|Y], X - \mathbb{E}[X|Y] \rangle = \mathbb{E}[\text{Cov}(W, X|Y)]
\] (15)

and

\[
\text{Cov}(W, X) = \mathbb{E}[\text{Cov}(W, X|Y)] + \text{Cov}[\mathbb{E}[W|Y], \mathbb{E}[X|Y]],
\]

respectively.

### 3 Greatest Accuracy Credibility Theory

Following Klugman, Panjer, and Willmot (1998, Chapter 5), let \(X_j\) denote the claim amount in the \(j\)th period, \(j = 1, 2, 3, ...\). In greatest accuracy credibility theory the objective is to determine the coefficients...
\( \alpha_0, \alpha_1, \ldots, \alpha_n \) of the credibility premium for period \((n + 1)\) given the losses in the previous \(n\) periods,

\[
P_{n+1} = \alpha_0 + \sum_{j=1}^{n} \alpha_j X_j
\]

so that the norm

\[
\|X_{n+1} - P_{n+1}\|
\]

is minimized. Because \(P_{n+1}\) is a function of the random variables \(X_1, X_2, \ldots, X_n\), we have a special case of equation (14):

\[
\|X_{n+1} - P_{n+1}\|^2 \geq \|X_{n+1} - \mathbb{E}[X_{n+1}|X_1, X_2, \ldots, X_n]\|^2 \\
+ \|\mathbb{E}[X_{n+1}|X_1, X_2, \ldots, X_n] - P_{n+1}\|^2.
\]

Hence, the credibility premium \(P_{n+1}\) can be determined by minimizing

\[
\|\mathbb{E}[X_{n+1}|X_1, X_2, \ldots, X_n] - P_{n+1}\|,
\]

which is not a surprising result.

As in Section 5.4 of Klugman, Panjer, and Willmot (1998), we assume the existence of a risk parameter random variable \(\Theta\), conditional on which the random variables \(X_1, X_2, \ldots, X_j, \ldots\) are independent. We write

\[
\mu_j(\Theta) = \mathbb{E}[X_j|\Theta].
\]

Thus,

\[
\mu_{n+1}(\Theta) = \mathbb{E}[X_{n+1}|\Theta] = \mathbb{E}[X_{n+1}|\Theta, X_1, X_2, \ldots, X_n]
\]
because of the conditional independence assumption. By the law of iterated expectations,

\[
\mathbb{E}[\mu_{n+1}(\Theta)|X_1, X_2, \ldots, X_n] = \mathbb{E}[\mathbb{E}[X_{n+1}|\Theta, X_1, X_2, \ldots, X_n]|X_1, X_2, \ldots, X_n]
\]

\[
= \mathbb{E}[X_{n+1}|X_1, X_2, \ldots, X_n].
\]

This shows that expression (19) is the same as

\[
\|\mathbb{E}[\mu_{n+1}(\Theta)|X_1, X_2, \ldots, X_n] - P_{n+1}\|.
\]

Similar to equation (18), we have
\[ \| \mu_{n+1}(\Theta) - P_{n+1} \|^2 = \| \mu_{n+1}(\Theta) - \mathbb{E} [\mu_{n+1}(\Theta) | X_1, X_2, \ldots, X_n] \|^2 \]
\[ + \| \mathbb{E} [\mu_{n+1}(\Theta) | X_1, X_2, \ldots, X_n] - P_{n+1} \|^2. \] (20)

Therefore, an alternative way to determine the credibility premium is to minimize

\[ \| \mu_{n+1}(\Theta) - P_{n+1} \|. \] (21)

By equation (15),

\[ \langle X_j - \mu_j(\Theta), X_k - \mu_k(\Theta) \rangle = \mathbb{E} [\text{Cov}(X_j, X_k | \Theta)], \]

which is zero because of the conditional independence assumption. Hence, the random variables \{X_j - \mu_j(\Theta)\} are mutually orthogonal. This fact will play a key role in determining the credibility premium.

We now follow Klugman, Panjer, and Willmot (1998, Section 5.4) and assume that \( \mu_j(\Theta) = \mu(\Theta) \) for \( j = 1, 2, 3, \ldots \), and write \( \mathbb{E} [\mu(\Theta)] = \mu \).

Thus, \( \mathbb{E} [X_j] = \mu \) for \( j = 1, 2, 3, \ldots \), and expression (21) becomes

\[ \| \mu(\Theta) - P_{n+1} \|. \] (22)

If we fix \( \alpha_1, \alpha_2, \ldots, \alpha_n \), which are the coefficients of \( \{X_j\} \) in \( P_{n+1} \), then the minimum of expression (22) is attained with

\[ \alpha_0 = \mathbb{E} \left[ \mu(\Theta) - \sum_{j=1}^{n} \alpha_j X_j \right] = \left( 1 - \sum_{j=1}^{n} \alpha_j \right) \mu, \]

because the mean of a random variable is its orthogonal projection onto the subspace of constants. With the definition

\[ c_0 = 1 - \sum_{j=1}^{n} \alpha_j, \] (23)

equation (16) becomes

\[ P_{n+1} = c_0 \mu + \sum_{j=1}^{n} \alpha_j X_j \]

and, hence,

\[ P_{n+1} - \mu(\Theta) = c_0 [\mu - \mu(\Theta)] + \sum_{j=1}^{n} \alpha_j [X_j - \mu(\Theta)]. \] (24)
It follows from equations (24) and (23) that $\hat{P}_{n+1}$ is the credibility premium minimizing expression (22) if and only if $\hat{P}_{n+1} - \mu(\Theta)$ is the minimum-norm vector in the affine space spanned by $\mu - \mu(\Theta)$ and $X_j - \mu(\Theta)$, $j = 1, 2, \ldots, n$.

We have pointed out earlier that the $\{X_j - \mu(\Theta)\}$ are mutually orthogonal. Also, $X_j - \mu(\Theta) = X_j - E[X_j|\Theta]$ is orthogonal to $\mu - \mu(\Theta)$, because $\mu - \mu(\Theta)$ is a function of $\Theta$. Therefore, we can apply the system of equations (7) to obtain the optimal coefficients:

\begin{align*}
\hat{c}_0 &= \frac{1}{\|\mu - \mu(\Theta)\|^2 + \sum_{j=1}^{n} \frac{1}{\|X_j - \mu(\Theta)\|^2}}, \\
\hat{c}_k &= \frac{1}{\|X_k - \mu(\Theta)\|^2 + \sum_{j=1}^{n} \frac{1}{\|X_j - \mu(\Theta)\|^2}}
\end{align*}

for $k = 1, 2, \ldots, n$.

To express the premium in the form $\hat{P}_{n+1} = (1 - Z)\mu + ZX$, we set

\begin{equation}
Z = 1 - \hat{c}_0 = \frac{\sum_{j=1}^{n} \frac{1}{\|X_j - \mu(\Theta)\|^2}}{\|\mu - \mu(\Theta)\|^2 + \sum_{j=1}^{n} \frac{1}{\|X_j - \mu(\Theta)\|^2}},
\end{equation}

and

\begin{equation}
\hat{X} = \frac{\sum_{j=1}^{n} \frac{X_j}{\|X_j - \mu(\Theta)\|^2}}{\sum_{k=1}^{n} \frac{1}{\|X_k - \mu(\Theta)\|^2}}.
\end{equation}

Thus, $\hat{X}$ is a weighted average of the $X_j$s with the weight attached to $X_j$ being inversely proportional to $\|X_j - \mu(\Theta)\|^2$. Also, note that

$$\|\mu - \mu(\Theta)\|^2 = \|\mu(\Theta) - \mu\|^2 = E[(\mu(\Theta) - \mu)^2] = \text{Var}[\mu(\Theta)],$$
and, by equation (12),

\[ \|X_j - \mu(\Theta)\|^2 = \mathbb{E} \left[ \text{Var} \left[ X_j | \Theta \right] \right]. \]

An illustration of this geometric approach to credibility theory is shown in Figure 1. The affine space spanned by \( \mu - \mu(\Theta) \) and \( X_j - \mu(\Theta) \), \( j = 1, 2, \ldots, n \), is the linear space spanned by \( \mu \) and \( X_j \), \( j = 1, 2, \ldots, n \), translated by \( -\mu(\Theta) \). The vector \( \hat{p}_{n+1} - \mu(\Theta) \), being the minimum-norm vector in the affine space, is orthogonal to all vectors in the linear space spanned by \( \mu \) and \( X_j \), \( j = 1, 2, \ldots, n \); see also condition (10).

4 Applications

The purpose of this section is to derive some of the results in Klugman, Panjer, and Willmot (1998, Chapter 5) using the results above.

(i) In the Bühlmann model as explained in Section 5.4.3 of Klugman, Panjer, and Willmot (1998),

\[ \|\mu - \mu(\Theta)\|^2 = \text{Var} [\mu(\Theta)] = a \]
and
\[ \|X_j - \mu(\Theta)\|^2 = \mathbb{E} \left[ \text{Var} \left[ X_j | \Theta \right] \right] = \mathbb{E} [\nu(\Theta)] = \nu. \]

Hence, equation (27) becomes
\[ Z = \frac{\sum_{j=1}^{n} \frac{1}{v}}{1 + \sum_{j=1}^{n} \frac{1}{v}} = \frac{n}{\frac{1}{a} + n} = \frac{n}{\frac{1}{a + n}}, \]
and equation (28) is
\[ \bar{X} = \frac{n \sum_{j=1}^{n} \frac{X_j}{v}}{\sum_{k=1}^{n} \frac{1}{v}} = \frac{n \sum_{j=1}^{n} X_j}{n}. \]

As a check, we evaluate equation (26),
\[ \hat{\alpha}_k = \frac{1}{\frac{1}{a + n} \frac{1}{v}} = \frac{Z}{\frac{1}{a + n}} = \frac{Z}{n}, \quad k = 1, 2, \ldots, n. \]

(ii) In the Bühlmann-Straub model as explained in Section 5.4.4 of Klugman, Panjer, and Willmot (1998),
\[ \|\mu - \mu(\Theta)\|^2 = \text{Var} [\mu(\Theta)] = a \]
and
\[ \|X_j - \mu(\Theta)\|^2 = \mathbb{E} \left[ \text{Var} \left[ X_j | \Theta \right] \right] = \mathbb{E} \left[ \nu(\Theta)/m_j \right] = \nu/m_j. \]

Hence, with \( m = \sum_{j=1}^{n} m_j \), we have from equation (27)
\[ Z = \frac{\sum_{j=1}^{n} \frac{m_j}{v}}{1 + \sum_{j=1}^{n} \frac{m_j}{v}} = \frac{n \sum_{j=1}^{n} m_j}{\frac{1}{a} + \sum_{j=1}^{n} m_j} = \frac{m}{\frac{1}{a} + m}, \]
and from equation (28)
\[ \bar{X} = \frac{\sum_{j=1}^{n} \frac{m_jX_j}{v}}{\sum_{j=1}^{n} \frac{m_j}{v}} = \frac{\sum_{j=1}^{n} m_jX_j}{m}. \]

As a check, we evaluate equation (26),
\[ \hat{\alpha}_k = \frac{\frac{m_k}{v}}{\frac{1}{a} + \sum_{j=1}^{n} \frac{m_j}{v}} = \frac{Z}{m}, \quad k = 1, \ldots, n. \]

(iii) In Example 5.40 of Klugman, Panjer, and Willmot (1998),
\[ \|\mu - \mu(\Theta)\|^2 = \text{Var}[\mu(\Theta)] = a \]
and
\[ \|X_j - \mu(\Theta)\|^2 = \mathbb{E}\left[\text{Var}(X_j|\Theta)\right] = \mathbb{E}\left[w(\Theta) + v(\Theta)/m_j\right] = w + v/m_j. \]

Hence, with
\[ m^* = \sum_{j=1}^{n} \frac{1}{\|X_j - \mu(\Theta)\|^2} = \sum_{j=1}^{n} \frac{m_j}{v + wm_j}, \]
we have
\[ Z = \frac{m^*}{1/a + m^*} = \frac{am^*}{1 + am^*}, \]
and
\[ \bar{X} = \frac{\sum_{j=1}^{n} \frac{m_jX_j}{v + wm_j}}{\sum_{k=1}^{n} \frac{m_k}{v + wm_k}} = \frac{\sum_{j=1}^{n} m_jX_j}{m^*}. \]
As a check, we evaluate equation (26),

\[ \hat{\alpha}_k = \frac{m_k}{1 + m^*} = Z \frac{1}{v + w} \cdot \frac{m_k}{m^*}, \quad k = 1, \ldots, n. \]

(iv) In Example 5.41 of Klugman, Panjer, and Willmot (1998),

\[ m = \sum_{j=1}^{n} m_j, \quad \|\mu - \mu(\Theta)\|^2 = \text{Var}(\mu(\Theta)) = a + b/m \]

and

\[ \left\| X_j - \mu(\Theta) \right\|^2 = \mathbb{E} \left[ \text{Var}(X_j|\Theta) \right] = w + v/m_j. \]

Hence,

\[ Z = \frac{m^*}{\frac{1}{a + b/m} + m^*} = \frac{(a + b/m)m^*}{1 + (a + b/m)m^*}, \]

and

\[ \hat{X} = \frac{\sum_{j=1}^{n} m_j X_j}{\frac{1}{v + w} m_j}. \]

As a check, we evaluate equation (26),

\[ \hat{\alpha}_k = \frac{m_k}{1 + m^*} = Z \frac{1}{v' + w'} \cdot \frac{m_k}{m^*}, \quad k = 1, \ldots, n. \]

(v) To solve Exercise 5.51 in Klugman, Panjer, and Willmot (1998), consider \( X_j/\beta_j \) in the exercise as \( X_j \) in Section 3 above.

(vi) To solve Exercise 5.56 in Klugman, Panjer, and Willmot (1998), consider \( X_j/\tau^j \) in the exercise as \( X_j \) in Section 3 above.

5 Similar Triangles

Similar triangles are now used to derive several equivalent forms for the credibility factor, \( Z \), and, hence, several equivalent forms for the credibility premium,
It follows from equation (29) that

\[ Z = \frac{\|\hat{P}_{n+1} - \mu\|}{\|\bar{X} - \mu\|} \]  

(30)

and

\[ 1 - Z = \frac{\|\bar{X} - \hat{P}_{n+1}\|}{\|\bar{X} - \mu\|}. \]

(Thus, \( Z \) is the ratio of the standard deviation of \( \hat{P}_{n+1} \) to that of \( \bar{X} \).)

Now, equation (29) is equivalent to

\[ \hat{P}_{n+1} = Z\bar{X} + (1 - Z)\mu. \]  

(29)

As \( \bar{X} \) is an average of \( \{X_j\} \), we have \( E[\bar{X}|\Theta] = \mu(\Theta) \), from which it follows that \( [\bar{X} - \mu(\Theta)] \) and \( [\mu - \mu(\Theta)] \) are orthogonal to each other. Figure 2 illustrates the geometric relationships among the random variables; note that Figure 2 is a slice in Figure 1.

There are three similar right-angled triangles in Figure 2. We shall show that each triangle gives a different form for \( Z \) (and for \( 1 - Z \)). In each triangle, there are two acute angles complementary to each other. We shall also show that the square of the cosine of one of the acute angles gives the value of the credibility factor \( Z \), while the square of the cosine of the other is \( 1 - Z \).

The three triangles yield three equivalent sets of ratios,
\[
\|\hat{X} - \mu\| : \|\hat{X} - \mu(\Theta)\| : \|\mu(\Theta) - \mu\|
\]
\[
= \|\hat{X} - \mu(\Theta)\| : \|\hat{X} - \hat{\mu}_{n+1}\| : \|\mu(\Theta) - \hat{\mu}_{n+1}\|
\]
\[
= \|\mu(\Theta) - \mu\| : \|\mu(\Theta) - \hat{\mu}_{n+1}\| : \|\hat{\mu}_{n+1} - \mu\|. \quad (31)
\]

In particular, we have the equation
\[
\frac{\|\hat{X} - \mu\|}{\|\mu(\Theta) - \mu\|} = \frac{\|\mu(\Theta) - \mu\|}{\|\hat{\mu}_{n+1} - \mu\|}, \quad (32)
\]

which applied to equation (30) yields
\[
Z = \frac{\|\mu(\Theta) - \mu\|^2}{\|\hat{X} - \mu\|^2} \quad (33)
\]

and
\[
Z = \frac{\|\hat{\mu}_{n+1} - \mu\|^2}{\|\mu(\Theta) - \mu\|^2}. \quad (34)
\]

From (30) and (31), we also obtain
\[
Z = \frac{\|\hat{\mu}_{n+1} - \mu(\Theta)\|^2}{\|\hat{X} - \mu(\Theta)\|^2}. \quad (35)
\]

Corresponding to equations (33), (34), and (35), we have
\[
1 - Z = \frac{\|\hat{X} - \mu(\Theta)\|^2}{\|\hat{X} - \mu\|^2}, \quad (36)
\]
\[
1 - Z = \frac{\|\mu(\Theta) - \hat{\mu}_{n+1}\|^2}{\|\mu(\Theta) - \mu\|^2} \quad (37)
\]

and
\[
1 - Z = \frac{\|\hat{X} - \hat{\mu}_{n+1}\|^2}{\|\hat{X} - \mu(\Theta)\|^2}, \quad (38)
\]

respectively.

The usual credibility premium equation is obtained by applying equations (33) and (36),
\[ \hat{P}_{n+1} = \frac{\|\mu(\Theta) - \mu\|^2}{\|\bar{X} - \mu\|^2} \bar{X} + \frac{\|\bar{X} - \mu(\Theta)\|^2}{\|\bar{X} - \mu\|^2} \mu. \] 

(39)

\[ = \frac{\text{Var}[\mu(\Theta)]}{\text{Var}[\bar{X}]} \bar{X} + \frac{\mathbb{E}[\text{Var}[\bar{X}|\Theta]]}{\text{Var}[\bar{X}]} \mu. \] 

(40)

The credibility premium can thus be viewed as a weighted average of \( \bar{X} \) and \( \mu \), with weights distributed according to the Pythagorean equation

\[ \|\bar{X} - \mu\|^2 = \|\bar{X} - \mu(\Theta)\|^2 + \|\mu(\Theta) - \mu\|^2, \]

or its equivalent variance-decomposition equation

\[ \text{Var}[\bar{X}] = \mathbb{E}[\text{Var}[\bar{X}|\Theta]] + \text{Var}[\mathbb{E}(\bar{X}|\Theta)]. \]

Equation (39) follows from equations (6) and (7), with \( m = 2, v_1 = [\bar{X} - \mu(\Theta)] \), and \( v_2 = [\mu - \mu(\Theta)] \).

The cosine of the angle between \( [\mu(\Theta) - \mu] \) and \( [\bar{X} - \mu] \) is the correlation coefficient between \( \mu(\Theta) \) and \( \bar{X} \), which we call \( \rho_{\bar{X},\mu(\Theta)} \). Hence, it follows from equation (33) that \( Z \) is the square of the correlation coefficient, i.e.,

\[ Z = \rho_{\mu(\Theta),\bar{X}}^2, \]

and the credibility premium is

\[ \hat{P}_{n+1} = \rho_{\mu(\Theta),\bar{X}}^2 \bar{X} + (1 - \rho_{\mu(\Theta),\bar{X}}^2) \mu. \]

Also, it follows from equation (34) that the credibility factor \( Z \) is the square of the correlation coefficient between \( \mu(\Theta) \) and \( \hat{P}_{n+1} \),

\[ Z = \rho_{\mu(\Theta),\hat{P}_{n+1}}^2. \]

We remark that

\[ \text{Cov}[\bar{X},\mu(\Theta)] = \text{Cov}[\mathbb{E}[\bar{X}|\Theta],\mu(\Theta)] = \|\mu(\Theta) - \mu\|^2, \]

which may be viewed as a consequence of equation (2). Also,

\[ \text{Cov}[\hat{P}_{n+1},\mu(\Theta)] = \|\hat{P}_{n+1} - \mu\|^2. \]
We conclude this paper with some equations that readily follow from the discussion above. These equations provide further insights for understanding credibility theory.

From the ratios (31) we can obtain

\[ \| \mu(\Theta) - \mu \| \| \bar{X} - \mu(\Theta) \| = \| \mu(\Theta) - \hat{P}_{n+1} \| \| \bar{X} - \mu \|. \]  

(41)

If we divide both sides of equation (41) by 2, then the two sides of the equation represent two ways for finding the area of the largest triangle in Figure 2. Another consequence of the ratios (31) is

\[ \frac{1}{\| \mu(\Theta) - \hat{P}_{n+1} \|^2} = \frac{1}{\| \mu(\Theta) - \mu \|^2} + \frac{1}{\| \bar{X} - \mu(\Theta) \|^2}, \]

which also follows from equation (9).

From equation (32) we see that \( \text{Var} [\mu(\Theta)] \) is the geometric mean of \( \text{Var} [\bar{X}] \) and \( \text{Var} [\hat{P}_{n+1}] \). Let us rewrite equations (33) and (34) as

\[ \text{Var} [\mu(\Theta)] = Z \text{Var} [\bar{X}] \]  

(42)

and

\[ \text{Var} [\hat{P}_{n+1}] = Z \text{Var} [\mu(\Theta)], \]  

(43)

respectively. Applying equation (42) to (43) yields

\[ \text{Var} [\hat{P}_{n+1}] = Z^2 \text{Var} [\bar{X}], \]

which is also a consequence of equation (30).

Recall that \( \hat{P}_{n+1} \) is the solution in minimizing (17). Thus, it follows from equation (3) that

\[ \| X_{n+1} - \mu \|^2 = \| X_{n+1} - \hat{P}_{n+1} \|^2 + \| \hat{P}_{n+1} - \mu \|^2, \]

or

\[ \text{Var}(X_{n+1}) = \| X_{n+1} - \hat{P}_{n+1} \|^2 + \text{Var} (\hat{P}_{n+1}). \]

Also, if we write expression (17) as
\[ \|X_{n+1} - [Z\bar{X} + (1 - Z)\mu]\| = \|(X_{n+1} - \mu) - Z(\bar{X} - \mu)\|, \]

we see from the left side of equation (4) that the coefficient of \((\bar{X} - \mu)\) is

\[ Z = \frac{\langle X_{n+1} - \mu, \bar{X} - \mu \rangle}{\langle \bar{X} - \mu, \bar{X} - \mu \rangle} = \frac{\text{Cov}[X_{n+1}, \bar{X}]}{\text{Var}(\bar{X})}. \] (44)

Equation (44) can be found in Fuhrer (1989, equation 1). Fuhrer (1989, page 84) derived the equation without assuming the existence of the risk parameter \(\Theta\); he also made some interesting remarks concerning the equation. A parameter-free approach to credibility theory can be found in Jones and Gerber (1975) and in Section 6.3 of Gerber (1979). Jones and Gerber (1975) also provided an appendix entitled "credibility theory ... in the light of functional analysis."

For further discussions on credibility and geometry, we refer the reader to De Vylder (1976a, 1976b, 1996), Gisler (1990), Hiss (1991), Jones and Gerber (1975), Norberg (1992), and Taylor (1977). We also recommend the book by Small and McLeish (1994).

References


