

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Journal of Actuarial Practice 1993-2006

Finance Department


2001

Journal of Actuarial Practice, Volume 9 (2001)

Colin Ramsay, Editor

University of Nebraska - Lincoln, cramsay@unl.edu

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

 Part of the [Accounting Commons](#), [Business Administration, Management, and Operations Commons](#), [Corporate Finance Commons](#), [Finance and Financial Management Commons](#), [Insurance Commons](#), and the [Management Sciences and Quantitative Methods Commons](#)

Ramsay, Colin, Editor, "Journal of Actuarial Practice, Volume 9 (2001)" (2001). *Journal of Actuarial Practice 1993-2006*. 43.
<http://digitalcommons.unl.edu/joap/43>

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

ARTICLES

**Analyzing Management Fees of Pension Funds:
A Case Study of Mexico**

Tapen Sinha 5

**Premium Earning Patterns for Multi-Year
Policies with Aggregate Deductibles**

Thomas Struppeck 45

**Exponential Bonus-Malus Systems Integrating
A Priori Risk Classification**

Lluís Bermúdez, Michel Denuit, and Jan Dhaene 67

**Fitting Loss Distributions
in the Presence of Rating Variables**

Farrokh Guiahi 97

**Linear Empirical Bayes Estimation
of Survival Probabilities with Partial Data**

Mostafa Mashayekhi 131

**Controlling the Solvency Interaction
Among a Group of Insurance Companies**

Alexandros Zimbidis and Steven Haberman 151

**A Sensitivity Analysis of the Premiums for a
Permanent Health Insurance (PHI) Model**

Ben D. Rickayzen 189

**Premium Calculation Using the
Probability of Ruin**

K.C. Yuen, H. Yang, and K.L. Chu 213

Analyzing Management Fees of Pension Funds: A Case Study of Mexico

Tapen Sinha*

Abstract[†]

Though the rates of return for public pension funds have been high over the past two decades, one critical aspect of the financing of this type of fund is often overlooked: high management fees. As a result, the rates of return for workers who have invested in these funds have not necessarily been high. Management fees charged on pension funds in Mexico result in a leakage of funds in the order of 20-30% of the fund. That is, the amount at retirement would have been 20-30% higher had there been no fees.

A model is developed that includes all the diverse fees and discounts. No other model of the Mexican system contains all of these fees and discounts. Therefore, simulations from other studies do not yield reliable results. Our simulation results show that it is rarely optimal (from the point of view of minimizing lifetime management fees) to stay with one company. Also, no company dominates all others with respect to the minimization of its fees. Unfortunately, because of the complexity of the fee structure, it is difficult to say much beyond this. This research shows that the risks that the privatized system carries may be much higher than what appears at first sight.

Key words and phrases: privatization, defined contribution plan, Chile, retirement fund, pay-as-you-go, AFORE, simulation

*Tapen Sinha, Ph.D., is the Seguros Comercial America Chair Professor of actuarial studies at the Instituto Tecnológico Autónomo de Mexico (ITAM). He obtained his B.Sc. and M.Sc. degrees from the Indian Statistical Institute and his Ph.D. in Economics from the University of Minnesota. Dr. Sinha has written over 80 research papers that have appeared in economics, finance, marketing, accounting, insurance, and medical journals. He is the author of *Pension Reform in Latin America and Its Lessons for International Policymakers* (Kluwer Academic Publishers, Boston, 2000). His forthcoming book on the Mexican pension system will be published by the Society of Actuaries. He has wide consulting experience in the U.S., Singapore, Australia, and in Mexico.

Dr. Sinha's address is: Department of Actuarial Studies, Instituto Tecnológico Autónomo de Mexico (ITAM), Rio Hondo No. 1, Col. Tizapán San Ángel, C.P. 01000 MEXICO, D.F. Internet address: tapen@itam.mx

[†]The author thanks the anonymous referees for their valuable input, Felipe Martínez and Connie Barrios-Muñoz for their help with the simulations, and the editor of this journal for his meticulous editing of the paper and for pointing out several mistakes.

1 Introduction

Privatization of pensions has become an important issue around the world. From Chile to China, from Argentina to Zimbabwe, privatization of pensions either has been implemented or is being contemplated (Schwarz and Demirguc-Kunt, 1999).

Nowhere in the world has privatization of state-run pension schemes been undertaken with more zeal than in Latin America. Ten countries in the world have privatized their pension plans (Social Security Administration, 1999)—eight of them are from Latin America (Argentina, Bolivia, Chile, Colombia, El Salvador, Mexico, Peru, and Uruguay). The other two are from Eastern Europe (Hungary and Poland).

In 1997, a new privatized (but government-mandated) system of retirement program was created in Mexico. The system is essentially a defined contribution pension plan in which private companies operate pension funds. Each company operating a pension fund is called an *Administradora de Fondos de Retiro* or an AFORE. Under the system each worker will have his or her own account with an AFORE, and this account accumulates the individual and government contributions and the investment returns generated by these contributions. Thus, the contributions and the performance of the fund solely will determine each worker's pension benefit.

In some sense, the Mexican model can be viewed as an adaptation of the Chilean model. The Chilean model is the most decentralized model of pension. Chile has succeeded in delivering many benefits for which privatized pension plans strive. Most policy makers in Mexico are familiar with the system in Chile and are influenced by it. Economists (such as Diamond, 1994, 1999) have criticized the Chilean system because of its high transaction cost. In some ways, the high growth rate in real wages and high real rates of return have obscured high transactions costs for Chile.

Like the Chilean system, an unfortunate feature of the new Mexican system is its relatively high management fees. Management fees imposed on the pension funds in Mexico are the most complex in Latin America. It is difficult for anyone other than sophisticated investors to disentangle the effects of various charges and determine which fund offers the best rate of return. There are several types of fees and discounts. These high fees result in severe losses to the development of the workers' funds.

The objective of this paper is to analyze the impact of management fees on funds available at retirement for Mexican workers. To this end a model is developed in which all the diverse fees and discounts are ac-

counted for. This model is used to calculate future values of the fund, taking into account all the complexities of the Mexican system. The model is also used to compare funds over various horizons under a variety of scenarios. In all other models developed in the Mexican context, many of these cost elements have been ignored. Due to the complexity of the management fee structure, analytic results are difficult to obtain; simulations are used instead.

This paper is organized as follows: Section 2 discusses some of the reasons for moving from a state-run pay-as-you-go system to a privatized system. Section 3 reviews the basics of the new Mexican system and comments on some of its deficiencies.

2 Why Privatize Social Security?

Why are Latin American countries enthused about privatizing social security? There are four related reasons:

- Policy makers have recognized that their current state-run pay-as-you-go systems will be bankrupt within the next decade or so.
- The pioneering privatization plan in Chile has been extolled for its success. The Chilean example has given privatization a new sense of urgency in neighboring countries.
- Privatization systems seem to increase national saving.
- Privatization helps develop long-term capital markets.

2.1 What's Wrong With the Pay-as-You-Go System?

Developed countries (such as the United States) are beginning to experience problems with pay-as-you-go retirement systems (such as the U.S. social security system) largely due to a mismatch of benefits paid to retirees compared with the revenue generated from the working population.

Similar problems are in the offing for other countries. Therefore, aging itself provides a strong incentive for fixing the systems in some ways.

These problems can arise in a number of different ways:

- The government increases the benefits of the retired population by indexing benefits to inflation without indexing revenue in the same way.

- The government relaxes eligibility (for example, by relaxing the age of retirement, by making the definition of disability or poor health broader etc.).
- Government, directly or indirectly, reduces its revenue base. For example, a rise in the marginal tax rates may force some people to leave the formal employment sector (where they finance such a scheme through payroll taxes) and enter the informal (cash-only) sector where they avoid paying payroll taxes thus reducing the government's revenue base.
- The population is aging, i.e., the percentage of the aged in the population is increasing. Aging is caused mainly by falling birth rates and falling mortality rates at the older ages.

Table 1 illustrates how the projected proportion of older persons will rise (in some cases, dramatically) in Latin American countries. For comparison, the United States is included in Table 1. There are two striking features of Table 1:

1. From looking at the column for 2050, all the countries appear to be converging to a similar population structure. The proportion of people over 60 is similar across countries. If, instead of this proportion, we look at the entire distribution (say separated by five years), it will be similar too.
2. Not all countries have the same degree of population aging in 1990.

As Argentina and Uruguay have population structures that are similar to the United States' population structure today, there is a certain urgency of reform for their state-run pension scheme that exceeds those of other Latin American countries. On the other hand, even though Peru has a much younger population structure today, its population is expected to experience the benefits of better health care and medicines and will age rapidly over the next 50 years. A similar experience is expected to occur in the other countries in Latin America.

From the point of view of demographics (i.e., population structure), the potential problem may seem to be far in the future. But many Latin American countries will face the problem sooner rather than later. There are many inefficiencies in their public pension systems, including a large informal (cash-only) sector, that make the problem more acute than ever before (Vittas, 1994).

Bolivia is a classic example of how things can go wrong, even when the population structure is young. Bolivia has had a defined-benefit

Table 1
Projections of Percentage of Population
Over Age 60 from 1990–2050

Year	1990	2030	2050
Argentina	13.1	19.3	25.9
Bolivia	5.4	10.0	17.6
Brazil	6.7	16.9	24.2
Chile	8.7	20.8	26.4
Colombia	6.0	18.0	25.5
Ecuador	5.5	13.7	22.4
Mexico	5.7	15.7	24.6
Paraguay	5.2	10.4	16.1
Peru	5.8	13.7	21.5
Uruguay	16.4	22.5	27.8
Venezuela	5.6	15.5	23.6
U.S.	16.6	28.2	29.8

Source: World Bank (1994).

pay-as-you-go scheme for many years. In 1997 the number of people contributing to the system was 300,000. The number of people drawing a pension from the system was 120,000. Thus, the pensioner to contributor dependency ratio of the system was 40 (120,000/300,000) percent. If we look at the ratio of number persons age 60+ to the number age less than 60 in the population as a whole, however, it is 5.8 ($100 \times 0.054 / (1 - 0.054)$) percent (Table 1). The percentage of GDP covered by the system was less than 12 percent (von Gersdorff, 1997). Most pensioners were either government employees (65 percent of the total) or schoolteachers (30 percent). The Bolivian economy, however, is dominated by the informal sector.

2.2 Why is Everyone Looking at Chile?

The Chilean system has produced spectacular results in terms of rates of return on funds (Table 2). The system also has created deeper financial markets; markets for long-term bonds have developed as a direct consequence of the system. The saving rate in Chile also has risen spectacularly over the same period, from 8.2 percent of GDP in 1982 to 23.3 percent in 1996. Real GDP has increased at an average

annual rate of 7.7 percent from 1980 to 1997. [For an illuminating discussion on the Chilean system, see Edwards (1996).] GDP growth has slowed to 3.1 percent in 1998 and -1.4 percent in 1999. Many commentators have jumped to the conclusion that the rise in saving and GDP are (partly) consequences of privatization of pension (Piñera, 2000). This conclusion, however, is not supported by statistical evidence (Holzmann, 1996).

Table 2
Percentage Rates of Return
For Pension Funds in Chile

Year	Weighted	
	Average	Range
1982	28.8	23.2 to 30.2
1983	21.2	18.5 to 24.7
1984	3.6	2.2 to 5.1
1985	13.4	13.0 to 14.3
1986	12.3	10.6 to 15.5
1987	5.4	4.8 to 8.5
1988	6.5	5.9 to 8.7
1989	6.9	4.0 to 9.5
1990	15.6	13.3 to 19.4
1991	29.7	25.8 to 34.3
1992	3.0	0.9 to 4.2
1993	16.2	14.6 to 16.9
1994	18.2	15.7 to 21.1
1995	-2.5	-4.6 to -1.8
1996	3.5	2.9 to 4.1
1997	4.7	-0.2 to 5.5
1998	-1.1	-2.7 to -0.4
1999	12.31	11.99 to 14.16

Source: Banco Central de Chile, *Boletín Mensual* (various issues). *Notes:* Rates of return (in %) are weighted by the asset value in each pension fund. The data for 1999 are through August 1999.

There are several notable features of Table 2. First, the average rates of return for funds in Chile have been high. This has impressed many

foreign observers; however, there is a large year-to-year variation. At the same time, the rates of return for different funds in any particular year have not varied a great deal in that year (especially early years). The rate of return for funds is misleading, however, as it does not necessarily mean the same thing for the workers who have contributed to these funds. This difference is discussed further below.

2.3 Saving and Capital Market Developments

In theory, under certain conditions savings may rise as a result of privatization. Such results are sensitive to model specification. A change in model specification can lead to a collapse of the result (Sinha, 2000, Chapter 2). As Chile has the longest experience of privatized pensions, it is natural that researchers have turned to Chile to investigate the question of whether savings rise under privatization. The Chilean evidence, when carefully analyzed, shows that national saving rate does not increase when social security is privatized (Holzmann, 1996, Agosin et al., 1997).

Do capital market developments follow from pension privatization? It is clear that privatization needs to be preceded by some capital market development. For example, there has to be a well-functioning government bond market (Vittas, 1996).

3 Mexican System

3.1 The Pre-Reform System

Since 1943 Mexico had a system run by the IMSS (Instituto Mexicano del Seguro Social). There are four pillars of this system: (i) disability, old age, severance, and life insurance; (ii) maternity and health insurance; (iii) workplace insurance; and (iv) childcare centers.

The disability, old age, severance, and life insurance component (also known as Seguro de Invalidez, Vejez, Cesancia en Edad Avanzada y Muerte abbreviated to IVCN) is the largest program for social security in Mexico. IVCN is a pay-as-you-go scheme that has protected workers in the formal sector since 1943. In addition, there are separate disability, old age, severance, and life insurance programs for government employees, for the armed forces, and others.

Total contribution to IVCN was 8.5 percent of base salary in 1996. There is a tripartite split between the employers, employees, and the

government. Employers paid 5.95 percent, employees paid 2.125 percent, and the government paid 0.425 percent of the base salary. In addition, there was an additional payment of 2 percent of base salary in the SAR (Sistema para el retiro, the *retirement account*).

There were many problems associated with the old IVCM system. Only a small percentage of workers was covered. For example, in 1999 fewer than 30 percent of the workers in the labor force were covered. In addition, it is estimated that without any reform, revenue for the IMSS in 1999 would have fallen short of the cost in 1999, a classic problem of pay-as-you-go schemes.

3.2 Pension Reform in Mexico

On July 1, 1997, a new privatized (but government-mandated) retirement system came into existence in Mexico, replacing IVCM. This system consists of private companies operating pension funds. Each company operating a pension fund is called an *Administradora de Fondos de Retiro* or an AFORE. The investment fund, run by the company independent of the parent AFORE company, is called a *Sociedad de Inversión en Fondos de Retiro* (a SIEFORE).

Each worker in the system is assigned a retirement fund account with an AFORE. Funds in the account accumulate through periodic employer, employee, and government contributions and from the yield generated by investment in the AFORE. Thus, the contributions and the investment performance of the fund alone determine each worker's pension benefits at retirement. This individual defined contribution pension scheme contrasts sharply with the old pay-as-you-go scheme ran by IMSS.

Among the four pillars mentioned earlier, only IVCM was privatized through AFORES. The other three pillars are still being operated by IMSS. We will not consider the other three pillars of the IMSS. [See Banco de Mexico (1996) for further discussion on reform in the other three pillars.]

There are two elements of contribution to an account: contribution of 6.5 percent of wages by the employee/employer and a government contribution of 5.5 percent of national minimum salary (regardless of the worker's actual salary). For a worker who earns exactly the minimum salary, the contribution to an AFORE will be 11.5 percent (6.5 + 5.5) of his or her salary. For a worker earning ten times the minimum salary, the contribution will be 7.05 percent (6.5 + 5.5/10) of his or her actual salary. For the average worker, the government contribution amounts to 2.2 percent of salary. For high-income workers, the

government contribution is a relatively insignificant percentage of their salary.

The new system is compulsory for persons entering the workforce on or after July 1, 1997. Individuals who have contributed to the old system have a choice: they can opt for the benefits under the old scheme or they can receive benefits from the new scheme, whichever is larger. The majority that have contributed to the old system for at least 20 years will fare better under the old scheme. For others, it depends critically on the rates of return that the new scheme will earn. Thus, there will be additional costs incurred during the transition. The cost will rise up to 4 percent of GDP during the early part of the 21st century (Sales-Sarrapy et al., 1996).

The new system has spawned many AFOREs. Seventeen AFOREs have been given licenses to operate (although four since have merged). Mexican companies (mainly banks) own some of them (wholly). Others have large (although not majority) foreign shareholders. Table 3 lists the AFOREs in operation at the end of 2000.

The Mexican government has created a separate division to oversee the activities of the AFOREs: the Comisión Nacional del Sistema de Ahorro para el Retiro (CONSAR). CONSAR has the critical role of overseeing all the activities of AFOREs. For example, CONSAR has established general rules of operation of the AFOREs.

The objectives of these institutions include:

- Open, administer, and manage the individual retirement accounts in agreement with provisions in social security laws. Regarding housing-promotion sub-accounts, the AFOREs will register each worker's contributions, and the interest paid thereon, based on information provided by social security institutions.
- Receive, from social security institutions, the contributions made, in accordance with the law, by the government, employers, and workers, as well as voluntary contributions by workers and employers.
- Itemize the amounts received periodically from social security institutions and deposit them into each worker's individual retirement account, as with the returns obtained on the investment of these funds.
- Provide administrative services to mutual investment funds. (Banco de México, 1996).

Table 3
AFORES Authorized by CONSAR and Their Compositions

AFORE	Main Shareholders and Percentage Holding
Atlántico Promex	Banca Promex 50, Banco del Atlántico 50
Banamex	Grupo Financiero Banamex-Accival 100
Bancomer	Grupo Financiero Bancomer 51, Aetna Internacional, Inc. 49
Bancrecer-Dresdner	Grupo Financiero Bancrecer 51, Dresdner Pension Fund Holdings 44, Allianz México, S. A. 5
Bital	Grupo Financiero BITAL 51, ING America Insurance Holding, Inc. 49
Capitaliza	General Electric Capital Assurance Co. 100
Confia-Principal	Abaco Grupo Financiero 51, Principal International 49
Garante	Grupo Financiero Serfin 51, Grupo Financiero Citibank 40, Hábitat Desarrollo Internacional 9
Génesis	Seguros Génesis, S. A. 100
Inbursa	Grupo Financiero INBURSA 100
Previnter	Boston AIG Company 90, The Bank of Nova Scotia 10
Profuturo GNP	Grupo Nacional Provincial 51, Banco Bilbao Vizcaya-México, S. A. 25, Provida Internacional, S. A. 24
Santander Mexicano	Grupo Financiero Invermexico 75, Santander Investment, S. A. 25
Siglo XXI	Instituto Mexicano del Seguro Social 50, IXE Grupo Financiero 50
Sólida Banorte	Grupo Financiero Banorte
Tepeyac	Seguros Tepeyac
Zurich	Zurich Vida, Compañía de Seguros 77, Gabriel Monterrubio Guasque 10

Note: No mention is made of shareholders with equity participation under 5 percent of the total capital of the respective AFORE

Table 4
Administrative Costs as a Percentage of Expenditure

Latin America		OECD	
Argentina	2.30	Australia	1.22
Bolivia	21.39	Canada	2.80
Chile	8.00	France	4.18
Colombia	81.80	Germany	2.86
El Salvador	33.40	Italy	2.20
Mexico	23.55	Japan	1.79
Peru	130.98	Spain	2.81
Uruguay	6.51	Switzerland	3.04
		United Kingdom	3.10
		United States	3.28

Source: Mitchell (1996)

The cost of administering the new system is high by OECD standards. When compared with other Latin American countries, however, administrative costs are not out of line (Table 4).

Because charges apply to different parts of the AFORE, it is not easy to compare charges across AFOREs. If we examine the system as a whole, however, the charges appear too high at this early stage of the system's development. In Chile, for example, in 1984 charges amounted to 9 percent of wages or 90 percent of contributions to the retirement system (Edwards, 1996, p. 17). The costs dropped to about 15 percent of contributions in 1990 (World Bank, 1994, p. 224).

3.3 Organization and Investment Activities of AFOREs

Some AFOREs are fully owned by Mexican companies, while other AFOREs are partly owned by foreign companies. For example, AFORE Bancomer is 51 percent owned by the second largest banking group in Mexico and the other 49 percent is owned by Aetna, one of the largest insurance companies in the United States. Garante has the most interesting ownership structure. It has the majority shareholding by a Mexican group; it is partly owned by Citibank; and it is partly owned by a pension fund from Chile, AFP Habitat.

On one hand, the Mexican government was keen to have foreign companies participate in this sector, because foreign participation usu-

ally signals a faith in the system. On the other, the government was also keen on keeping the majority shareholding within the country for political reasons. By the end of 1999, three of the AFOREs have already merged with others. Atlantico has been sold to Confia; Genesis has been sold to Santander; and Previnter has been sold to Profuturo.

Although CONSAR is clear on ownership rules, it has been ambiguous on the issue of prevention of monopoly rule. It states:

CONSAR will establish procedures to prevent absolute or relative monopolistic practices resulting from the behavior of individual market participants or due to market concentration. In doing so, the CONSAR will abide by the Economic Competition Federal Act. Accordingly, *no single AFORE may have more than 20 percent of the retirement saving system's market*. Subject to prior authorization from its Consultative and Surveillance Committee, the CONSAR may authorize greater market concentration ratios, as long as this does not harm workers' interests.

The rule initially did not define the phrase "no more than 20 percent of the market." Later, CONSAR ruled that it meant 20 percent of the total number of individual accounts (rather than 20 percent of market share in terms of value). CONSAR also left the question of some AFOREs operating with more than 20 percent of all individual accounts open by adding the phrase "as long as this does not harm workers' interests."

At present, AFOREs do not have much freedom in choosing their investment portfolios. Basically, all of their investments have to be in the form of Mexican government bonds (called CETES) and price-indexed linked bonds (such as UDIBONOS).

CETES (Certificados de la Tesorería de la Federación) are peso-denominated money market instruments issued by the Mexican Treasury in 28-day, 91-day, 182-day, 364-day, and 728-day maturities. CETES are considered to be the short-term interest rate benchmark in Mexico and, with rare exceptions, are auctioned on a weekly basis. CETES are similar to U.S. Treasury bills. The market for CETES is the most important capital market instrument available in Mexico. It is also one of the few Mexican capital market instruments with an active futures market: CETES futures are traded in the Chicago Mercantile Exchange.

As a consequence, CONSAR has chosen CETES to be the first instrument for the AFOREs. Because there are CETES of differing maturities, it is possible to obtain different rates of return on CETES, as the term structure of interest rates does not stay constant over time.

Table 5
Annualized Rates of Return
(July 1997-June 1999)

Name	Nominal	Real
Banamex	28.83%	8.38%
Bancomer	29.12%	8.59%
Bancrecer	25.12%	5.64%
Bital	29.90%	9.17%
Garante	29.21%	8.66%
Génesis	28.29%	7.98%
Inbursa	25.26%	5.75%
Principal	27.54%	7.43%
Profuturo	29.92%	9.19%
Santander	26.48%	6.64%
Banorte	28.19%	7.91%
Tepeyac	26.48%	6.64%
XXI	27.27%	7.23%
Zurich	26.79%	6.87%
Average	28.33%	8.01%

Source: CONSAR

About 35 percent of total investment by AFORES has been in CETES. Another 48 percent has been in five-year inflation-indexed government bonds called Bonde91, while another 10 percent has been in convertible bonds called Udibonos (July 2000).

Restrictions on the use of financial instruments by the AFORES have reduced the variability in the before-charges rates of return of the funds (Table 5). With the restrictions imposed, one important question arises: why should different AFORES charge such high fees? After all, their roles have been reduced to (almost) nothing but bookkeeping (Espinosa and Sinha, 2000).

Though there have been high rates of return of the funds, this does not automatically imply a high rate of return for workers who have money in those funds. The basic problem is the high management fees charged by private pension funds. Shah (1997) has calculated these rates of return after charges for Chile (Table 6). Table 6 shows that even though the real rates of return of funds have been large and positive for the funds, they have not been so for the affiliates.

Table 6
Comparing Real Rates of Return of Funds
And Cumulative Real Rates of Return of Affiliates in Chile

Year	Rates For Funds	Cumulative Rates For Affiliates
1982	28.8%	-3.2%
1983	21.3%	-1.3%
1984	3.5%	-5.9%
1985	13.4%	-2.3%
1986	12.3%	0.3%
1987	5.4%	0.5%
1988	6.4%	1.4%
1989	6.9%	2.1%
1990	15.5%	4.2%
1991	29.7%	7.9%
1992	3.1%	6.9%
1993	16.2%	8.0%
1994	18.4%	9.1%
1995	-2.5%	7.4%

Source: Shah (1997). *Notes:* The first column gives the rate of return of the fund in a given year. The second column gives the cumulative rate of return. Thus, for example, the figure for 1995 for the affiliates is the real rate of return the affiliate would have between 1982 and 1995. As a result, it is possible for the second column to have a bigger number than the first column.

The basic features of individual accounts are similar in Mexico. Therefore, it should not be surprising that the Mexican system will not produce positive real rates of return in the next decade.

4 Calculating Future Values of AFORE

Individual retirement benefits are essentially calculated using an accumulated value formula. This formula must account for wages, contributions, fees, and discounts. In particular, the following are peculiarities of the Mexican system:

- The government contribution to the individual account is made every two months, and indexing is not applicable monthly.

- Commissions come in three basic varieties:
 - Commissions on the flow of funds,
 - Commissions on the account balance, and
 - Commissions on the real rate of return.
 - Some companies charge commissions combining all these options.
- In addition, these commissions may vary with the number of years one stays in the fund.

For these reasons, the following discussion will be devoted to a step by step development of the formula for calculating retirement benefits.

4.1 The Basic Formula

There are two components of the new system: the contribution by the worker and the contribution by the government. The contribution by the worker is 6.5 percent of his or her base wage. The contribution by the government is 5.5 percent of the minimum salary indexed to the rate of inflation. There are two additional complications: the interest rate is calculated for every account every two months, and indexation of the government contribution to inflation occurs every three months.

For $k = 1, 2, \dots$, let S_k denote the accumulated sum in the k^{th} month; BW_k denote the worker's base wage in the k^{th} month; G_k denote the government's contribution in the k^{th} month; $i_k^{(12)}$ denote the nominal annual rate of interest compounded monthly that is in effect in the k^{th} month (see, for example, Kellison (1991) for more on nominal interest rates); and CP is the number of months of contribution by an affiliate. Therefore, we can write the accumulated value in the AFORE as:

$$S_k = \begin{cases} (0.065BW_k + G_k) & k = 1; \\ S_{k-1}(1 + \frac{i_k^{(12)}}{12}) & k = 2, 4, \dots, CP; \\ (S_{k-1} + 0.065(BW_{k-1} + BW_k) + G_k) \\ \quad \times (1 + \frac{i_k^{(12)}}{12}) & k = 3, 5, \dots, CP. \end{cases} \quad (1)$$

Note that, for $k = 1, 2, \dots, CP$,

$$G_k = CS_{k-1} + CS_k$$

where, CS_k is defined as:

$$CS_k = \begin{cases} 0.055MW & k = 1; \\ CS_{k-1} \left(1 + \frac{\pi_k^{(4)}}{4}\right) & k = 3, 5, \dots, CP; \\ CS_{k-1} & k=2, 4, \dots, CP, \end{cases} \quad (2)$$

where $\pi^{(4)}$ is the nominal annual adjustment compounded quarterly that is in effect in the k^{th} month (every quarter the government's contribution is adjusted according to the consumer price index), and MW_k is the (national or regional) minimum wage in effect in the k^{th} month. The government's contribution is set at 5.5 percent of the minimum salary in Mexico City for the year 1997 (about U.S. \$1 per day under the exchange rate at the end of 1997).

The idea behind equation (1) is simple. Every affiliate gets his/her contribution plus the government's contribution. The way the interest is credited and the way the government's contribution is credited makes it complicated. The wage (BW_k) is added every other month. Government contributions are adjusted every three months for inflation. Thus, every third month, a bit extra is added using the consumer price index.

4.2 The Inclusion of Charges

Equation (1) does not take into account charges that funds impose on account holders (affiliates). Some AFOREs have charges on contribution as a percentage of wages (for example, Banamex). Others have charges on the balance in the AFORE account (such as Bancrecer). Still others have charges on the real interest rate (such as Inbursa).

Let CW_k be the charge on wage (rate) and CB_k be the charge on the account balance in effect in the k^{th} month. Equation (1) is modified as follows:

$$S_k = \begin{cases} \left((0.065BW_k \left(1 - \frac{CW_k}{0.065}\right) + G_k) \right. \\ \quad \times \left(1 + \frac{i_k^{(12)}}{12}\right) \left(1 - \frac{CB_k}{12}\right) & k = 1; \\ S_{k-1} \left(1 + \frac{i_k^{(12)}}{12}\right) & k = 2, 4, \dots, CP; \\ \left(S_{k-1} + 0.065(BW_{k-1} + BW_k) \left(1 - \frac{CW_k}{0.065}\right) \right. \\ \quad \left. + G_k \right) \times \left(1 + \frac{i_k^{(12)}}{12}\right) \left(1 - \frac{CB_k}{12}\right) & k = 3, 5, \dots, CP; \end{cases} \quad (3)$$

Table 7
Fee Structure of AFORES
As Charges on Annual Flow, Account Balance, and Real Returns

AFORES	Charges on		
	Annual Flow (% of Wages)	Account Balance	Real Rate Of Return
Atlantico Promex	1.40%		20.00%
Banamex	0.002% in 1997 0.85% in Jan. 1998 1.70% in March 1998 and onward		
Bancomer	1.70%		
Bancrecer Dresdner	1.60%	0.50%	
Banorte	1.00%	1.50%	
Bital	1.68%		
Capitaliza	1.60%		
Confia Principal	0.90%	1.00%	
Garante	1.68%		
Genesis	1.65%		
Inbursa			33.00%
Previnter	1.55%		
Profuturo GNP	1.70%	0.50%	
Santander	1.70%	1.00%	
XXI	1.50%	0.99%	
Tepeyac	1.17%	1.00%	
Zurich	0.95%	1.25%	

Source: CONSAR website at <http://www.consar.gob.mx>

There is a third element of charges. For two funds (Inbursa and Atlantico) charges apply to the real rate of return. Incorporating the charges on the real interest rate yields

$$S_k = \begin{cases} \left[0.065BW_k \left(1 - \frac{CW_k}{0.065} \right) + G_k \right] \\ \times \left[\left(1 + \frac{i_1^{(12)}}{12} \right) \left(1 - \frac{CB_k}{12} \right) - \frac{i_R^{(12)}}{12} CY \right] \\ \text{for } k = 1; \\ S_{k-1} \left[\left(1 + \frac{i_1^{(12)}}{12} \right) \left(1 - \frac{CB_k}{12} \right) - \frac{i_R^{(12)}}{12} CY \right] \\ \text{for } k = 2, 4, \dots, CP; \\ \left[S_{k-1} + 0.065(BW_{k-1} + BW_k) \left(1 - \frac{CW_k}{0.065} \right) \right. \\ \left. + G_k \right] \times \left[\left(1 + \frac{i_1^{(12)}}{12} \right) \left(1 - \frac{CB_k}{12} \right) - \frac{i_R^{(12)}}{12} CY \right] \\ \text{for } k = 3, 5, \dots, CP; \end{cases} \quad (4)$$

where $\pi^{(12)}$ is the annual inflation rate compounded monthly, CY is the charge on the real interest rate, and $i_R^{(12)}$ is the nonnegative real interest rate

$$i_R^{(12)} = \max \left\{ 0, \frac{i_1^{(12)} - \pi^{(12)}}{12 \left(1 + \frac{\pi^{(12)}}{12} \right)} \right\}. \quad (5)$$

One assumption made here is that the charges remain fixed for the total life of the system. In practice, however, the charges for each company depend on the number of years a person has been in the AFOR. For example, AFOR Banamex charges 1.70 percent of wages up to the fourth year. A person who stays with the AFOR for the fifth year gets a reduction in charges. Thus, the fifth year charge becomes 1.68 percent of wages; the sixth year charge becomes 1.66 percent of wages; and so on. This process continues until year 39 with the AFOR with a reduction of 0.02 percent of wages for every additional year.

The final realistic element missing from equation (4) is growth in wages. In Chile, for example, the average wage rate has grown at a rate of 6 percent per year over the last 20 years. But the rise in the average wage rate is not important here, as it represents the average across many individuals at a given point of time. For individuals, the more meaningful number is the growth of wage rate longitudinally. Hence, equation (4) must be modified to take the reductions and wage growth into account:

$$S_k = \begin{cases} \left[0.065BW_k \left(1 - \frac{CW_k(1-f_k)}{0.065} \right) + G_k \right] \\ \times \left[\left(1 + \frac{i_1^{(12)}}{12} \right) \left(1 - \frac{(1-f_k)CB_k}{12} \right) - \frac{i_R^{(12)}}{12} (1-f_k)CY \right] \\ \text{for } k = 1; \\ S_{k-1} \left[\left(1 + \frac{i_1^{(12)}}{12} \right) \left(1 - \frac{(1-f_k)CB_k}{12} \right) - \frac{i_R^{(12)}}{12} (1-f_k)CY \right] \\ \text{for } k = 2, 4, \dots, CP; \\ \left[S_{k-1} + 0.065(BW_{k-1} + BW_k) \left(1 + \frac{\Delta s_k^{(6)}}{6} \right) \left(1 - \frac{(1-f_k)CW_k}{0.065} \right) \right. \\ \left. + G_k \right] \times \left[\left(1 + \frac{i_1^{(12)}}{12} \right) \left(1 - \frac{(1-f_k)CB_k}{12} \right) - \frac{i_R^{(12)}}{12} (1-f_k)CY \right] \\ \text{for } k = 3, 5, \dots, CP \end{cases} \quad (6)$$

where f_k is the discount rate at month k , and $\Delta s_k^{(6)}$ is the annual growth rate of wages, compounded bimonthly, of an individual worker salary over his or her lifetime. Note that f_k is not the same for all funds. For example, AFORE Bancomer offers a rising discount rate starting with 0.01 percent of wages up to 0.05 percent of wages.

In some countries (Chile, South Korea), average wage rates have risen more than 6 percent in real terms per year. In others (Mexico), the average real wage rate has fallen over the past two decades. We should look at the wage rate for each individual longitudinally and not the average wage for the population.

Equation (6) is called the *comprehensive model* and will be used in the simulation study of the Mexican fee structure.

5 Simulation of the Comprehensive Model

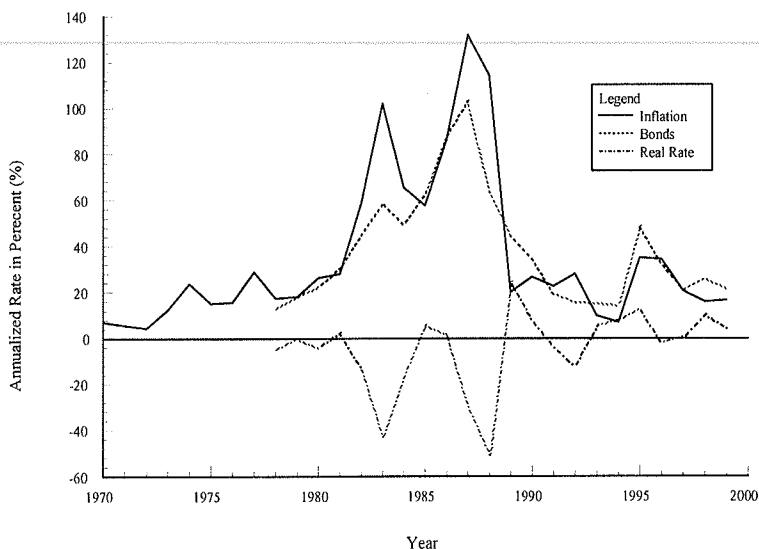
5.1 Simulation Assumptions

As the simulation is based on equation (6), assumptions must be made about many items, including the rates of return for an AFORE. Separate assumptions must be made about the rate of inflation and the real rate of return because two of the 17 AFORES have charges based on the real rate of return (Inbursa and Atlantico). The growth of individual wages rate and the specific charges that apply also must be considered in our list of assumptions.

Assumptions About Inflation and Wage Growth: The inflation rate is assumed to be constant,¹ and the real growth rate is assumed to be zero over the lifetime of the affiliates.

Assumptions About Interest: It is a daunting task to predict inflation and interest rates for a country that has seen triple digit inflation rates and negative real interest rates over a number of years in the last 20 years; see Figure 1.²

Figure 1
Annualized Inflation Rates in Mexico (1970–1999), Annualized Rates of Return for Mexican Government Bonds (CETES) (1978–1999), and Mexican Real Interest Rates (1978–1999)



Source: Banco de Mexico

¹The author experimented with stochastic inflation rates that have truncated normal and uniform distributions. For each 1,000 simulations, the majority of the cases produced results that were either identical or similar to the ones reported with constant inflation.

²Few forecasters are brave enough to predict Mexican rates more than three years. Even the Central Bank of Mexico is reluctant to venture into such an exercise!

The simulations are performed under three sets of interest rate scenarios: fixed interest rate, stochastic but time-independent interest rates, and stochastic and time-dependent interest rate. The fixed interest rate scenario is used to provide a benchmark to measure our results. A study of month-to-month changes in the (nominal) interest rate shows that they are a dependent time series process. There is clear evidence of first order autocorrelation.³ Therefore, the first order autoregressive time series model is used for interest:

$$x_t = 0.7x_{t-1} + 0.015 + \epsilon_t$$

where ϵ_t is normally distributed with mean zero and variance σ^2 . Under this assumption, the long-term interest rate converges to 5 (0.015/(1 - 0.7)) percent.

Assumptions About Charges: In Mexico commissions often are expressed as a percentage of wages and not as a percentage of contributions. Thus, if a person earns 1,000 pesos a month, the actual contribution will be 6.5 percent of 1,000 pesos or 65 pesos. Hence, the charges in some cases will be a straight percentage of the 65 pesos. Of the 17 AFOREs, 15 charge on the flow of wages. Eight of the AFOREs charge only on the wages and nothing else. These companies, therefore, do not have schemes based on performance of the funds. Regardless of the performance of the fund, charges apply. It is easy to compare across those funds: we simply choose the fund with the lowest charges. In this case, the winner is Previnter with 23.85 percent of contributions. By international standards, however, even Previnter's rate is high. In addition, there are service fees, some of which are expressed in pesos, and some of which are expressed in UDIs (these are inflation-indexed rates). Table 8 shows the discount factors obtained by staying with the same fund. Table 9 shows the various charges levied by each AFORE.

³See Sinha, T. and Escoto, Y. "Oil Price and Economic Growth: A View from the South." Paper presented at the Southern Economic Association Annual Conference, November 17-19, 2001

Table 8
Partial List of Discounts
Given by Various AFOREs

Year	Banamex	Bital	Confia	Bancrecer
1	1.70	1.68	0.90	1.60
2	1.70	1.68	0.85	1.60
3	1.70	1.68	0.80	1.60
4	1.70	1.68	0.75	1.60
5	1.70	1.68	0.70	1.60
6	1.68	1.66	0.65	1.58
7	1.66	1.64	0.60	1.56
8	1.64	1.62	0.55	1.54
9	1.62	1.60	0.50	1.52
10	1.60	1.58	0.45	1.50
11	1.58	1.58	0.45	1.48
12	1.56	1.58	0.45	1.46
13	1.54	1.58	0.45	1.44
14	1.52	1.58	0.45	1.42
15	1.50	1.58	0.45	1.40
16	1.48	1.58	0.45	1.38
17	1.46	1.58	0.45	1.36
18	1.44	1.58	0.45	1.34
19	1.42	1.58	0.45	1.32
20	1.40	1.58	0.45	1.30
21	1.38	1.58	0.45	1.28
22	1.36	1.58	0.45	1.26
23	1.34	1.58	0.45	1.24
24	1.32	1.58	0.45	1.22
25	1.30	1.58	0.45	1.20
26	1.28	1.58	0.45	1.18
27	1.26	1.58	0.45	1.16
28	1.24	1.58	0.45	1.14
29	1.22	1.58	0.45	1.12
30	1.20	1.58	0.45	1.10

Table 8 (continued)
Partial List of Discounts
Given by Various AFOREs

Year	Banamex	Bital	Confia	Bancrecer
31	1.18	1.58	0.45	1.08
32	1.16	1.58	0.45	1.06
33	1.14	1.58	0.45	1.04
34	1.12	1.58	0.45	1.02
35	1.10	1.58	0.45	1.00
36	1.08	1.58	0.45	0.98
37	1.06	1.58	0.45	0.96
38	1.04	1.58	0.45	0.94
39	1.02	1.58	0.45	0.92
40	1.00	1.58	0.45	0.90
41	0.98	1.58	0.45	0.88
42	0.96	1.58	0.45	0.86
43	0.94	1.58	0.45	0.84
44	0.92	1.58	0.45	0.82
45	0.90	1.58	0.45	0.80
46	0.88	1.58	0.45	0.78
47	0.86	1.58	0.45	0.76
48	0.84	1.58	0.45	0.74
49	0.82	1.58	0.45	0.72
50	0.80	1.58	0.45	0.70
51	0.78	1.58	0.45	0.68
52	0.76	1.58	0.45	0.66
53	0.74	1.58	0.45	0.64
54	0.72	1.58	0.45	0.62
55	0.70	1.58	0.45	0.60
56	0.68	1.58	0.45	0.58
57	0.66	1.58	0.45	0.56
58	0.64	1.58	0.45	0.54
59	0.62	1.58	0.45	0.52
60	0.60	1.58	0.45	0.50

Table 9
Commissions as Percentages of Contribution

AFORE	Commissions as a Percentage of Wage	Charges as a Percentage of Contributions
Banamex	1.70%	26.15%
Bancomer	1.70%	26.15%
Profuturo	1.70% plus others	26.15% plus others
Santander	1.70% plus others	26.15% plus others
Bitel	1.68%	25.85%
Garante	1.68%	25.85%
Genesis	1.65%	25.38%
Previnter	1.55%	23.85%
XXI	1.50% plus others	23.08% plus others
Capitaliza	1.50%	23.08%
Atlantico	1.40%	21.54%
Tepeyac	1.17% plus others	18.00% plus others
Banorte	1.00% plus others	15.38% plus others
Zurich	0.95%	14.62%
Confia	0.90% plus others	13.85% plus others
Bancrecer	Charges on balance	Charges on balance
Inbursa	Charges on real return	Charges on real return

Source: CONSAR website at <<http://www.consar.gob.mx>>

5.2 Results of the Simulations

Though the simulations are performed under various scenarios with fixed interest rates, stochastic but independent interest rates, and stochastic-dependent independent interest rates, only the results of the the deterministic case are presented here.

For most income levels, Inbursa performs the best at the beginning because Inbursa's charges are based only on account balances, and balances are usually small in the early stages. Funds that charge on contributions only have the opposite trend: their charges appear relatively high when the balance is low (compared with the contributed amount). Three factors determine how the balance grows: (1) the real interest rate, (2) the level of income, and (3) the inflation rate.

- **Impact of real interest rate:** If the real interest rate is high and stays high (for example, more than 6 percent), the charges of Inbursa become significant within five to ten years. If the real interest rate is low (3 percent or less), Inbursa remains the top performer for 20 years.
- **Impact of income level:** If the income level rises, the cost benefits from staying with Inbursa rise. For example, for persons earning the minimum wage, the benefits of low fees from Inbursa evaporate after ten years. But, for people earning at least ten times the minimum wage, the benefits (such as lower management fees) from staying with Inbursa are evident for 20 years.
- **Impact of inflation rate:** Except for Inbursa, all other funds charge a fee regardless of how well the funds are performing. (Atlantico's charges are based on the real rate and the contribution.) Therefore, if the real rate is zero or negative, Inbursa will not charge anything, while other funds will still charge a fee.

The simulation results show that no single fund dominates all others under all scenarios. Our results do, however, suggest an interesting strategy: it is optimal to switch to a different fund after ten to 20 years (depending on level of income). The best fund to shift to depends the person's level of income and the level of real interest rates.

We do not show each fund's accumulated values under each scenario because the actual accumulated values are scenario dependent. Instead, the overall ranking of each fund is reported to see if any fund dominates. Clearly the rankings do not tell us how far apart the funds are in their final balances, nor do they tell us how accumulated values compare with a fund with zero fees. After 25 years or so, the differences between consecutively ranked funds are in the order of magnitude of 1 to 3 percent.

Tables 10, 11, and 12 show the best performing AFOREs for various levels of interest, inflation, and salaries. For example, Panel A of Table 10 shows that Inbursa is the best performing fund when the nominal interest rate is 3 percent and inflation is 0 percent and a person with income equivalent to the minimum salary leaves his or her money in the AFORE for five years. For investments for five, ten, and 15 years, Inbursa is the best performer. The best AFORE with 0 percent inflation is Zurich, but Banamex leads in other scenarios. A 3 percent real rate is used in Table 10 because the Mexican government's national development plan projects a long-term real rate of 3 percent in Mexico.

Table 10
Different Scenarios with a 3% Real Interest Rate and a Minimum Salary of 768.5

Panel A: Initial Wage = Minimum Salary								
Nominal Rates	Inflation	Time (in years)						
		5	10	15	20	25	30	35
3%	0%	Inbursa	Inbursa	Inbursa	Inbursa	Zurich	Zurich	Zurich
		Confia	Confia	Zurich	Zurich	Banamex	Banamex	Banamex
		Bancrecer	Zurich	Confia	Banamex	Inbursa	Previnter	Previnter
9%	6%	Inbursa	Inbursa	Inbursa	Banamex	Banamex	Banamex	Banamex
		Confia	Confia	Banamex	Previnter	Previnter	Previnter	Previnter
		Bancrecer	Banamex	Previnter	Inbursa	Zurich	Capitaliza	Capitaliza
15%	12%	Inbursa	Inbursa	Inbursa	Banamex	Banamex	Banamex	Banamex
		Confia	Banamex	Banamex	Previnter	Previnter	Previnter	Previnter
		Zurich	Confia	Previnter	Capitaliza	Capitaliza	Capitaliza	Capitaliza
21%	18%	Inbursa	Inbursa	Inbursa	Banamex	Banamex	Banamex	Banamex
		Confia	Banamex	Banamex	Previnter	Previnter	Previnter	Previnter
		Zurich	Previnter	Previnter	Capitaliza	Capitaliza	Capitaliza	Capitaliza

Table 10 (continued)

Different Scenarios with a 3% Real Interest Rate and a Minimum Salary of 768.5

Panel B: Initial Wage = 10 × Minimum Salary								
Nominal		Time (in years)						
Rates	Inflation	5	10	15	20	25	30	35
3%	0%	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Zurich
		Bancrecer	Bancrecer	Bancrecer	Bancrecer	Bancrecer	Zurich	Inbursa
		Confía	Confía	Confía	Zurich	Zurich	Bancrecer	Banamex
9%	6%	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Zurich
		Bancrecer	Bancrecer	Bancrecer	Zurich	Zurich	Zurich	Banamex
		Confía	Confía	Confía	Bancrecer	Bancrecer	Banamex	Inbursa
15%	12%	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Zurich	Zurich
		Bancrecer	Bancrecer	Bancrecer	Zurich	Zurich	Inbursa	Banamex
		Confía	Confía	Zurich	Bancrecer	Banamex	Banamex	Previnter
21%	18%	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Zurich	Zurich
		Bancrecer	Confía	Zurich	Zurich	Zurich	Inbursa	Banamex
		Confía	Bancrecer	Confía	Banamex	Banamex	Banamex	Previnter

Table 10 (continued)
Different Scenarios with a 3% Real Interest Rate and a Minimum Salary of 768.5

Panel C: Initial Wage = 100 × Minimum Salary									
Nominal Rates	Inflation	Time (in years)							
		5	10	15	20	25	30	35	
3%	0%	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa
		Bancrecer	Bancrecer	Bancrecer	Bancrecer	Bancrecer	Zurich	Zurich	
		Confia	Confia	Confia	Zurich	Zurich	Bancrecer	Bancrecer	
9%	6%	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Zurich	
		Bancrecer	Bancrecer	Bancrecer	Bancrecer	Zurich	Zurich	Inbursa	
		Confia	Confia	Confia	Zurich	Bancrecer	Banamex	Banamex	
15%	12%	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Zurich	
		Bancrecer	Bancrecer	Bancrecer	Zurich	Zurich	Zurich	Inbursa	
		Confia	Confia	Confia	Bancrecer	Banamex	Banamex	Banamex	
21%	18%	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Inbursa	Zurich	
		Bancrecer	Bancrecer	Bancrecer	Zurich	Zurich	Zurich	Inbursa	
		Confia	Confia	Zurich	Bancrecer	Banamex	Banamex	Banamex	

Table 11
Different Scenarios with a 6% Real Interest Rate and a Minimum Salary of 768.5

Panel A: Initial Wage = Minimum Salary								
Nominal		Time (in years)						
Rates	Inflation	5	10	15	20	25	30	35
6%	0%	Inbursa	Inbursa	Zurich	Zurich	Zurich	Zurich	Banamex
		Confia	Confia	Previnter	Banamex	Banamex	Banamex	Zurich
		Bancrecer	Zurich	Banamex	Previnter	Previnter	Previnter	Previnter
12%	6%	Inbursa	Inbursa	Banamex	Banamex	Banamex	Banamex	Banamex
		Confia	Confia	Previnter	Previnter	Previnter	Previnter	Previnter
		Bancrecer	Banamex	Capitaliza	Capitaliza	Capitaliza	Capitaliza	Capitaliza
18%	12%	Inbursa	Inbursa	Banamex	Banamex	Banamex	Banamex	Banamex
		Confia	Banamex	Previnter	Previnter	Previnter	Previnter	Previnter
		Zurich	Previnter	Capitaliza	Capitaliza	Capitaliza	Capitaliza	Capitaliza
24%	18%	Inbursa	Banamex	Banamex	Banamex	Banamex	Banamex	Banamex
		Confia	Previnter	Previnter	Previnter	Previnter	Previnter	Previnter

Table 11 (continued)

Different Scenarios with a 6% Real Interest Rate and a Minimum Salary of 768.5

Nominal		Time (in years)						
Rates	Inflation	5	10	15	20	25	30	35
6%	0%	Inbursa	Inbursa	Bancrecer	Bancrecer	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Inbursa	Zurich	Bancrecer	Bancrecer	Banamex
		Confia	Confía	Confía	Confia	Confia	Banamex	Previnter
12%	6%	Inbursa	Inbursa	Inbursa	Zurich	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Bancrecer	Bancrecer	Banamex	Banamex	Banamex
		Confia	Confía	Confía	Confia	Previnter	Previnter	Previnter
18%	12%	Inbursa	Inbursa	Inbursa	Zurich	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Zurich	Banamex	Banamex	Banamex	Banamex
		Confia	Confía	Bancrecer	Previnter	Previnter	Previnter	Previnter
24%	18%	Inbursa	Inbursa	Inbursa	Zurich	Zurich	Zurich	Banamex
		Bancrecer	Confía	Zurich	Banamex	Banamex	Banamex	Zurich
		Confia	Bancrecer	Confía	Previnter	Previnter	Previnter	Previnter

Table 11 (continued)

Different Scenarios with a 6% Real Interest Rate and a Minimum Salary of 768.5

Panel C: Initial Wage = 100 × Minimum Salary								
Nominal		Time (in years)						
Rates	Inflation	5	10	15	20	25	30	35
6%	0%	Inbursa	Inbursa	Bancrecer	Bancrecer	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Inbursa	Zurich	Bancrecer	Bancrecer	Banamex
		Confía	Confía	Confía	Confía	Confía	Banamex	Previnter
12%	6%	Inbursa	Inbursa	Inbursa	Bancrecer	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Bancrecer	Zurich	Bancrecer	Banamex	Banamex
		Confía	Confía	Confía	Confía	Banamex	Previnter	Previnter
18%	12%	Inbursa	Inbursa	Inbursa	Zurich	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Bancrecer	Bancrecer	Banamex	Banamex	Banamex
		Confía	Confía	Confía	Inbursa	Previnter	Previnter	Previnter
24%	18%	Inbursa	Inbursa	Inbursa	Zurich	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Bancrecer	Inbursa	Banamex	Banamex	Banamex
		Confía	Confía	Zurich	Banamex	Previnter	Previnter	Previnter

Table 12 (continued)

Different Scenarios with a 9% Real Interest Rate and a Minimum Salary of 768.5

Panel B: Initial Wage = 10 × Minimum Salary								
Nominal		Time (in years)						
Rates	Inflation	5	10	15	20	25	30	35
9%	0%	Inbursa	Bancrecer	Bancrecer	Bancrecer	Zurich	Zurich	Zurich
		Bancrecer	Inbursa	Confía	Zurich	Bancrecer	Banamex	Banamex
		Confía	Confía	Zurich	Confía	Banamex	Previnter	Previnter
18%	9%	Inbursa	Inbursa	Bancrecer	Zurich	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Zurich	Banamex	Banamex	Banamex	Banamex
		Confía	Confía	Confía	Previnter	Previnter	Previnter	Previnter
27%	18%	Inbursa	Inbursa	Zurich	Zurich	Zurich	Banamex	Banamex
		Bancrecer	Confía	Confía	Banamex	Banamex	Zurich	Zurich
		Confía	Bancrecer	Banamex	Previnter	Previnter	Previnter	Previnter

Table 12 (continued)

Different Scenarios with a 9% Real Interest Rate and a Minimum Salary of 768.5

Panel C: Initial Wage = $100 \times$ Minimum Salary

Nominal		Time (in years)						
Rates	Inflation	5	10	15	20	25	30	35
9%	0%	Inbursa	Bancrecer	Bancrecer	Bancrecer	Zurich	Zurich	Zurich
		Bancrecer	Inbursa	Confia	Zurich	Bancrecer	Banamex	Banamex
		Confia	Confia	Zurich	Confia	Confia	Bancrecer	Previnter
18%	9%	Inbursa	Inbursa	Bancrecer	Zurich	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Confia	Bancrecer	Banamex	Banamex	Banamex
		Confia	Confia	Zurich	Confia	Previnter	Previnter	Previnter
27%	18%	Inbursa	Inbursa	Zurich	Zurich	Zurich	Zurich	Zurich
		Bancrecer	Bancrecer	Bancrecer	Banamex	Banamex	Banamex	Banamex
		Confia	Confia	Confia	Previnter	Previnter	Previnter	Previnter

Not surprisingly, the rankings change when the scenarios change. Once again, Inbursa does well for short time periods such as five or ten years. Banamex is better for all the long horizon scenarios. For money invested in an AFOREs for ten years when there is a 6 percent nominal interest rate and 0 percent inflation rate, for example, Confia comes out at the top, followed by Zurich and Banamex.

If the real interest rate stays high (9 percent) for a number of years, the advantage of Inbursa erodes. There is no single winning AFORE under all possible alternatives.

6 Alternatives to the Decentralized Pension Model

The Chilean-influenced model adopted by Mexico is not the only model available. Other models have been tried successfully in different countries. The two most cited alternatives are the Singaporean Central Provident Fund (CPF) model and the employer-based Australian-Swiss model.

6.1 The Singaporean Central Provident Fund

As the name suggests, the CPF model has only one fund. This fund is centralized and controlled by the government. Investment by the CPF has been mainly in foreign government bonds and some foreign stocks. The real rate of return for the fund has been less than 3 percent per year over a period of 25 years. The transactions cost has been low as well.

To implement the Singaporean model, people have to have faith in government. In Mexico (and in other parts of Latin America), the population has had little faith in government. In the past governments in these countries have not been efficient or open. Therefore, implementing a model with a central and crucial role for the government was not a viable option.

There have been criticisms of the Singapore model. Two comparisons can be made: one with other private pension funds operating in Singapore and the other with holding a mostly-bonds fund. On both counts, CPF account holders are penalized 1 to 3 percent per annum (Valdés-Prieto, 1998).

6.2 The Australian-Swiss Employer Based Model

In the Australian-Swiss model each employer (rather than each employee) chooses a fund. Every employee for the employer is assigned the same fund. In this case, the transactions cost is low. Funds do not have to seek each account holder. They can concentrate on a few thousand employers rather than on millions of employees. Therefore, the costs of obtaining additional accounts are significantly lower. In these systems of pension, there is some choice by the superannuation account holders. Each pension fund is floated as a separate entity. In each entity employees (mostly through the unions) choose half of the members of the board of directors, and the employer chooses the rest. Hence, it is possible for workers to have (at least) indirect influence on the fund. From the complaints received by the Commissioner of Superannuation in Australia, it seems that many persons are deeply dissatisfied with the lack of choice. As a result, new legislation is being considered that would force each superannuation fund to offer a menu of at least five separate funds for employees.

Early evidence on management fees in Australia indicates that costs are low. A recent study conducted by the Association of Superfunds of Australia reveals that earlier estimates may have severely underestimated management fees. This study, reported by Quinlivan (1998), argues that the pension fund industry in Australia has approximately \$350 billion under management (Australian dollars). Cost of administration and management is estimated at \$4 billion. The annual inflow is around \$33 billion. Therefore, charges are 12 percent of annual inflow and 1.15 percent of account balance. These charges in Australia are not spectacularly lower than what we observe in Latin America. The results from Murthi et al., (1999) for the United Kingdom are similar. The cost of fund management (without including fees for changing funds) is of the same order of magnitude in the United Kingdom.

7 Closing Comments

In studying the Mexican model, we compare the performance of the AFORE funds under various economic scenarios. The results show that an optimal strategy for individuals who want to minimize the impact of Mexico's high management fees is to switch funds periodically. The point of switching depends on assumptions about the scenarios. Moreover, in some scenarios, the optimal strategy is to switch more than

once. Still, the effect of these fees is to reduce workers' retirement funds by as much as 30 percent.

In contrast, Mitchell (1999) conducts a set of simulations with a person earning average income and without the discount given to persons with long tenure in one fund. Mitchell (page 16) wrongly concludes that "... Plan ranking by commissions prove rather stable across simulated holding periods and interest rates."

As other countries contemplate privatization of their public pension systems they must be wary because privatization is a double edged sword. First, privatization brings the risk of adverse selection that is well known in the insurance literature. Second, privatization does not solve the problem of transition generation, the obligations of the government to pay the promised benefits under the old pay-as-you-go system. If the government issues bonds to finance the transition, then the system is not privatized (Espinosa and Sinha, 2000). Third, if privatization entails huge transaction costs, then another (perhaps more insidious) problem may have been created.

There are generally two circumstances under which high transactions cost or low rates of return credited to workers' funds may be obscured: (i) during periods of rapidly growing wages, or (ii) during periods of rapidly growing contribution rates.

In Chile, high transactions costs were obscured by the rapidly growing wage rate. In addition, the real rates of return on the funds were high. Therefore, account holders ignored costs because the growth in their fund balances was high.

In Singapore the real wage rate grew rapidly in the late 1980s while the rate of contributions grew rapidly (from 11 percent of salary to 45 percent of salary) over a period of 25 years. Yet during the same period low rates of return were credited on the workers' accounts. Account holders did not protest because their balances grew steadily.

In closing, we must emphasize that differences exist among the funds and using calculated average fees tend to mask these differences.

References

- Agosin, M.R., Gustavo, C.T., and Leonardo. L.S. "Análisis Sobre el Aumento del Ahorro en Chile." Inter-America Development Bank, Working Paper No. R-309 (1997).
- Banco de México. "The Mexican Economy." Mexico, D.F., Mexico (1996).

- Diamond, P.A. "Privatization of Social Security: Lessons from Chile." *Revista de Análisis Económico* 9 (June 1994): 21-34.
- Diamond, P.A. "The Economics of Social Security Reform." National Bureau of Economic Research Working Paper No. 6719 (1998).
- Diamond, P.A. "Administrative Costs and Equilibrium Charges with Individual Accounts." National Bureau of Economic Research Working Paper No. 7050 (1999).
- Edwards, S. "The Chilean Pension Reform: A Pioneering Program." National Bureau of Economic Research Working Paper No. 5811 (1996).
- Espinosa, M., and Sinha, T. "A Primer and Assessment of Social Security Reform in Mexico." *Quarterly Review of the Federal Reserve Bank of Atlanta* (First Quarter, 2000): 1-23.
- Holzmann, R. "Pension Reform, Financial Market Development, and Economic Growth: Preliminary Evidence from Chile." International Monetary Fund Working Paper: WP/96/94 (August 1996): 34.
- Kellison, S.G. *The Theory of Interest* Second Edition. Homewood, Ill.: Irwin, 1991.
- Mitchell, O. "Administrative Costs in Public and Private Retirement Systems." National Bureau of Economic Research Working Paper No. 5734 (August 1996).
- Mitchell, O. "Evaluating Administrative Costs in Mexico's AFOREs Pension System." Pension Research Council Working Paper No. PRC WP 99-1. The Wharton School, University of Pennsylvania, 1999.
- Murthi, M., Orszag, J.M. and Orszag, P.R. "The Charge Ratio on Individual Accounts: Lessons from the U.K. Experience." Birkbeck College Working Paper 2/99 (March 1999).
- Piñera, J. "A Chilean Model for Russia." *Foreign Affairs* (September/October 2000): 1-12.
- Quinlivan, B. "Costs Too Big a Slice of the Super Cake Superfunds." *Superfunds* (November 1998): 5-9. (Available online at <<http://www.asfa.asn.au/press/sfs9811.htm>>).
- Sales-Sarrapy, C., Solís-Soberón, F., and Villagómez-Amerzcua, A. "Pension System Reform: The Mexican Case." Paper presented for the National Bureau of Economic Research Working Project on Social Security (June 1996).
- Schwarz, A., and Demirguc-Kunt, A. "Taking Stock of Pension Reforms Around the World." World Bank Working Paper (1999).

- Shah, H. "Towards Better Regulation of Private Pension Funds." Policy Research Working Paper No. 1791, World Bank (1997).
- Sinha, T. *Pension Reform in Latin America and Its Lessons for International Policymakers*. Boston, Mass.: Kluwer Academic Publishers, 2000.
- Social Security Administration. *Social Security Programs Throughout the World*. SSA Publication No. 13-11805 (Washington, DC, Social Security Administration, 1999).
- Valdés-Prieto, S. "Private Sector in Social Security: Latin American Lessons for APEC." Paper presented at the APEC Regional Forum on Pension Fund Reforms (1998).
- Vittas, D. "Sistema Swiss Chilanpore, ¿El Camino Hacia Una Reforma de las Pensiones?" In Antonio Ruezga Barba (ed.) *Administracion Publica y Privada de Los Seguros Sociales en America Latina*. Conferencia Interamericana de Seguridad Social, Mexico DF. (1994).
- Vittas, D. "Private Pension Funds in Hungary: Early Performance and Regulatory Issues." World Bank Policy Research Working Paper No. 1638 (1996).
- Von Gersdorff, H. "The Bolivian Pension Reform." World Bank, Policy Research Working Paper No. 1832 (1997).
- World Bank. *Averting the Old Age Crisis*. New York, N.Y.: Oxford University Press, 1994.

Premium Earning Patterns for Multi-Year Policies with Aggregate Deductibles

Thomas Struppeck*

Abstract[†]

Multi-year policies with large aggregate deductibles or multiple triggers raise some interesting issues about the correct amount of unearned premium reserve that a company should carry. Examples in this paper illustrate some of the difficulties that arise when trying to establish such reserves. The basic approach taken here is that the pure premium portion of the unearned premium reserve should always be adequate to cover the remaining risk. This approach, however, can lead to some unusual and controversial earning patterns; there are even situations where a negative premium is earned. In addition, the earning pattern for a particular loss scenario can differ materially from the earning pattern that is expected when the contract is written.

Key words and phrases: *unearned premium, unearned premium reserve, premium deficiency reserves, required pure premium reserve*

*Thomas Struppeck, Ph.D., F.C.A.S., A.S.A., M.A.A.A., is an actuary with Centre Re. He obtained his B.Sc. from Tulane University. After graduating he served as a computer consultant specializing in digital cartography. He obtained his Ph.D. in number theory from the University of Texas at Austin and then taught mathematics at Rutgers University. Dr. Struppeck's first actuarial position was with Republic Insurance in Dallas. He then joined Centre Re in its Zurich, Switzerland office. There he priced international reinsurance treaties before transferring to his current position in Centre's New York office. His current assignment includes capital allocation and financial reporting duties.

Dr. Struppeck's address is: Centre Insurance Co., One Chase Manhattan Plaza, 35th Floor, New York, NY 10005, USA. Internet address: tom.struppeck@zurich.com

[†]The author would like to thank the reviewers and colleagues who read and commented on early versions of this paper.

1 Introduction

Statutory accounting requires that reserves be established for covered losses that have occurred but are unpaid (loss reserves) and effectively for losses that have not yet occurred, but will be covered by policies already on the books (unearned premium reserves). Furthermore, these reserves need to be separate.

A problem can arise, however, when a multi-year contract has a large aggregate deductible. If losses depleting the deductible occur faster than expected, the premium reserve at some point may be inadequate. Of course, it is also possible that those losses occur more slowly than anticipated, in which case the premium reserve may be redundant.

To deal with this potential problem with multi-year contracts, we recommend that at each point in time (or at the end of each accounting period) the pure premium portion of the unearned premium reserve should be adequate. This, in turn, implies a certain earning pattern for the premium that, in some cases, requires that a negative premium be earned.

The problems associated with the adequacy of the pure premium reserve were captured in the spirit of a hypothetical question put forward by Ruy Cardoso on CASNET in 1999. Mr. Cardoso's question is paraphrased here:

Losses are certain at \$10 per month. You cover \$20 excess \$100 in aggregate. The contract begins 7/1/xx. What is the loss reserve at 12/31/xx (ignore investment income)?

Most of this paper illustrates, using numerical examples, some of the consequences of taking the "adequate pure premium reserve" approach to establishing the unearned premium reserve (UEPR). These examples are designed to illustrate how the experience early in a multi-year excess contract affects the expected losses (to the contract, not ground-up)¹ that occurs later in the contract and how this, in turn, should affect premium reserving and earning patterns. While the examples could be made more realistic, such realism could introduce complications not relevant to the central issue. For example, in our simplification of Ruy Cardoso's question above, we assume that there are certain losses of \$20 per month. If the losses are certain, there are questions of risk-transfer.

Similarly, in Section 4, the single premium policy has an indefinite term—even though such a policy would be highly unusual. Despite

¹The losses "ground-up" refer to the losses from first dollar. The losses to the contract are those losses (limited by the limit) that are above the attachment point.

the simplifications, the examples and the technical considerations they illustrate are relevant.

Section 7 provides some comments on practical considerations, including remarks relevant to the new requirement that an actuary opine on the adequacy of the unearned premium reserve under certain circumstances.

In some cases, the approach contained herein might result, for example, in earning a premium faster than some state's regulations would allow. Naturally, one should consult with qualified accounting professionals to decide how to properly record the financials of complex or difficult contracts.

2 The Unearned Premium Reserve (UEPR)

2.1 What is Unearned Premium?

According to the glossary of the 1994 property-casualty insurance accounting text published by the Insurance Accounting and Systems Association (IASA), "Unearned premium [is] the portion of the premium applicable to the unexpired period of the policy." What is the unearned premium reserve (UEPR)? Again from the glossary, "The sum of all premiums representing the unexpired portions of the policies or contracts which the insurer or reinsurer has on its books as of a certain date . . ." UEPR is a liability that represents the premium for the unexpired risks on the insurer's books.

The American Academy of Actuaries' *Statement of Principles Regarding P&C Insurance Ratemaking* (1999) states that ratemaking is prospective, and that a rate is an estimate of the expected value of future costs. Also, a rate provides for all costs associated with the transfer of risk. This paper is concerned primarily with the pure premium portion of the rate—i.e., the expected loss and loss adjustment expense, not including other expenses.

Combining these two concepts, we see that UEPR consists of the pure premiums and the other expenses for the unexpired portion of the risks that are currently on the insurer's books. From one valuation date to another, the amount of unexpired risk on an insurer's books changes: new risks may be written, and the unexpired portion of those risks that were on the books at the beginning of the period generally decreases. This is captured in the familiar accounting identity:

$$EP = WP + UEPR_{\text{begin}} - UEPR_{\text{end}} \quad (1)$$

where:

- EP = The premium earned during the period;
- WP = The premium written during the period;
- UEPR_{begin} = UEPR at the beginning of the period; and
- UEPR_{end} = UEPR at the end of the period.

Thus, other things being equal, UEPR_{end} is inversely related to the amount of premium earned. Should it happen that the UEPR_{end} for a certain policy is larger than its UEPR_{begin} without any new premium being written (we shall see below how this might happen), then equation (1) implies that the premium earned on the policy during this particular period is negative.

2.2 Example 1

We now turn to the question of the indicated UEPR for multi-year policies. For ease of exposition, let's first examine a simplified version of the problem. We will assume

- there are no reporting lags and that losses are paid as they are incurred;
- there is a maximum of one loss in each year, each loss is exactly \$1,000;
- there is no investment income; and
- the probability that a loss occurs in any given year is 10 percent and that different years are independent.

For this simplified set of assumptions, we want to compute the pure premium for the k^{th} loss during the next n years; we will denote this pure premium $PP(k, n)$. Let Policy (k, n) denote a policy covering the k^{th} loss.

To illustrate:

- $PP(1, 1)$ is the pure premium for a policy that pays \$1,000 if there is at least² one loss during year one, so $PP(1, 1) = \$1,000 \times 0.1 = \100 .

²In this first example, there can be only one loss per year so for the first year "at least one" implies "exactly one."

- $PP(1, 2)$ is the pure premium for a policy that pays \$1,000 if there is a loss during year one or year two (as we discount flows at 0 percent it does not matter which). The probability that there is no loss in two years is $0.9^2 = 81$ percent, so the probability of at least one loss is 19 percent and $PP(1, 2) = \$190$.
- $PP(2, 2)$ is the pure premium for a policy that pays \$1,000 if there are at least two losses during years one and two. As we are assuming at most one loss per year, this can happen only if there is exactly one loss in each of years one and two. The probability of this is $0.10 \times 0.10 = 1$ percent and the pure premium is \$10.

Suppose that you purchased both Policy (1,2) and Policy (2,2). You would have full coverage for two years. In fact, your coverage would be identical to first purchasing Policy (1,1) and then one year later purchasing a second Policy (1,1). Your pure premium for the first set of policies would be $\$190 + \$10 = \$200$. For the second your pure premium would be $\$100 + \$100 = \$200$ once more. This is no coincidence. Identical coverages must have identical pure premiums.

In a world ignoring transaction costs, risk and profit loads, and other expenses, where risk carriers are willing to cede or assume risks for their pure premiums, the following principle holds: If two sets of policies give identical coverage, they must have the same premium charge. If this were not so, a portfolio consisting of a long position (assumed risk) and a short position (ceded risk) could be assembled that has positive net (pure) premium, but no net risk. This would violate the economic principle of no risk-free arbitrage, also referred to as the *no arbitrage principle*.

2.3 The Required Pure Premium Reserve (RPPR)

The pure premium for a policy is equal to the expected losses at contract inception. As time passes, however, the pure premium for the remaining losses will change. The *required pure premium reserve* (RPPR) is the expected future losses (ignoring transactions costs and other expenses) over the remaining lifetime of the insurance contract. The required pure premium reserve at time t ($RPPR_t$) is the amount that a hypothetical risk carrier would require to assume the risk at time t , ignoring transactions costs and other expenses. $RPPR_t$ may depend on the loss experience up to time t .

At policy inception, the required premium reserve equals the pure premium for the policy. At policy termination, when no more losses can occur, the required premium reserve is zero. (Here and throughout

the paper we assume that losses are paid as they are incurred and that there is no reporting lag.) RPPR is similar to the unearned premium reserve (UEPR), but it has one important difference. UEPR contains premium elements other than pure premium (such as expense loads and risk loads). In world with no transactions costs, an exactly adequate UEPR is equal to RPPR; in the following discussion the terms are used interchangeably.

RPPR may depend on loss experience, as the following continuation of example 1 illustrates. The RPPR for Policy (1,2) at time $t = 0$ is the pure premium, which we computed above as \$190. After one year, we are in one of two states:

State	Probability	RPPR ₁
Loss	10%	No more cover remains; RPPR ₁ = 0
No Loss	90%	Remaining cover is Policy (1,1); RPPR ₁ = 100.

The decrease in RPPR during the first year is analogous to the (pure) premium earned during that period. The decrease in RPPR in the loss case is 190 and in the no-loss case is 90. The probability of the loss case is 10 percent, so the expected change in RPPR is $0.1 \times 190 + 0.9 \times 90 = 100$, which must be equal to the pure premium for a one-year cover (i.e., the coverage that you receive during the first year of Policy (1,2)). In fact, it is always true that the a priori expected value of the change in RPPR during a period is equal to the a priori expected value of the losses occurring during that period.

In the above example, expected losses are \$100 and the expected change in RPPR is also \$100. While the expected change is \$100, an actual change of \$100 is not possible in this example. (It is either \$90 or \$190.)

3 The Adequate Pure Premium Reserve Approach

Using this approach, the change in RPPR is a correct measure for pure premium earned during the period, and the pure premium portion of UEPR should be RPPR. Applying this approach to the example of the previous section: in the no-loss case, we would earn premium of \$90 during the first period. In the loss case we would earn premium of \$190.

Under current accounting rules: in the loss case, because there is no more cover, all future premiums would be accrued and earned in the current period,³ so earned (pure) premium would be \$190, just as the adequate pure premium reserve approach indicates. In the no-loss case, I believe that most companies would simply earn half of the pure premium (\$95) during the first year (and some might recognize that they have a \$5 premium deficiency, as the pure premium for year two is \$100).

My view is that at policy inception we expect to earn \$100, but that in fact we earn either \$190 or \$90 depending on our experience. This however can lead to some odd results.

Consider the expected change in RPPR for Policy (2,2) during year one. This policy pays \$1,000 for the second loss in two years. The pure premium for this policy is \$10, so this is RPPR at time 0.

After one year we are again in one of two states:

State	Probability	RPPR ₁
Loss	10%	Remaining cover is Policy (1,1); RPPR ₁ = 100
No Loss	90%	As there can be only one loss per year, there can now be no second loss: RPPR ₁ = 0.

In the no-loss case, which occurs 90 percent of the time, the decrease in RPPR is \$10. In the loss case, the decrease in RPPR for Policy (2,2) is -\$90. The expected decrease in RPPR is $0.9 \times 10 + 0.1 \times -\$90 = 0$.

The premium earning principle tells us that this must be the expected value of losses occurring during the first year. Does this make sense? Yes! This policy pays only on the second loss, and because we assume there can be only one loss per year, the second loss cannot occur during year one. That is why the expected losses during year one are zero.

3.1 Standard Premium-Accrual Methodology Considerations

I am not certain how companies would account for the above cover today. Some would argue that because the second loss cannot occur in

³Under U.S.-GAAP, at least for reinsurers, this is the content of EITF93-6, Issue 3 "How should the ceding and assuming companies account for changes in future coverage resulting from experience under the reinsurance contract?"

year one, no premium should be earned in year one on this cover; they would earn all \$10 in year two. Others might earn \$5 in the first year and \$5 in the second year.

I would argue that in the no-loss case all \$10 should be earned in the first year, but that in the loss case $-\$90$ should be earned in the first year. The adequate pure premium reserve approach implies that the amount of pure premium earned during a period must be that amount such that the remaining RPPR contains exactly the expected pure premium required for the remaining policy period given the losses that have occurred to date.

At inception, the company's expectation is to earn nothing during year one on this policy because the insured event could not occur during this period. But in fact one of two things happens: they have either an underwriting gain of \$10 or an underwriting loss of \$90.

The standard premium accrual procedure referred to earlier (i.e., accruing all future premium when no more cover remains) together with an application of the no arbitrage principle leads to the same conclusion as the adequate pure premium reserve approach, as we will now illustrate.

Recall that the portfolio consisting of Policy (1,2) and Policy (2,2) together give identical coverage to the portfolio consisting of Policy (1,1) along with a one year deferred Policy (1,1). By the no arbitrage principle, the premiums and how they are earned should be the same. During year one, the premium earned on Policy (1,1) is equal to 100. The premium earned during year one on each of Policy (1,2) and Policy (2,2) depends on the results of year one:

(i) The Loss Case:

Probability = 10 percent

Policy (1,2) earns a premium of \$190 implies

Policy (2,2) earns a premium of $-\$90$

or

(ii) The No-Loss Case:

Probability = 90 percent

Policy (2,2) earns a premium of \$10 implies

Policy(1,2) earns a premium of \$90.

In the loss case, the premium earned on Policy (1,2) is \$190 by the standard premium accrual procedure. Using the no arbitrage principle,

because the total premium earned on the two policies during year one must be \$100, the premium earned on Policy (2,2) must be $-\$90$.

Similarly, in the no-loss case, the premium earned on Policy (2,2) should be all \$10, because no coverage remains. No arbitrage forces the premium earned on Policy (1,2) to be \$90, because the sum must be \$100.

If one is uncomfortable with earning all of the premium for Policy (2,2) in the no-loss case in year one, consider what happens to the pair of policies in year two given that there is no loss in year one. The coverage is identical to the coverage afforded by a one year deferred Policy (1,1), so the earned premium in year two must be the same: \$100. The coverage during year two for Policy (1,2) is the same as for a Policy (1,1) because we are given that there is no loss in year one. The premium earned on Policy (1,2) during year two must be \$100. Because the total premium earned is also \$100, no premium can be earned on Policy (2,2). Over the life of Policy (2,2) \$10 must be earned; if none is earned in year two, all of it must be earned in year one.

3.2 Reconciling Total Earnings

The total amount of pure premium earned during the life of the policy is always equal to the initial pure premium. If some negative premium is earned during one period, it is recovered in later periods (or is balanced by some overearning in prior periods). The total change in RPPR from contract inception to contract termination is the a priori pure premium. This is an important point. The negative premium earned is not new premium, the written premium stays the same—it is just earned in a different pattern.⁴

UEPR for a given policy is amortized over the policy's term. This amortization occurs according to some amortization schedule. For most lines of business this amortization schedule is linear over the term. This linearity produces the familiar pro-rata earning pattern. This pattern is theoretically correct for a policy with no aggregate deductible, no aggregate limit, and an underlying loss process that has a compound Poisson distribution. For a further discussion of compound distributions see, for example, Bowers et al., (1997, Chapter 12). For certain lines of business (e.g., extended warranty, ocean marine cargo cover, credit insurance on a declining balance) other amortization patterns and, hence, earning patterns are used. The adequate unearned

⁴It should be noted that the process of setting the UEPR to the currently required pure premium reserve is nothing more than a mark-to-market of the outstanding UEPR.

premium reserve process described above can be thought of as adjusting this amortization schedule to include the latest data.

Traditionally, one thinks of unearned premium reserves flowing into loss reserves and surplus as the policy term progresses. Sometimes the losses occur more slowly than expected, and an unexpectedly large portion of this flow goes to surplus. Other times losses occur more rapidly than expected, and (unfortunately) in these cases surplus may flow into loss reserves. In the example above, it is the unearned premium reserve, not the loss reserve, that has become inadequate and requires supplementation from surplus.

4 More Examples

4.1 Example 2: A Less Simplified Example

This example allows for more than one loss in each year. For simplicity, we assume that in each year there are 0, 1, or 2 losses with probabilities $1/2$, $1/3$, and $1/6$, respectively. Losses are still constant but the constant loss amount will be \$216 instead of \$1,000. We continue to ignore investment income.

The pure premiums for Policy (k, n) may be computed as follows. First compute the probability of having exactly k losses by the end of year n ; the result of this calculation⁵ is displayed in Table 1. Then sum the probabilities in Table 1 to produce the probability of having at least k losses in n years; see Table 2 for these values. Finally, multiply the probabilities in Table 2 by the constant loss amount of \$216 to compute the pure premiums shown in Table 3.

Consider Policy (2,3), which covers the second loss in three years. The pure premium for this coverage is \$135. How much of this premium do we expect to earn during the first year?

Half of the time there will be no loss during the first year, and RPPR for the last two years of the policy must be \$90—the pure premium for Policy (2,2). In this case $\$135 - \$90 = \$45$ would be earned in the first year.

⁵The probabilities are most easily computed recursively. For example:

$$\Pr(2, 2) = 1/2 \times \Pr(2, 1) + 1/3 \times \Pr(1, 1) + 1/6 \times \Pr(0, 1).$$

That is, the only way to have exactly two losses at the end of year two is to have had no loss in year two and exactly two losses in year one, or exactly one loss in year two and one loss in year one, or two losses in year two and no loss in year one. (Here the events joined by “and” are independent and the events joined by “or” are mutually exclusive.)

Similarly, one-third of the time there will be one loss during the first year; then RPPR for the last two years must be \$162 (the pure premium for Policy (1,2), which is equivalent to the remaining coverage) and $\$135 - \$162 = -\$27$ would be earned during the first year.

Finally, one-sixth of the time there are two losses in year one. In this case there is no more coverage available. RPPR for the last two years is zero, and the full \$135 would be earned during year one.

Combining the above calculations for the first year earned premiums we find that *at policy inception* the expected earned premium for year one is

$$1/2 \times \$45 + 1/3 \times -\$27 + 1/6 \times \$135 = \$36.$$

Year three's expected earnings are similarly easy to calculate: during the first two years of the cover there is a $1/2 \times 1/2 = 1/4$ chance that there have been no losses and a $1/2 \times 1/3 + 1/3 \times 1/2 = 1/3$ chance of exactly one loss. From Table 2, we see that the pure premium for Policy (2,1) is 36 and for Policy (1,1) is 108. From this we see that *at policy inception* we expect to earn $1/4 \times \$36 + 1/3 \times \$108 = \$45$ during year three.

During the life of the policy we will earn exactly \$135. If *at policy inception* we expect to earn \$36 in year one and \$45 in year three, it follows that we must expect *at policy inception* to earn $\$135 - \$36 - \$45 = \54 during year two.

Does this mean that we should earn the premium over the three years in this pattern: \$36, \$54, \$45? No, because these are a priori expectations. As we have seen in earlier sections, the premium earned during year one need not equal the a priori expected earned premium. Also, at the end of year one our expectations for the earnings in years 2 and 3 will probably be different than they were at inception.

The first two rows of Table 3 contain all the information needed to compute the actual amount of premium earned to date at the end of each year. For example, suppose there is exactly one loss, and it occurs in year two. Then we should earn \$45 in the first year, because when we start year two, the remaining coverage is the second loss in two years: a Policy (2,2). During year three we are in a first-loss position, so we need to earn \$108 because at the start of year three, the remaining coverage is the first loss in one year: a Policy (1,1). Because the total amount earned over the three years must be \$135, we find that the year two (actual) earnings must be $-\$18$. So the actual earning pattern observed in this case would be (\$45, $-\$18$, \$108), which differs markedly from the a priori expectation.

Table 1
Probability of Exactly k Losses in n Years

k	$n = 1$	$n = 2$	$n = 3$
0	50.00%	25.00%	12.50%
1	33.33%	33.33%	25.00%
2	16.67%	27.78%	29.17%
3	0.00%	11.11%	20.37%
4	0.00%	2.78%	9.72%
5	0.00%	0.00%	2.78%
6	0.00%	0.00%	0.46%

Table 2
Probability of at Least k Losses in n Years

k	$n = 1$	$n = 2$	$n = 3$
0	100.00%	100.00%	100.00%
1	50.00%	75.00%	87.50%
2	16.67%	41.67%	62.50%
3	0.00%	13.89%	33.33%
4	0.00%	2.78%	12.96%
5	0.00%	0.00%	3.24%
6	0.00%	0.00%	0.46%

Table 3
Pure Premiums for Policy (k, n)

Loss k	$n = 1$	$n = 2$	$n = 3$
1	108	162	189
2	36	90	135
3	0	30	72
4	0	6	28
5	0	0	7
6	0	0	1

4.2 Example 3: An Indefinite-Term Example

In this example we will assume a $1/10$ chance of loss each year and return to the simplified model of at most one loss per year. Loss severity is assumed constant at \$3,000. We will continue to ignore investment income. The policy that we consider in this example covers one loss, but has no time limit. The policy will stay in effect until there is a loss, at which time it will pay \$3,000.⁶

4.2.1 Pure Premium and Earning Patterns

What is the pure premium for this coverage? Let P be this premium. Then P must pay for two things. One-tenth of the time there is a loss during year one of \$3,000 and $RPPR_1 = 0$. The other nine-tenths of the time, there is no loss in the first year, and $RPPR_1$ is the pure premium for a policy that pays \$3,000 whenever the loss occurs—but this is exactly what P is. We have:

$$P = 1/10 \times (\$3,000 + 0) + 9/10 \times (0 + P).$$

Solving for P , one finds $P = \$3,000$.

Upon reflection this is not surprising, as \$3,000 will be paid out eventually. (Recall that we are still ignoring investment income.) The pure premium equals the expected loss, which is \$3,000.

How does one earn the premium for such a policy? In the loss case, the premium earned in year one is \$3,000; in the no-loss case the premium earned in year one is \$0 (because $RPPR_1$ remains at \$3,000). At policy inception the expected earned premium for the first year is \$300.

What about later years? The answer depends on when you ask the question.

At the start of the first year, we expect to earn \$270 during the second year and \$243 during the third. But these are the a priori expectations at the start of the first year; after one year has passed there has been either one loss or no loss, and with this additional information the expected values for earned premium change.

At the start of the second year there are two possibilities: either there is a loss in year one (in which case no coverage remains) or there is no loss in year one (in which case there is coverage for year two). Also, because we are assuming no late reporting, you will know which case applies. The conditional expectation (given no loss in year one)

⁶This example is akin to a single premium whole life insurance policy.

for the premium earned in year two is \$300. Similarly, the conditional expectation (given no loss in year one) for the premium earned in year three is \$270. Similarly, the conditional expectation (given no loss in years one and two) for the earned premium in year three is \$300.

The expected earning pattern at the start of any year, for that and subsequent years, is (\$300, \$270, \$243, ...), with each term being 9/10 of the previous term. When a year passes without loss, each of these terms shifts forward. It should come as no surprise that this infinite geometric series sums to \$3,000.

Why is no premium earned during no-loss years? Because RPPR at the start of the no-loss year is \$3,000, and it is also \$3,000 at the end of the year. The change in RPPR, in this case \$0, is the earned premium. During a loss year, RPPR is \$3,000 at the start of the year, and it is \$0 at the end of the year (because no more coverage remains). The amount earned during the year is \$3,000.

The company shows no underwriting gain or loss, regardless of the outcome. In the no-loss case there is no movement in the reserves; in the loss case RPPR becomes the loss reserve. This is a consequence of the indefinite policy term. Because the cover continues until there is a loss, having a no-loss year only delays the inevitable payment; without investment income, the delay does not benefit us. We relax this restriction below.

4.2.2 The Impact of Investment Income

Let's take into account investment income. Assume that all losses are paid at the end of the year and that invested funds earn interest at a rate of 5 percent. The equation for the present value of the pure premium then becomes:

$$P = 1/10 \times \frac{\$3,000}{1.05} + 9/10 \times \frac{P}{1.05}.$$

One-tenth of the time we pay a loss of \$3,000 (discounted one year) and nine-tenths of the time the present value of RPPR₁ is P (discounted one year). Solving for P , we find that $P = \$2,000$.

How should this premium be earned? Should the fact that we now consider investment income affect how we earn the premium?

Suppose that we have a loss in year one. Then, as before, RPPR₁ = 0, so we earn the full \$2,000 during year one. We also have investment income of \$100. On the other hand, suppose that we have no loss in the

first year. Then $RPPR_1 = \$2,000$, and again we have investment income of \$100. What should be done with the investment income?

To investigate that question, we examine an alternative way to construct this same coverage. Consider an annual policy that pays \$1,000 at the end of the year if there is a loss, for a premium payable at the end of the year⁷ of \$100 (the pure premium for the policy). In effect, this policy provides similar coverage to the first year of the original policy, subject to a \$2,000 self-insured retention. Imagine that the insured sets aside this \$2,000 in a special account. During the year, \$100 in investment income is earned on the \$2,000 (this is paid to the insurer as premium) and, if there is a loss, the \$2,000 set aside and the \$1,000 from the insurer combine to provide the \$3,000.

With a one-time premium of \$2,000 and a limit of \$3,000, the insurer has only \$1,000 at risk. So in this second set-up, the insurer is entitled to only \$100 (= \$1,000 \times 10%) in annual pure premium. This, as we have seen, is the investment income generated by the one-time premium payment of \$2,000.

We see that the insured can obtain identical coverage in two ways: by setting aside the \$2,000 and paying an annual premium of \$100 in arrears or by paying a one-time premium of \$2,000. The no arbitrage principle says that because the two coverages are identical, their pure premiums must be equal. In order for this to work, we need to view the investment income on (discounted) premium as premium—this is implicit in the pricing equation.

Now we can determine the earning pattern for the original multi-year policy and answer the question about what to do with the investment income. In a year with no loss, premium of \$100 is earned. In a loss year, premium of \$2100 (the original premium plus one year's investment income) is earned.

This result is related to the paid-up insurance formula for life reserves; see, for example, Bowers et al., (1997, Chapter 7).

4.3 Example 4: An Example with Expenses

In the real world, UEPR contains many components in addition to RPPR's pure premium. There may be, for example, on-going contract maintenance expense.⁸ Effectively, such expense forms an annuity that runs until contract termination. One quick example will give a flavor of the complications.

⁷The premium is made payable at the end of the year to remove timing effects.

⁸Had these expenses have been deferred policy acquisition expenses, there would be additional accounting complications.

Recall the earlier example of an indefinite-term policy that pays \$3,000 when the loss occurs, has annual loss probability of 10 percent, and no investment income. Assume that on-going contract maintenance expense is \$150 per year. Letting G stand for the expense-loaded premium, the premium equation now reads:

$$G = 1/10 \times (\$3,000 + \$150) + 9/10 \times (G + \$150).$$

That is, one-tenth of the time we have expenses of \$150 and a loss of \$3,000, and the other nine-tenths of the time we have expenses of \$150 and $RPPR_1 = G$ (because of the indefinite term). Solving for G , we find that $G = \$4500$.

The company with this risk on its books suffers an underwriting loss (after expenses) of \$150 each year that there is no loss, but has an underwriting gain of \$1350 the year that the loss occurs!⁹

The interested reader may find it amusing to calculate the effect on this example of including 5 percent investment income.

5 Some Practical Ramifications

Though the preceding examples illustrate some of the theoretical issues, the practicing actuary must consider the broader practical effects of any change to common practice. Questions of materiality and practicality also should be addressed.

5.1 Actuarial Reserve Opinions

The National Association of Insurance Commissioners' (NAIC) *SAO Instructions for Property-Casualty* (1998) specifies that the SCOPE paragraph include the reserve for direct, ceded, and net unearned premiums. It also specifies that these three items must be covered in the opinion and relevant comments paragraphs. This applies to all insurers that write direct and/or assumed contracts or policies (excluding financial guaranty, mortgage guaranty, and surety contracts) with terms of

⁹What's happening here is that we have an annuity with an expected life of ten years funding the expenses. When we have a no-loss year, the expected life of the annuity stays at ten (instead of decreasing to nine) and we show an underwriting loss of the difference. When we have a loss year, the expense annuity is no longer needed (its expected life drops from ten to zero). The release of the reserve supporting this annuity yields the underwriting gain.

13 months or more, which the insurer cannot cancel and for which the insurer cannot increase premiums during the term.

The insurer is required to establish an adequate unearned premium reserve. For each of the three most recent policy years, the gross unearned premium reserve must be no less than the largest result of three tests. The three tests (in slightly simplified form) are:

1. The best estimate of the amounts refundable to the contract holders at the reporting date.
2. The gross premium multiplied by the ratio of (a) over (b) where:
 - (a) Equals the projected future gross losses and expenses to be incurred during the unexpired term of the contracts; and
 - (b) Equals the projected total gross losses and expenses under the contracts.
3. The amount of the projected future gross losses and expense to be incurred during the unexpired term of the contracts (as adjusted), reduced by the present value of the future guaranteed gross premiums.

The examples in this paper are intended to be non-cancelable insurance contracts with fixed premiums. The contract terms are more than 13 months in length. The rule applies, except for the proscribed lines of business. How do our examples fare under these tests?

For simplicity, we shall assume that there are no refund provisions in the policy, so the Test 1 lower bound on the unearned premium reserve is zero.

Test 2 requires that we estimate gross losses and expense. The examples in this paper for the most part have been concerned with pure premiums (i.e., only the expected losses, with no provision for expenses). Under the simplifying assumption that expenses are zero, Test 2 tells us to estimate the projected future gross loss to be incurred and to divide this by the projected total gross loss. This ratio is then multiplied by the gross premium to obtain the second lower bound on the unearned premium reserve.

Test 3 requires that the unearned premium reserve be at least as large as the expected future losses and expenses to be incurred during the contract (as adjusted). The amount of the projected future gross losses to be incurred is exactly RPPR at the statement date. The adjustments in question are for future premiums and for investment income until the loss is incurred but not beyond. Our examples have no future

premiums and our losses are assumed to be immediately payable. [The test also specifies a company-specific maximum interest rate. We will assume that 5 percent meets this test.]

In our examples, RPPR is the lower bound on the unearned premium reserve specified by Test 3.

5.2 Perspectives on Aggregate Deductible Business

In a multi-year contract with an aggregate deductible, the experience of the first few years can influence the required premium reserve in two ways. First, the aggregate deductible may be depleted more rapidly or more slowly than planned; second, adverse or favorable experience during the initial period may influence one's view of the future ground-up experience. This paper addresses only the former.

There is an additional way to view such policies. The later years of a multi-year policy with an aggregate deductible can be thought of as excess layers, each year/layer having a retention that depends on the earlier years' experience. If the total losses to date have been small, little of the aggregate deductible has been eroded and the retention (the remaining aggregate deductible) for the later years is higher. Because higher layers have lower premiums, RPPR is small. Similarly, if early experience has been unfavorable, much of the aggregate deductible will have been eroded. The retention will be lower and RPPR will be large. In essence, early experience determines to which layers the later years' coverage corresponds.

5.3 What to Do About Negative Premium?

In chapter 14 of the IASA text, David L. Holman and Chris C. Stroup discuss U.S.-GAAP accounting for P&C insurers. Under U.S.-GAAP there is a notion of a premium deficiency reserve (PDR). Holman and Stroup write:

Projections, therefore, are periodically updated, based on new information about expected cash flows. GAAP requires that a premium deficiency be recognized if the sum of expected loss and loss adjustment expenses, expected dividends to policyholders, maintenance costs, and unamortized (or deferred) policy acquisition costs, exceed the related unearned premiums related thereto.

If there is a deficiency, the unamortized policy acquisition costs are reduced to make up the shortfall. If that alone is not sufficient, a liability

is reported for the remaining deficiency. Interestingly, Canadian statutory accounting provides a line item (Line 15) for premium deficiency (see chapter 18 of the IASA text). European actuaries speak of the reserve for unexpired risks, which is similar in concept to a combined unearned premium and premium deficiency reserve.

So, under U.S.-GAAP one might establish a PDR to handle negative premium earnings. Effectively, a negative premium is earned by the reduction of an asset (the unamortized policy acquisition cost) and/or the establishment of an additional liability.

Statutory accounting does not have the notion of a premium deficiency, although in principle one could include one by using the write-in lines. Due to U.S. income tax regulation, there may be a material difference between treating the shortfall as premium or as some other type of liability. The interested reader should see chapter 13 of the IASA text or Almagro and Ghezzi (1988).

5.4 Is It Loss or Is It Premium?

The argument can be made that instead of altering the premium earning methodology, we should establish loss reserves corresponding to the losses that are eroding the aggregate deductible. That is, there is an increase in expected losses to the cover caused by events that have occurred prior to the statement date. The amounts are not in dispute; they would be exactly the amount needed to make the booked reserve match RPPR. The difference is that these reserves would be characterized as loss instead of premium.

But these reserves behave more like premium than loss in two important ways. First, they amortize over the remaining policy period. To see the second reason, consider a two-trigger two-year policy. In order for the policy to pay, two events, A and B, must occur during a two-year period. Say event A occurs in year one, and as a result some additional reserve (either a loss reserve or a premium deficiency reserve) is needed. Suppose now you wanted to completely reinsure this risk. You could do this by purchasing cover for event B. Observe that this reinsurance is completely prospective. Being prospective, it should be funded from premium reserves, not loss reserves.¹⁰

¹⁰Claims-made policies and sunset clauses in reinsurance agreements can further blur the line between premium reserves and loss reserves. Suppose that an event has occurred, but that it has not been reported yet. Assuming that a reserve is appropriate, should it be premium or loss? This reserve amortizes over the remaining reporting period (acts like premium). On the other hand, the underlying loss event has already occurred. Is the reporting a second trigger?

6 Conclusions

We could use the adequate pure premium reserve approach to answer Mr. Cardoso's question, which was mentioned in Section 1 above: Losses are certain at \$10 per month. You cover \$20 excess \$100 in aggregate. The contract begins 7/1/xx. What is the loss reserve at 12/31/xx?

Assuming no expenses or investment income, UEPR would be \$20 (because that is RPPR remaining), and the loss reserve would be \$0 (because no covered loss has occurred). No premium (positive or negative) would have been earned to date.

The adequate pure premium reserve approach outlined in this paper is internally consistent, even though it leads to some controversial implications such as negative earned premium. But the idea of negative earned (and written) premium already is used in some instances, such as the treatment of ceded proportional reinsurance. U.S.-GAAP and Canadian accounting have a notion of a premium deficiency reserve (PDR), and in some European jurisdictions there is a notion of an unexpired risk reserve. These entries could be used to record unexpected changes in the required premium reserve.

There are some operational problems, however, with the negative premium approach: it may distort loss and expense ratios; it can make budgeting difficult; and, for U.S. taxpayers, the treatment of UEPR for U.S. taxation is different than for other reserves, which could lead to complications.

The good news is that, on average, the standard methodology should give the same results as this method for a large book of uncorrelated risks, written evenly throughout the year. The analysis outlined in this paper is probably justified for those risk carriers with a few large risks or for single risks that are large enough to distort the book.

References

- Almagro, M., and Ghezzi, T.L. "Federal Income Taxes-Provisions Affecting Property/Casualty Insurers." *Proceedings of the Casualty Actuarial Society* 75 (1988): 95-162.
- American Academy of Actuaries. *Property/Casualty Loss Reserve Law Manual 1998*. Washington, D.C.: American Academy of Actuaries, 1998.

Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A., and Nesbitt, C.J. *Actuarial Mathematics*, Second Edition. Schaumburg, Ill.: Society of Actuaries, 1997.

Casualty Actuarial Society. *1999 Yearbook*. Arlington, Va.: Casualty Actuarial Society, 1999.

Insurance Accounting and Systems Association (IASA). *Property-Casualty Insurance Accounting*, Sixth Edition. Durham, N.C.: Insurance Accounting and Systems Association, 1994.

National Association of Insurance Commissioners (NAIC). *SAO Instructions for Property-Casualty*. Kansas City, Mo.: NAIC, 1998.

Exponential Bonus-Malus Systems Integrating A Priori Risk Classification

Lluís Bermúdez,* Michel Denuit,[†] and Jan Dhaene[‡]

Abstract[§]

This paper examines an integrated ratemaking scheme including a priori risk classification and a posteriori experience rating. In order to avoid the high penalties implied by the quadratic loss function, the symmetry between the overcharges and the undercharges is broken by introducing parametric loss functions of exponential type.

Key words and phrases: quadratic loss function, exponential loss function, credibility estimation, explanatory variables, experience rating, risk classification

*Lluís Bermúdez, Ph.D., is a professor of economics at the University of Barcelona.

Dr. Bermúdez's address is: Departament de Matemàtica Econòmica i Empresarial, Facultat de Econòmiques, Universitat de Barcelona, Avda. Diagonal 690, S-08034 Barcelona, SPAIN. Internet address: lbermu@eco.ub.es

[†]Michel Denuit, Ph.D., F.A.R.A.B., is professor of actuarial mathematics at the Université Catholique de Louvain, Belgium. He received his B.Sc. in mathematics and his Ph.D. in statistics from the Université Libre de Bruxelles, Belgium, where he worked five years as a lecturer. Dr. Denuit is a fellow of the Belgian Society of Actuaries (KVBA-ARAB), a former member of the Directorial Board of the Society, a member of the Education Committee, and has been a Belgian representative to the IAA and Groupe Consultatif Education Committees.

Dr. Denuit's address is: Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, BELGIUM. Internet address: denuit@stat.ucl.ac.be

[‡]Jan Dhaene, Ph.D., is professor of actuarial mathematics at the Catholic University of Leuven, Belgium. He obtained his B.Sc. in mathematics from University of Ghent in 1985 and his Ph.D. in actuarial science from University of Leuven in 1991. His research interests include topics related to modeling dependencies, with applications in finance and insurance.

Dr. Dhaene's address is: Departement Toegepaste Economische Wetenschappen, Faculteit Economische And Toegepaste Economische Wetenschappen, Katholieke Universiteit Leuven, Minderbroedersstraat 5, B-3000 Leuven, BELGIUM. Internet address: Jan.Dhaene@econ.kuleuven.ac.be

[§]The authors would like to thank the referee for carefully reading the manuscript, as well as for several useful comments and suggestions.

1 Introduction and Motivation

1.1 A Priori Risk Classification Variables

One of the main tasks of an actuary is to design a tariff structure that fairly distributes the burden of current claims among its portfolio of current policyholders. If the insurance portfolio consists of heterogeneous risks (policyholders), then it is fair to partition the portfolio into homogeneous classes of policies with policyholders belonging to the same class paying the same premium.

The classification variables used to partition the portfolio into homogeneous cells are called a priori variables (as their values can be determined at the start of the policy). In automobile third-party liability insurance, for example, the commonly used classification variables include the age, gender, marital status, occupation, type and use of car, and residential address. Generalized linear models can be used to select the a priori classification variables.¹

In most practical situations many important factors (such as driving style, reflexes, or knowledge of rules of the road) cannot be taken into account when selecting the a priori classification variables. Consequently, even after the a priori classification variables have been chosen, tariff cells may still be heterogeneous. It is reasonable to believe, however, that these characteristics are revealed by the number and sizes of claims reported by the policyholders over the successive insurance periods. Hence, at the end of each insurance period the next period's premium is adjusted on the basis of the individual's claims experience in order to ensure fair premiums among policyholders.

It is interesting to mention that in North America, emphasis has traditionally been laid on a priori ratings using many classifying variables, while in continental Europe just a few a priori classifying variables were used and much importance was placed on the a posteriori evaluation of drivers. Since July 1994, however, European Union (EU) directives have introduced complete rating freedom. Insurance companies operating in EU countries are now (theoretically) free to set up their own rates, select their own classification variables, and design their own bonus-malus system.² Companies in most EU countries have taken advantage of this freedom by introducing more rating variables.

¹For more on generalized linear models, see, for example, Renshaw (1994) or Pinquet (1997,1999) for applications in actuarial science; or Mc Cullagh and Nelder (1989), Dobson (1990), or Fahrmeir and Tutz (1994).

²For a thorough presentation of the techniques relating to bonus-malus systems, we refer the interested reader to Lemaire (1995).

In a competitive insurance market the trend is toward portfolio segmentation because insurers tend to use all available relevant information to match the rating structure used by competitors. As the only item of interest is the unknown distribution function of the claim amounts produced by the driver during a period, it seems fair to correct the inadequacies of the *a priori* system by using an adequate bonus-malus system. Such an experience rating system should be better accepted by policyholders than arbitrary *a priori* classifications.

A bonus-malus system is a rating system based on the following mechanism:

1. Claim-free policyholders, i.e., those with zero claims within a single period, are rewarded by premium discounts called *bonuses*; and
2. Policyholders reporting one or more accidents at fault during a period are penalized by premium surcharges called *maluses*.

This a posteriori ratemaking system is an efficient way of classifying policyholders into cells according to their risk. As pointed out by Lemaire (1995), if insurers are allowed to use only one rating variable, it should be a merit rating variable because merit rating variables are the best predictor of the number of claims incurred by a driver. Besides encouraging policyholders to drive carefully (i.e., to counteract moral hazard), merit rating systems aim to better assess individual risks, so that everyone will pay in the long run a premium corresponding to her or his own claim frequency. Such systems are called no-claim discounts, experience rating, merit rating, or bonus-malus systems.

1.2 The Nature of Risk Transfers³

Consider a portfolio of automobile third-party liability insurance policies. Let Y denote a quantity of actuarial interest for a policy taken at random from the portfolio. For example, Y can be the amount of a claim, the aggregate claims in one period, or the number of accidents at fault reported by the policyholder during one period. The actuary has a set of observable risk classification variables, X , pertaining to the selected policyholder, which may include such items as age, gender, marital status, occupation, home address, type and use of her or his car. In addition, Y also depends on a set of unknown characteristics Z , which may include such items as annual mileage (i.e., risk exposure),

³The ideas presented in this section are inspired by De Wit and Van Eeghen (1984).

accuracy of judgment, aggressiveness behind the wheel, drinking behavior, etc. Some of the elements of Z are unobservable; others cannot be measured in a cost efficient way. If Ω denotes the entire set of risk factors for this policyholder then

$$\Omega = X \cup Z.$$

The true net premium for this policyholder is $\mathbb{E}[Y|\Omega]$; it is worth mentioning that this premium is a random variable but much less dispersed than Y itself, making insurance policies worth buying. The situation can be summarized as described in Table 1. In this case, the policyholder keeps the variations of the premiums due to the modifications in her or his personal characteristics Ω and transfers to the company the purely random fluctuations of Y (that is, the variance of the outcomes of Y once the personal characteristics X and Z have been taken into account). As the elements of Z are unknown to the insurer, the situation described in Table 1 is purely theoretical. Because the company only knows X , the actual reality of the insurance business is rather as depicted in Table 2.

Table 1
Risk Transfer Between Insurance Company
And Policyholder in Case of Full Information

	Amount Carried By	
	Policyholder	Insurer
Risk:	$\mathbb{E}[Y \Omega]$	$Y - \mathbb{E}[Y \Omega]$
Expectation:	$\mathbb{E}[Y]$	0
Variance:	$\text{Var}[\mathbb{E}[Y \Omega]]$	$\mathbb{E}[\text{Var}[Y \Omega]]$

It is well known to statisticians and actuaries that for a random variable A and a random vector B (possibly of dimension 1),

$$\mathbb{E}[A] = \mathbb{E}[\mathbb{E}[A|B]] \quad \text{and} \quad \text{Var}[A] = \mathbb{E}[\text{Var}[A|B]] + \text{Var}[\mathbb{E}[A|B]].$$

If we let $A = Y|X$ and $B = \Omega$, then

$$\mathbb{E}[\text{Var}[Y|X]] = \mathbb{E}[\text{Var}[Y|\Omega]] + \mathbb{E}[\text{Var}[\mathbb{E}[Y|\Omega]|X]].$$

The first term on the right, i.e., $\mathbb{E}[\text{Var}[Y|\Omega]]$, represents the purely random fluctuations of the risk and is supported by the insurance com-

pany. The second term on the right represents the variation in the expected claims due to the unknown risk characteristics \mathbf{Z} . This quantity should be corrected by an experience rating mechanism.

Table 2
Risk Transfer Between Insurance Company
And Policyholder in Case of Partial Information

	Amount Carried By	
	Policyholder	Insurer
Risk:	$\mathbb{E}[Y \mathbf{X}]$	$Y - \mathbb{E}[Y \mathbf{X}]$
Expectation:	$\mathbb{E}[Y]$	0
Variance:	$\text{Var}[\mathbb{E}[Y \mathbf{X}]]$	$\mathbb{E}[\text{Var}[Y \mathbf{X}]]$

Next, assume the insurance company incorporates more a priori variables in its pricing structure; that is, $\tilde{\mathbf{X}}$ (with $\mathbf{X} \subset \tilde{\mathbf{X}}$) is substituted for \mathbf{X} .

$$\mathbb{E}[\text{Var}[\mathbb{E}[Y|\Omega]|\tilde{\mathbf{X}}]] \leq \mathbb{E}[\text{Var}[\mathbb{E}[Y|\Omega]|\mathbf{X}]],$$

that is, the residual heterogeneity in the portfolio is reduced. Consequently, the variance of the insurer’s experience is also reduced, i.e.,

$$\mathbb{E}[\text{Var}[Y|\tilde{\mathbf{X}}]] \leq \mathbb{E}[\text{Var}[Y|\mathbf{X}]].$$

The severity of the a posteriori corrections thus decreases as the information used by the insurer increases.

1.3 Objectives

Let \mathcal{F}_t denotes the entire past claims experience available about Y at time t . The central idea behind experience rating is that \mathcal{F}_t reveals its hidden features \mathbf{Z} as $t \rightarrow \infty$, i.e., the information contained in $(\mathbf{X}, \mathcal{F}_t)$ becomes comparable to Ω as time goes on. Therefore, the a posteriori premium is $\mathbb{E}[Y|\mathbf{X}, \mathcal{F}_t]$.

The aim of this paper is to examine the interaction between a priori ratemaking (i.e., identification of the best predictors \mathbf{X} and of the risk premium $\mathbb{E}[Y|\mathbf{X}]$) and a posteriori ratemaking (i.e., premium corrections according to past claims history \mathcal{F}_t in order to reflect the unavailable information contained in \mathbf{Z}).

The paper is organized as follows: Section 2 contains a brief review of the current methodology of automobile ratemaking in EU countries. It considers risk classification and credibility as two separate problems. This approach has flaws because the aim of experience rating is to reduce the residual heterogeneity of the portfolio, which obviously depends on the degree of a priori segmentation. Therefore a priori and a posteriori ratemaking have to be integrated in a continuous risk evaluation mechanism. In Section 3, we present the results of Dionne and Vanasse (1989, 1992) and Gisler (1996), as well as an alternative approach based on an exponential loss function. Such loss functions have been considered by Ferreira (1977), Lemaire (1979), Young (1996), and Denuit and Dhaene (2001), among others.

Our methods are illustrated by an example using a Spanish insurance portfolio. This example considers only two risk factors and allows for a deeper understanding of the technical concepts introduced. Adaptation of the methodology to real-life portfolio is then straightforward. Several optimization programs are used extensively throughout this paper (some of them are standard in actuarial science, others are less common). The appendix contains a description of all results, together with proofs for the sake of completeness.

2 Current Methodology

2.1 The Model

Consider an automobile portfolio consisting of N independent policies. These policies are split into M homogeneous risk classes. The premium paid by each policyholder depends on the policyholder's rating factors for the current period and also on her or his claim history. The premium charged is the product of a risk classification base premium and of a bonus-malus coefficient. The base premium for a risk class is a function of the current rating factors, whereas the bonus-malus coefficient only depends on the policyholder's history of reported claims at fault.

We assume the insurance company determines its risk classification factor using generalized linear models; see, e.g., in Renshaw (1994). We suppose the N risks are partitioned into M distinct (disjoint) risk classes. In each risk class, the policies are identical from the company point of view, whereas policies in different risk classes have distinct risk profiles.

For $m = 1, 2, \dots, M$, the base premium for the m^{th} risk class is denoted by BP_m , which is the amount charged to a new policyholder entering the m^{th} risk class. Of course, inside each risk class, the policies are not strictly identical. Therefore, the premium is adjusted over time using a bonus-malus factor $BMF(k, t)$ where t is the number of years the policy is in force and k is the number of claims reported while the policy is in force.

Notice that while the base premium depends on the risk class, the *same* bonus-malus factor is applied to all drivers, i.e., it is independent of the risk class. This is erroneous because a bonus-malus system is supposed to correct the actual premium for the residual heterogeneity existing in the different risk classes, which implies that the severity of a bonus-malus system must depend on the policyholder's risk class. In fact, the more a priori risk factors used in the risk classification system, the less severe bonus-malus coefficients should be. Uniform bonus-malus systems imposed by regulatory authorities in some EU countries (e.g., Belgium and France) create cross-subsidization of insurance portfolios.

Let

K_{ij} = Number of claims incurred by the i^{th} policyholder during period $(j - 1, j)$;

n_{ik} = Number of policies from class i reporting k claims;

Θ_i = Risk proneness parameter of policyholder i . It captures the propensity of policyholder i to produce claims and is regarded as a random variable; and

Z_{ijk} = Size (severity) of the k^{th} claim produced by the i^{th} policyholder during year $(j - 1, j)$.

At the portfolio level, the vectors $(\Theta_i, K_{i1}, K_{i2}, K_{i3}, \dots)$ are assumed to be independent and identically distributed for $i = 1, 2, \dots, N$. Also, given $\Theta_i = \theta$, the random variables $K_{i1}, K_{i2}, K_{i3}, \dots$ are assumed to be independent and identically distributed for fixed i . Unconditionally, these random variables are dependent. For fixed i , the Z_{ijk} s are assumed to be independent and identically distributed and independent of the claim frequencies K_{ij} . This assumption has been questioned by several authors because it implies that the cost of an accident is, for the most part, beyond the control of a policyholder. Though the degree of care exercised by a driver may mostly influence the number of accidents, it has less influence on the cost of these accidents. Nevertheless,

this assumption seems acceptable in third-party liability insurance. The Z_{ijk} are also independent of Θ_i for any given i .

The total claim amount for policyholder i in year j is

$$S_{ij} = \sum_{k=1}^{K_{ij}} Z_{ijk}.$$

We put $\mathbb{E}[Z_{ijk}] = 1$, which means that the expected claim amount is chosen as monetary unit. The pure premium for policy i in year j is then given by

$$\mathbb{E}[S_{ij}|\Theta_i = \theta] = \mathbb{E}[K_{ij}|\Theta_i = \theta] = \theta.$$

A priori (i.e., without information about claims history), an identical amount of premium $\mathbb{E}[\Theta_i]$ is charged to new policyholders.

Given $\Theta_i = \theta$, the numbers of claims generated in $(j-1, j)$ by policyholder i are assumed to be independent and identically distributed (i.i.d.) Poisson random variables with mean θ , i.e.,

$$\mathbb{Pr}[K_{ij} = k|\Theta_i = \theta] = \exp(-\theta) \frac{\theta^k}{k!}, \quad (1)$$

where θ is the claim frequency of this policyholder. The cumulative distribution function (cdf) of Θ_i , $F_{\Theta}(\cdot)$, (often called the *structure function*), belongs to the two-parameter gamma family, i.e.,

$$F_{\Theta}(\theta) = \Gamma(\theta|\alpha, \tau) \quad (2)$$

where

$$\Gamma(\theta|\alpha, \tau) = \int_0^{\theta} \frac{\tau^{\alpha} v^{\alpha-1} e^{-\tau v}}{\Gamma(\alpha)} dv, \quad \alpha, \tau, \theta > 0. \quad (3)$$

Combining equations (1) and (2) yields the well-known result that the number of claims for a policyholder randomly drawn from the portfolio follows a negative binomial distribution, i.e.,

$$\mathbb{Pr}[K_{ij} = k] = \frac{k + \alpha - 1}{k} \left(\frac{\tau}{1 + \tau} \right)^{\alpha} \left(\frac{1}{1 + \tau} \right)^k. \quad (4)$$

Though K_{i1}, K_{i2}, \dots are identically distributed, they are not independent, because they are generated by the same policyholder and thus contingent on the same risk parameter Θ_i .

2.2 A Posteriori Premiums

Suppose policyholder i has been observed for t years and the number of claims reported during this period is $k_{i1}, k_{i2}, \dots, k_{it}$. The premium for year $t + 1$ is defined as a function Ψ_t of the claims reported during the previous years, which is determined by minimizing $\mathbb{E}[L(\Theta_i - \Psi_t(k_{i1}, k_{i2}, \dots, k_{it}))]$ for some loss function L , taken to be non-negative, convex, and such that $L(0) = 0$. The loss functions considered in this paper are the quadratic loss where $L(x) = x^2$ and the exponential loss with positive parameter c where $L(x) = \exp(-cx)$.

From the results recalled in the appendix, we easily get the following proposition.

Proposition 1. *The best estimator of the pure premium Θ_i at time $t + 1$ is given by*

$$W_{t+1}^{(q)} = \frac{\alpha}{\tau} (1 - \rho_q) + \frac{k_{i\bullet}(t)}{t} \rho_q \tag{5}$$

for the quadratic loss function where

$$\rho_q = \frac{t}{\tau + t} \quad \text{and} \quad k_{i\bullet}(t) = \sum_{j=1}^t k_{ij}(t) \quad \text{while}$$

$$W_{t+1}^{(e)} = \frac{\alpha}{\tau} (1 - \rho_e(c)) + \frac{k_{i\bullet}(t)}{t} \rho_e(c) \tag{6}$$

for the exponential loss function with $c > 0$, and

$$\rho_e(c) = \frac{t}{c} \ln \left(1 + \frac{c}{\tau + t} \right). \tag{7}$$

Notice that in Proposition 1, both expressions for W_{t+1} are convex combinations of the portfolio mean α/τ and the observed average number of claims $k_{i\bullet}(t)/t$ over the period $[0, t]$. In both cases the weight given to the past claims tends to 1 as t goes to ∞ . The weight given to the claim history with the exponential loss function is smaller than the weight given to the claim history a quadratic loss function. i.e.,

$$\frac{t}{c} \ln \left(1 + \frac{c}{\tau + t} \right) \leq \frac{t}{\tau + t}.$$

Note that in the Poisson-gamma model, the Bayesian approach coincides with the linear credibility estimator. In other words, Proposition

1 can be interpreted in a semi-parametric framework, as in the classical Bühlmann-Straub approach. Notice that

$$\lim_{c \rightarrow 0} W_{t+1}^{(e)} = W_{t+1}^{(q)},$$

i.e., the a posteriori premium associated with the exponential loss function converges to that associated with the quadratic loss function.

Also, as $c \rightarrow +\infty$ we have that

$$\lim_{c \rightarrow +\infty} \rho_e(c) = 0 \text{ so that } W_{t+1}^{(e)} \rightarrow \alpha/\tau.$$

This provides an intuitive meaning of the parameter c : if c increases, then the a posteriori merit-rating scheme becomes less severe, and at the limit, the premium no longer depends on the incurred claims. Moreover, routine calculations show that

$$\frac{d}{dc} \rho_e(c) < 0,$$

so that the weight given to the observed average claim number decreases as c increases.

Let $I_i(t) \in (1, 2, \dots, M)$ denote the index of the risk class occupied by policyholder i during year t . Now, the a posteriori premium for year $t + 1$ (i.e., for the time period $(t, t + 1)$) charged to policyholder i having reported $k_{i\bullet}(t)$ claims during the first t years is given by

$$P_{t+1}^{(q)}(k_{i\bullet}(t), t) = \text{BP}_{I_i(t+1)} \text{BMF}^{(q)}(k_{i\bullet}(t), t) \quad (8)$$

with

$$\begin{aligned} \text{BMF}^{(q)}(k_{i\bullet}(t), t) &= \frac{W_{t+1}^{(q)}}{\mathbb{E}[\Theta_t]} \\ &= \frac{\alpha + k_{i\bullet}(t)}{\tau + t} \times \frac{\tau}{\alpha} \end{aligned} \quad (9)$$

under a quadratic loss. Under an exponential loss, we get

$$P_{t+1}^{(e)}(k_{i\bullet}(t), t) = \text{BP}_{I_i(t+1)} \text{BMF}^{(e)}(k_{i\bullet}(t), t) \quad (10)$$

with

$$\begin{aligned} \text{BMF}^{(e)}(k_{i\bullet}(t), t) &= \frac{W_{t+1}^{(e)}}{\mathbb{E}[\Theta_i]} \\ &= 1 - \frac{t}{c} \ln \left(1 + \frac{c}{\tau + t} \right) \\ &\quad + \ln \left(1 + \frac{c}{\tau + t} \right) \frac{k_{i\bullet}(t)}{c} \frac{\tau}{\alpha}. \end{aligned} \quad (11)$$

The model used to determine the bonus-malus coefficients assumes that all the risks of the portfolio have the same a priori claim frequency and that the differences in the claim frequency between the risks are only due to differences in the individual risk characteristics Θ_i . Hence, the model implicitly assumes that the tariff takes into account differences in claim frequencies only through the bonus-malus payments and that such differences are not reflected in the base premiums.

This approach is erroneous because the aim of the bonus-malus system is to adjust the amount of premium according to past claim experience. The effect of this premium adjustment is to reduce the residual heterogeneity within the different risk classes of the portfolio. As the bonus-malus coefficients of Proposition 1 do not take into account explanatory variables, they are functions of the total heterogeneity of the portfolio, before tariff segmentation. In other words, the bonus-malus factors penalize bad risks and reward good risks.

2.3 A Numerical Illustration

Consider Table 3, which displays data from a Spanish insurance company. As can be seen from Table 3, policies have been categorized into 12 classes according to the age of the driver (three categories) and the power of the car (four categories). The three age categories are “Age ≤ 35 ,” “ $36 \leq \text{Age} \leq 49$,” and “Age ≥ 50 .” The four power categories are “Power ≤ 53 ,” “ $54 \leq \text{Power} \leq 75$,” “ $76 \leq \text{Power} \leq 118$,” and “Power ≥ 119 .”

Let n_{ik} represent the number of policies from class i reporting k claims, $i = 1, 2, \dots, 12$, and

$$n_{i\bullet} = \sum_{k=0}^{\infty} n_{ik}$$

is the number of policies in the i^{th} class, $i = 1, 2, \dots, 12$.

Again, we assume that the number of claims reported by a policyholder in class i during a year follows a Poisson distribution with mean

Table 3
The Twelve Risk Classes
For Classification Factors Age and Power

Power of Car (In Horsepower)	Age of Driver (in Years)		
	Age \leq 35	36 \leq Age \leq 49	Age \geq 50
Power \leq 53	1	2	3
54 \leq Power \leq 75	4	5	6
76 \leq Power \leq 118	7	8	9
Power \geq 119	10	11	12

Table 4
Observed Mean Claim Frequencies
For Classification Factors Age and Power

Power of Car (In Horsepower)	Age of Driver (in Years)		
	Age \leq 35	36 \leq Age \leq 49	Age \geq 50
Power \leq 53	0.1866	0.1572	0.1283
54 \leq Power \leq 75	0.2685	0.2279	0.1986
76 \leq Power \leq 118	0.2992	0.2526	0.2386
Power \geq 119	0.3217	0.2846	0.2483

λ_i . Moreover, the random variables K_{i1}, K_{i2}, \dots are assumed to be independent. Therefore, the total number of claims $K_{i\bullet} = \sum_{j=1}^{n_i} K_{ij}$ reported by the n_i policyholders in class i has a Poisson distribution with mean $n_i \lambda_i$. The realization of $K_{i\bullet}$ is $k_{i\bullet} = \sum_{k=1}^{\infty} k n_{ik}$.

Next we introduce the indicator variable J_{ik} such that

$$J_{ik} = \begin{cases} 1 & \text{if policyholder } i \text{ is in age category } k \text{ for } k = 2, 3; \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, define L_{ik} as

$$L_{ik} = \begin{cases} 1 & \text{if policyholder } i \text{ drives a car in category } k \text{ for } k = 2, 3, 4; \\ 0 & \text{otherwise} \end{cases}$$

The i^{th} policyholder is represented by a vector of classification information:

Table 5
Observed Claims Distribution of Number of Policyholders
Submitting k Claims in the i^{th} Risk Class, n_{ik} , $i = 1, 2, \dots, 12$
The First Six Risk Classes, $i = 1, 2, \dots, 6$

k	n_{1k}	n_{2k}	n_{3k}	n_{4k}	n_{5k}	n_{6k}
0	3,316	7,797	10,437	9,470	21,031	22,788
1	548	1,063	1,159	1,916	3,775	3,766
2	61	140	143	445	720	591
3	15	17	15	84	143	109
4	4	6	2	21	36	24
5	1	0	1	7	11	5
6	0	0	1	0	2	4
7	0	0	0	1	1	0
8	0	0	0	3	0	0
≥ 9	0	0	0	0	0	0
$n_{i\cdot} =$	3,945	9,023	11,758	11,947	25,719	27,287
$k_{i\cdot} =$	736	1,418	1,509	3,208	5,862	5,420
$\bar{x}_i =$	0.1866	0.1751	0.1283	0.2685	0.2279	0.1986
$s_i^2 =$	0.227	0.1828	0.1501	0.3635	0.2946	0.2451

$$\mathbf{X}_i = (1, J_{i2}, J_{i3}, L_{i2}, L_{i3}, L_{i4}),$$

and a corresponding vector of unknown parameters is

$$\boldsymbol{\eta}^T = (\epsilon, \gamma_2, \gamma_3, \delta_2, \delta_3, \delta_4)$$

where T denotes the transposed matrix.

When the claim numbers are small, which is typically the case in automobile insurance, the normal approximation is poor and fails to account for the discreteness of the data. Normal regression should be avoided in this case. Generalized linear models provide an appropriate framework for the analysis of discrete data. A linear model for $\ln(\lambda_i)$ is often used in actuarial science. [See, e.g., Pinquet (1997)]. This provides a regression model for count data analogous to the usual normal regression for continuous data. In addition, the standard methodology of generalized linear models uses the logarithmic function as the natural link function for the Poisson distribution. [See, e.g., Dobson (1990).] Thus, we specify a linear model for $\ln(\lambda_i) + \ln(n_{i\cdot})$ as

Table 5 (Continued)
Observed Claims Distribution of Number of Policyholders
Submitting k Claims in the i^{th} Risk Class, n_{ik} , $i = 1, 2, \dots, 12$
The Second Six Risk Classes, $i = 7, 8, \dots, 12$

k	n_{7k}	n_{8k}	n_{9k}	n_{10k}	n_{11k}	n_{12k}
0	6,570	15,702	15,158	1,125	4,554	4,680
1	1,423	3,112	2,848	274	902	900
2	321	603	510	69	224	187
3	89	148	123	9	55	25
4	33	31	33	7	15	12
5	6	11	11	1	9	5
6	3	2	1	1	2	1
7	1	0	3	0	0	1
8	1	0	1	0	1	1
≥ 9	0	0	0	0	0	0
$n_{i\bullet} =$	8,447	19,609	18,688	1,486	5,762	5,812
$t_{i\bullet} =$	2,527	4,953	4,459	478	1,640	1,443
$\bar{x}_i =$	0.2992	0.2526	0.2386	0.3217	0.2846	0.2483
$s_i^2 =$	0.4322	0.3288	0.3200	0.4376	0.4214	0.3408

$$\ln(\lambda_i) + \ln(n_{i\bullet}) = \mathbf{X}_i \boldsymbol{\eta} = \epsilon + \sum_{k=2}^3 \gamma_k J_{ik} + \sum_{k=2}^4 \delta_k L_{ik}. \quad (12)$$

In order to determine the maximum likelihood estimator of the parameter $\boldsymbol{\eta}$, we have to maximize $L(\boldsymbol{\eta})$ where

$$L(\boldsymbol{\eta}) = \prod_{i=1}^{12} \exp(-\lambda_i n_{i\bullet}) \frac{(\lambda_i n_{i\bullet})^{k_{i\bullet}}}{k_{i\bullet}!}.$$

The regularity conditions satisfied by the Poisson distribution ensure there is a unique solution to the system of equations $\partial \ln L / \partial \boldsymbol{\eta} = 0$. It is easy to check that the maximum likelihood estimator $\hat{\boldsymbol{\eta}}$ of the parameter $\boldsymbol{\eta}$ is the solution of the equations

$$\sum_{i=1}^{12} (k_{i\bullet} - n_{i\bullet} \lambda_i) X_{ij} = 0 \quad (13)$$

for $j = 1, 2, \dots, 6$, where X_{ij} is the j^{th} element of \mathbf{X}_i . As pointed out by Pinquet (1997), equation (13) can be interpreted as an orthogonality relation between the residuals and the covariates. The estimates and the standard deviation and 95% confidence interval for the estimates are displayed in Table 6.

As the rating factors have a finite number of levels and the explanatory variables are indicators of these levels, equation (13) implies that, for every sub-portfolio corresponding to a given level, the sum of the fitted claim numbers is equal to the total number of claims incurred in that sub-portfolio for the observation period. As an example, equation (13) with $j = 2$ ensures that as far as policyholders in age category 2 are concerned the sum of the fitted claim frequencies equals the total number of claims. Consequently, such a system is expected not to create cross-subsidization in the portfolio

Table 6
Parameters Estimates of η in Equation (12)

η	$\hat{\eta}$	Standard Deviation	95% Confidence Interval
ϵ	-1.7219	0.0198	[-1.7607, -1.6831]
γ_2	-0.1634	0.0147	[-0.1922, -0.1345]
γ_3	-0.2800	0.0149	[-0.3093, -0.2508]
δ_2	0.3987	0.0185	[0.3625, 0.4350]
δ_3	0.5324	0.0189	[0.4953, 0.5694]
δ_4	0.6150	0.0236	[0.5688, 0.6611]

It is well known that the vector $\hat{\eta}$ is approximately normal for large sample sizes, with mean η and variance-covariance matrix $\hat{\mathbf{V}}$, which is the inverse of the Fisher information matrix. The element (j, k) of $\hat{\mathbf{V}}$ is

$$V_{jk} = \sum_{i=1}^{12} X_{ij} X_{ik} n_i \lambda_i.$$

Computing the variance-covariance matrix yields

$$\hat{\mathbf{V}}^{-1} = 10^{-3} \begin{pmatrix} 0.392 & -0.144 & -0.151 & -0.277 & -0.277 & -0.265 \\ -0.144 & 0.217 & 0.145 & 0.002 & 0.000 & -0.014 \\ -0.151 & 0.145 & 0.223 & 0.008 & 0.009 & -0.006 \\ -0.277 & 0.002 & 0.008 & 0.342 & 0.274 & 0.273 \\ -0.277 & 0.000 & 0.009 & 0.274 & 0.357 & 0.273 \\ -0.265 & -0.014 & -0.006 & 0.273 & 0.273 & 0.555 \end{pmatrix}$$

Considering Table 6, all the parameters are significantly different from 0 (because no confidence interval overlaps 0), so that all the covariates are statistically significant. The expected claim numbers for each of the 12 cells are given in Table 7. (It is interesting to compare the fitted results to their empirical counterparts given in Table 3.) Table 7 thus gives the base premiums attached to each of the 12 risk classes.

Table 7
Estimated Mean Claim Frequencies
Based on Classification Factors Age and Power

Power of Car (In Horsepower)	Age of Driver (in Years)		
	Age ≤ 35	36 ≤ Age ≤ 49	Age ≥ 50
Power ≤ 53	0.1787	0.1518	0.1351
54 ≤ Power ≤ 75	0.2663	0.2262	0.2013
76 ≤ Power ≤ 118	0.3044	0.2585	0.2300
Power ≥ 119	0.3306	0.2808	0.2498

In order to calculate the bonus-malus factors, let us consider the claim distribution for the whole portfolio, which is given in Table 8. The negative binomial is fitted using the maximum likelihood approach and is displayed in the third column. The a posteriori premiums are then given by equations (8) and (10) with the estimated values of α and τ given by $\hat{\alpha} = 0.8665$ and $\hat{\tau} = 3.9097$.

Consider for instance a 30-year-old female driver whose car is in the power category "≤ 53." Her a priori expected number of accidents is 0.1787 for the first five years; upon reaching age 35 her expected number of accidents becomes 0.1518. In the first half of Table 9, one can see the bonus-malus coefficients and premiums for that individual. The second column (entitled "BP_t") represents the expected number of accidents (i.e., the base premium) for each period. The BMF_{t+1} column represents the bonus-malus factor in case the policyholder does not cause

Table 8
Observed and Fitted Claim
Distribution Using Data in Table 5

k	n_k	\hat{n}_k
0	122,628	122,713
1	21,686	21,656
2	4,014	4,116
3	832	801
4	224	158
5	68	31
6	17	6
7	7	1
8	7	0
≥ 9	0	0

Notes: The fit is a negative binomial distribution with parameters $\hat{\alpha} = 0.8665$ and $\hat{\tau} = 3.9097$.

any claims during $(0, t)$ computed on the basis of equation (8). Column $P_{t+1}^{(q)}$ gives the total corresponding premium ($P_{t+1}^{(q)} = BP_{t+1} \times BMF_{t+1}$). For power category “ ≥ 119 ,” her expected claim frequency for the first five periods is 0.3306 and 0.2808 after. The second half of Table 9 shows the evolution of the premium amounts for this policyholder.

Table 10 is similar to Table 9 except an exponential loss function is used. The bonus-malus factors are computed from equation (10) with $c = 12.93$. This parameter has been set in such a way that the variance of the a posteriori premiums paid by a policyholder during the first 10 years represents 50% of the variance if the premiums were computed under a quadratic loss; for more details.[See Denuit and Dhaene (2001).] It is interesting to compare the bonus-malus factors in Tables 9 and 10. Notice that her bonus-malus factors are identical whatever the power of the car but the premiums differ substantially. When an exponential loss is used, the size of the maluses is reduced. Because the system is financially balanced, this implies that the size of the bonuses is also reduced.

Table 9
Bonus-Malus Coefficients and A Posteriori Premiums
Quadratic Loss Function for Policyholder Age 30

Car in Power Category "Power ≤ 53 "							
t	BP_{t+1}	0 Claim in $(0, t)$		1 Claim in $(0, t)$		2 Claims in $(0, t)$	
		$BMF_{t+1}^{(q)}$	$P_{t+1}^{(q)}$	$BMF_{t+1}^{(q)}$	$P_{t+1}^{(q)}$	$BMF_{t+1}^{(q)}$	$P_{t+1}^{(q)}$
1	0.1787	0.7963	0.1423	1.7154	0.3065	2.6344	0.4708
2	0.1787	0.6616	0.1182	1.4251	0.2547	2.1887	0.3911
3	0.1787	0.5658	0.1011	1.2189	0.2178	1.8719	0.3345
4	0.1787	0.4943	0.0883	1.0648	0.1903	1.6352	0.2922
5	0.1787	0.4388	0.0784	0.9453	0.1689	1.4517	0.2594
6	0.1518	0.3945	0.0599	0.8499	0.1290	1.3052	0.1981
7	0.1518	0.3584	0.0544	0.7720	0.1172	1.1856	0.1800
8	0.1518	0.3283	0.0498	0.7072	0.1073	1.0860	0.1649
9	0.1518	0.3028	0.0460	0.6524	0.0990	1.0019	0.1521
10	0.1518	0.2811	0.0427	0.6055	0.0919	0.9299	0.1412
Car in Power Category "Power ≥ 119 "							
1	0.3306	0.7963	0.2633	1.7154	0.5671	2.6344	0.8709
2	0.3306	0.6616	0.2187	1.4251	0.4711	2.1887	0.7236
3	0.3306	0.5658	0.1871	1.2189	0.4030	1.8719	0.6189
4	0.3306	0.4943	0.1634	1.0648	0.3520	1.6352	0.5406
5	0.3306	0.4388	0.1451	0.9453	0.3125	1.4517	0.4799
6	0.2808	0.3945	0.1108	0.8499	0.2386	1.3052	0.3665
7	0.2808	0.3584	0.1006	0.7720	0.2168	1.1856	0.3329
8	0.2808	0.3283	0.0922	0.7072	0.1986	1.0860	0.3050
9	0.2808	0.3028	0.0850	0.6524	0.1832	1.0019	0.2813
10	0.2808	0.2811	0.0789	0.6055	0.1700	0.9299	0.2611

Table 10
Bonus-Malus Coefficients and A Posteriori Premiums
Exponential Loss Function ($c = 12.93$) for Policyholder Age 30

Car in Power Category "Power ≤ 53 "							
t	$BP_{t+1}^{(e)}$	0 Claim in $(0, t)$		1 Claim in $(0, t)$		2 Claims in $(0, t)$	
		$BMF_{t+1}^{(e)}$	$P_{t+1}^{(e)}$	$BMF_{t+1}^{(e)}$	$P_{t+1}^{(e)}$	$BMF_{t+1}^{(e)}$	$P_{t+1}^{(e)}$
1	0.1787	0.9002	0.1609	1.3505	0.2413	1.8007	0.3218
2	0.1787	0.8207	0.1467	1.2253	0.2190	1.6299	0.2913
3	0.1787	0.7553	0.1350	1.1234	0.2007	1.4915	0.2665
4	0.1787	0.7003	0.1251	1.0384	0.1856	1.3765	0.2460
5	0.1787	0.6533	0.1167	0.9662	0.1727	1.2791	0.2286
6	0.1518	0.6125	0.0930	0.9039	0.1372	1.1953	0.1815
7	0.1518	0.5768	0.0876	0.8496	0.1290	1.1224	0.1704
8	0.1518	0.5452	0.0828	0.8017	0.1217	1.0583	0.1606
9	0.1518	0.5170	0.0785	0.7591	0.1152	1.0013	0.1520
10	0.1518	0.4916	0.0746	0.7210	0.1095	0.9504	0.1443
Car in Power Category "Power ≥ 119 "							
1	0.3306	0.9002	0.2976	1.3505	0.4465	1.8007	0.5953
2	0.3306	0.8207	0.2713	1.2253	0.4051	1.6299	0.5388
3	0.3306	0.7553	0.2497	1.1234	0.3714	1.4915	0.4931
4	0.3306	0.7003	0.2315	1.0384	0.3433	1.3765	0.4551
5	0.3306	0.6533	0.2160	0.9662	0.3194	1.2791	0.4229
6	0.2808	0.6125	0.1720	0.9039	0.2538	1.1953	0.3356
7	0.2808	0.5768	0.1620	0.8496	0.2386	1.1224	0.3152
8	0.2808	0.5452	0.1531	0.8017	0.2251	1.0583	0.2972
9	0.2808	0.5170	0.1452	0.7591	0.2132	1.0013	0.2812
10	0.2808	0.4916	0.1381	0.7210	0.2025	0.9504	0.2669

3 Integrated Ratemaking

3.1 Claim Frequency Model

In seminal papers, Dionne and Vanasse (1989, 1992) proposed a bonus-malus system that integrates a priori and a posteriori information on an individual basis. Their system introduces a regression component in the Poisson counting model in order to use all available information in the estimation of accident frequency.

Let us assume that the number of claims K_{it} for the i^{th} policyholder of the portfolio during the year t conforms to a Poisson distribution with mean $\lambda_{I_i(t)}$, where $I_i(t)$ is the index of the risk class occupied by policyholder i in year t . A common problem for count data is that the fits obtained are poor even after allowing for important explanatory variables using the Poisson regression model. This indicates that, conditional upon the explanatory variables included in the final model, the variance of an observation is greater than its mean, implying that the Poisson assumption is incorrect. Most often, this is due to the fact that important explanatory variables may not have been measured and are consequently incorrectly excluded from the regression relationship.

A convenient way to avoid this problem is to introduce a random effect in this model; see, e.g., Pinquet (1999). We assume that K_{it} follows a Poisson distribution with mean $\lambda_{I_i(t)}\Theta_i$, where Θ_i has a gamma distribution but with unit mean, i.e., with parameters (α, α) . Then, K_{it} follows a negative binomial law, i.e.,

$$\Pr[K_{it} = k | I_i(t)] = \frac{\alpha + k - 1}{k} \left(\frac{\lambda_{I_i(t)}}{\alpha + \lambda_{I_i(t)}} \right)^k \left(\frac{\alpha}{\alpha + \lambda_{I_i(t)}} \right)^\alpha.$$

We can view Θ_i as representing the impact on the mean claim frequency of all the policyholders' characteristics not taken into account a priori. Let us now derive the a posteriori distribution of Θ_i .

Lemma 1. *If the cdf of Θ_i is $\Gamma(\cdot | \alpha, \alpha)$ then the cdf of $[\Theta_i | K_{ij} = k_{ij}, j = 1, 2, \dots, t]$ is $\Gamma(\cdot | \alpha + k_{i\bullet}(t), \alpha + \lambda_{i\bullet}(t))$ where*

$$\lambda_{i\bullet}(t) = \sum_{j=1}^t \lambda_{I_i(j)}.$$

Proof: Bayes Theorem yields

$$\begin{aligned} d\Pr[\Theta_i \leq \theta | K_{ij} = k_{ij}, j = 1, 2, \dots, t] \\ &= \frac{\Pr[K_{ij} = k_{ij}, j = 1, 2, \dots, t | \Theta_i = \theta] d\Pr[\Theta_i \leq \theta]}{\Pr[K_{ij} = k_{ij}, j = 1, 2, \dots, t]} \\ &= \frac{\theta^{k_{i\bullet}(t)} \exp(-\theta \lambda_{i\bullet}(t)) \alpha^\alpha \theta^{\alpha-1} \exp(-\alpha \theta) d\theta}{\alpha^\alpha \int_{\xi \in \mathbb{R}^+} \xi^{k_{i\bullet}(t)+\alpha-1} \exp(-\alpha \lambda_{i\bullet}(t) \xi) d\xi}, \end{aligned}$$

and the result follows.

In order to estimate the parameter α describing the residual heterogeneity of the portfolio, we use the maximum likelihood method. We maximize

$$L(\alpha) = \prod_{i=1}^{12} \prod_{k=0}^{\infty} \left\{ \frac{\alpha + k - 1}{k} \left(\frac{\lambda_i}{\alpha + \lambda_i} \right)^k \left(\frac{\alpha}{\alpha + \lambda_i} \right)^\alpha \right\}^{n_{ik}},$$

which yields $\hat{\alpha} = 0.8157$.

3.2 A Posteriori Premium Using a Quadratic Loss Function

In the model described in the preceding section, Dionne and Vanasse (1989, 1992) and Gisler (1996) have obtained the following result; it can be seen as a direct consequence of Proposition 4 and its proof is thus omitted.

Proposition 2. *Assuming the cdf of Θ_i is $\Gamma(\alpha, \alpha)$, then under a quadratic loss function, the a posteriori premium for policyholder i is given by*

$$P_{t+1}^{(q)} = \lambda_{I_i(t+1)} BMF^{(q)}(k_{i\bullet}(t), \lambda_{i\bullet}(t)),$$

where the bonus-malus coefficient is given by

$$BMF^{(q)}(k_{i\bullet}(t), \lambda_{i\bullet}(t)) = \frac{\alpha + k_{i\bullet}(t)}{\alpha + \lambda_{i\bullet}(t)} = (1 - \rho_q) \times 1 + \rho_q \frac{k_{i\bullet}(t)}{\lambda_{i\bullet}(t)}$$

with

$$\rho_q = \frac{\lambda_{i\bullet}(t)}{\alpha + \lambda_{i\bullet}(t)}.$$

Note that the greater the variance of Θ_i (i.e., the smaller α) the greater ρ_q (i.e., the greater the weight given to the claim history of the policyholder). Moreover, ρ_q is clearly increasing in $\lambda_{i\bullet}$. If $\lambda_{i\bullet}$ is small, as is the case for policies with a high deductibles, then ρ_q is also small. The no-claim discount for such policies is thus also small and, as pointed out by Gisler (1996), the bonus-malus systems are of questionable utility.

3.3 A Posteriori Premium Using Exponential Loss

The use of a quadratic loss function leads to high maluses because of the symmetry of the loss function: overcharges and undercharges are equally penalized. Although theoretically correct, such a system is not accepted by policyholders. It is better to have a model with a parameter controlling the severity of the system. One approach is to incorporate a priori variables in the exponential loss function.

Proposition 3. *Assuming that the cdf of Θ_i is $\Gamma(\cdot|\alpha, \alpha)$, then under an exponential loss with parameter $c > 0$ the a posteriori premium for policyholder i is given by*

$$P_{t+1}^{(e)} = \lambda_{I_i(t+1)} BMF^{(e)}(k_{i\bullet}(t), \lambda_{i\bullet}(t)) \quad (14)$$

where the bonus-malus coefficient is given by

$$BMF^{(e)}(k_{i\bullet}(t), \lambda_{i\bullet}(t)) = (1 - \rho_e) \times 1 + \rho_e \times \frac{k_{i\bullet}(t)}{\lambda_{i\bullet}(t)} \quad (15)$$

with

$$\rho_e = \frac{\lambda_{i\bullet}(t)}{c} \ln \left(1 + \frac{c}{\alpha + \lambda_{i\bullet}(t)} \right). \quad (16)$$

Proof: From Lemma 1, we get

$$\mathbb{E} \left[e^{-c\Theta_i} | K_{ij} = k_{ij}, j = 1, 2, \dots, t \right] = \left(\frac{\alpha + \lambda_{i\bullet}(t)}{\alpha + \lambda_{i\bullet}(t) + c} \right)^{\alpha + k_{i\bullet}(t)}$$

It follows that

$$\begin{aligned}
& \ln \mathbb{E} \left[e^{-c\Theta_i} | K_{ij} = k_{ij}, j = 1, 2, \dots, t \right] \\
& \quad = -(\alpha + k_{i\bullet}(t)) \ln \left(1 + \frac{c}{\alpha + \lambda_{i\bullet}(t)} \right) \\
\mathbb{E} \left[\ln \mathbb{E} \left[e^{-c\Theta_i} | K_{ij}, j = 1, 2, \dots, t \right] \right] \\
& \quad = -(\alpha + \lambda_{i\bullet}(t)) \ln \left(1 + \frac{c}{\alpha + \lambda_{i\bullet}(t)} \right).
\end{aligned}$$

The result then follows from Proposition 4.

Comparing the bonus-malus coefficients obtained with a quadratic and exponential loss functions we have, for any $c \geq 0$,

$$\ln \left(1 + \frac{c}{\alpha + \lambda_{i\bullet}} \right) \leq \frac{c}{\alpha + \lambda_{i\bullet}},$$

so that $\rho_e(c) \leq \rho_q$; the weight given to past claims is thus smaller under an exponential loss.

It can be shown that $\rho_e(c) \rightarrow 0$ as $c \rightarrow +\infty$. If the asymmetry factor c tends to $+\infty$ then all the risks within the same tariff class pay the same premium: there is no more experience rating. Conversely, $\rho_e(c) \rightarrow \rho_q$ as $c \rightarrow 0$. The results obtained by Dionne and Vanasse (1989, 1992) also appear as limit cases of those obtained with an exponential loss function.

3.4 Numerical Illustration

Computing the premiums for a 30-year old female policyholder using Dionne-Vanasse's methodology yields the results in Table 11. Unlike Table 9, the bonus-malus factors are not the same for both categories of car. The differences are explained by the presence of personal characteristics in the calculation of the factors in Table 11. Once the a priori variables are introduced the sizes of the bonuses and the maluses are reduced. Technically, this means that part of the heterogeneity has been taken into account in the a priori differentiation of the premiums, so that the residual heterogeneity is smaller and the magnitude of the a posteriori corrections is reduced.

It is interesting to note that even if a policyholder whose car is in category "Power ≤ 53 " always pays a smaller premium than the corresponding premium for the driver in category "Power ≥ 119 ," her bonus-malus factors are always greater (i.e., she has less bonuses and more

maluses). This is because good risks are rewarded in their base premiums (through the a priori variables incorporated in the tariff). Consequently, the size of bonus they require for equity is reduced. In other words, the premium discount awarded to risks judged as good a priori has to be smaller than the bonus awarded to those judged as bad a priori. Conversely, the penalties assessed to risks judged as good are larger than the penalties assessed to those judged as bad.

The same remarks hold for the bonus-malus coefficients obtained with an exponential loss function presented in Table 12. The severity of the a posteriori corrections is weaker than with a quadratic loss function, as expected.

4 Summary and Conclusions

As was pointed out earlier, the aim of this paper is to examine the interaction between a priori ratemaking (i.e., identification of the best predictors X and of the risk premium $\mathbb{E}[Y|X]$) and a posteriori ratemaking (i.e., premium corrections according to the claims history up to time t). To this end, we propose an extension of the exponential bonus-malus systems introduced in Denuit and Dhaene (2001) in the presence of a priori risk classification. The main advantage of this extension is that it provides the actuary with a parameter for controlling the severity of the a posteriori corrections. The actuary is allowed to vary this parameter from one extreme where there is no a posteriori correction to the other extreme where the severity corresponds to the classical quadratic loss function. At the limit, previous results based on a quadratic loss function are thus obtained. The a posteriori corrections also depend on the a priori amount of premium, yielding an integrated ratemaking mechanism recognizing the continuous nature of risk evaluation.

To illustrate our methodology, an example is provided using data from a Spanish insurance portfolio. We show that good risks are rewarded in their base premiums and, consequently, they require a smaller bonus than the bonus awarded to those judged as bad a priori, as expected.

In the future, we purpose to study bonus-malus scales accounting for a priori risk classification in the spirit of Taylor (1997), substituting the exponential loss function for its classical quadratic counterpart.

Table 11
Bonus-Malus Coefficients and A Posteriori Premiums
Quadratic Loss Function for Policyholder Age 30

Car in Power Category "Power ≤ 53 "							
t	BP_{t+1}	0 Claim in $(0, t)$		1 Claim in $(0, t)$		2 Claims in $(0, t)$	
		$BMF_{t+1}^{(q)}$	$P_{t+1}^{(q)}$	$BMF_{t+1}^{(q)}$	$P_{t+1}^{(q)}$	$BMF_{t+1}^{(q)}$	$P_{t+1}^{(q)}$
1	0.1787	0.8203	0.1466	1.8259	0.3263	2.8316	0.5060
2	0.1787	0.6953	0.1243	1.5478	0.2766	2.4002	0.4289
3	0.1787	0.6034	0.1078	1.3432	0.2400	2.0829	0.3722
4	0.1787	0.5330	0.0952	1.1863	0.2120	1.8397	0.3288
5	0.1787	0.4772	0.0853	1.0623	0.1898	1.6474	0.2944
6	0.1518	0.4383	0.0665	0.9757	0.1481	1.5130	0.2297
7	0.1518	0.4053	0.0615	0.9021	0.1369	1.3989	0.2124
8	0.1518	0.3768	0.0572	0.8388	0.1273	1.3008	0.1975
9	0.1518	0.3521	0.0535	0.7838	0.1190	1.2155	0.1845
10	0.1518	0.3305	0.0502	0.7356	0.1117	1.1408	0.1732
Car in Power Category "Power ≥ 119 "							
1	0.3306	0.7945	0.2626	1.4162	0.4682	2.0379	0.6737
2	0.3306	0.6590	0.2179	1.1747	0.3884	1.6905	0.5589
3	0.3306	0.5630	0.1861	1.0036	0.3318	1.4442	0.4775
4	0.3306	0.4914	0.1625	0.8760	0.2896	1.2606	0.4168
5	0.3306	0.4360	0.1441	0.7772	0.2569	1.1184	0.3697
6	0.2808	0.3979	0.1117	0.7092	0.1992	1.0206	0.2866
7	0.2808	0.3659	0.1027	0.6522	0.1831	0.9386	0.2635
8	0.2808	0.3387	0.0951	0.6037	0.1695	0.8687	0.2439
9	0.2808	0.3152	0.0885	0.5619	0.1578	0.8085	0.2270
10	0.2808	0.2948	0.0828	0.5255	0.1476	0.7562	0.2123

Table 12
Bonus-Malus Coefficients and A Posteriori Premiums
Exponential Loss Function ($c = 12.93$) for Policyholder Age 30

Car in Power Category "Power ≤ 53 "							
t	$BP_{t+1}^{(e)}$	0 Claim in $(0, t)$		1 Claim in $(0, t)$		2 Claims in $(0, t)$	
		$BMF_{t+1}^{(e)}$	$P_{t+1}^{(e)}$	$BMF_{t+1}^{(e)}$	$P_{t+1}^{(e)}$	$BMF_{t+1}^{(e)}$	$P_{t+1}^{(e)}$
1	0.1787	0.9635	0.1722	1.1676	0.2087	1.3718	0.2451
2	0.1787	0.9313	0.1664	1.1236	0.2008	1.3159	0.2352
3	0.1787	0.9022	0.1612	1.0846	0.1938	1.2669	0.2264
4	0.1787	0.8758	0.1565	1.0495	0.1876	1.2232	0.2186
5	0.1787	0.8516	0.1522	1.0177	0.1819	1.1838	0.2115
6	0.1518	0.8324	0.1264	0.9927	0.1507	1.1531	0.1750
7	0.1518	0.8144	0.1236	0.9694	0.1472	1.1245	0.1707
8	0.1518	0.7974	0.1210	0.9476	0.1438	1.0978	0.1666
9	0.1518	0.7813	0.1186	0.9270	0.1407	1.0728	0.1628
10	0.1518	0.7660	0.1163	0.9076	0.1378	1.0492	0.1593
Car in Power Category "Power ≥ 119 "							
1	0.3306	0.9359	0.3094	1.1298	0.3735	1.3238	0.4377
2	0.3306	0.8835	0.2921	1.0597	0.3503	1.2359	0.4086
3	0.3306	0.8390	0.2774	1.0013	0.3310	1.1636	0.3847
4	0.3306	0.8003	0.2646	0.9513	0.3145	1.1023	0.3644
5	0.3306	0.7660	0.2532	0.9075	0.3000	1.0491	0.3468
6	0.2808	0.7396	0.2077	0.8743	0.2455	1.0089	0.2833
7	0.2808	0.7154	0.2009	0.8439	0.2370	0.9724	0.2731
8	0.2808	0.6931	0.1946	0.8161	0.2292	0.9391	0.2637
9	0.2808	0.6723	0.1888	0.7904	0.2219	0.9084	0.2551
10	0.2808	0.6530	0.1834	0.7665	0.2152	0.8800	0.2471

References

- Denuit, M., and Dhaene, J. "Bonus-Malus Scales Using Exponential Loss Functions." *Bulletin of the German Society of Actuaries* 25 (2001): 13-27.
- De Wit, G.W., and Van Eeghen, J. "Rate Making and Society's Sense of Fairness." *ASTIN Bulletin* 14 (1984): 151-163.
- Dionne, G., and Vanasse, C. "A Generalization of Actuarial Automobile Insurance Rating Models: The Negative Binomial Distribution with a Regression Component." *ASTIN Bulletin* 19 (1989): 199-212.
- Dionne, G., and Vanasse, C. "Automobile Insurance Ratemaking in the Presence of Asymmetrical Information." *Journal of Applied Econometrics* 7 (1992): 149-165.
- Dobson, A.J. *An Introduction to Generalized Linear Models*. London, England: Chapman & Hall/CRC, 1990.
- Fahrmeir, L., and Tutz, G. *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York, N.Y.: Springer Verlag, 1994.
- Ferreira, J. "Identifying Equitable Insurance Premiums for Risk Classes: an Alternative to the Classical Approach." At 23rd International Meeting, Institute of Management Sciences, Athens, Greece, 1977.
- Gisler, A. "Bonus-Malus and Tariff Segmentation." At 27th ASTIN Colloquium, Copenhagen, Denmark, September 1-5, 1996.
- Lemaire, J. "How to Define a Bonus-Malus System with an Exponential Utility Function." *ASTIN Bulletin* 10, (1979): 274-282.
- Lemaire, J. *Bonus-Malus Systems in Automobile Insurance*. Boston, Mass.: Kluwer, 1995.
- Mc Cullagh, P., and Nelder, J.A. *Generalized Linear Models*. New York, N.Y.: Chapman & Hall, 1989.
- Pinquet, J. "Allowance for Cost of Claims in Bonus-Malus Systems." *ASTIN Bulletin* 27 (1997): 33-57.
- Pinquet, J. "Experience Rating Through Heterogeneous Models." In *Handbook of Insurance* G. Dionne, Editor. Boston, Mass.: Kluwer, 1999.
- Renshaw, A.E. "Modelling the Claim Process in the Presence of Covariates." *ASTIN Bulletin* 24 (1994): 265-285.
- Taylor, G. "Setting a Bonus-Malus Scale in the Presence of Other Rating Factors." *ASTIN Bulletin* 27 (1997): 319-327.
- Young, V. "Credibility and Persistency." *ASTIN Bulletin* 26 (1996): 53-69.

Appendix: Credibility Models with Quadratic and Exponential Loss Functions

Let us consider a sequence of random variables $\{X_1, X_2, X_3, \dots\}$ and a risk parameter Θ where Θ is a random variable or possibly a sequence of random variables. We assume the sequence of random variables $\{X_1, X_2, X_3, \dots | \Theta\}$ are independent. The first two moments of the X_i s are assumed to be finite. Moreover, the conditional mean of the X_i s is given by

$$\begin{aligned}\mu_i(\Theta) &= \mathbb{E}[X_i | \Theta] \\ \mathbb{E}[\mu_i(\Theta)] &= \mu_i\end{aligned}$$

for $i = 1, 2, 3, \dots$

Proposition 4.

(i) The minimum of $\mathbb{E}[\mu_{n+1}(\Theta) - \Psi_n(X_1, X_2, \dots, X_n)]^2$ on all the measurable functions $\Psi_n : \mathbb{R}^n \rightarrow \mathbb{R}$ is obtained for

$$\Psi_n^*(X_1, X_2, \dots, X_n) = \mathbb{E}[\mu_{n+1}(\Theta) | X_1, X_2, \dots, X_n].$$

(ii) The minimum of $\mathbb{E}[\exp[-c(\mu_{n+1}(\Theta) - \Psi_n(X_1, X_2, \dots, X_n))]]$ on all the measurable functions $\Psi_n : \mathbb{R}^n \rightarrow \mathbb{R}$ satisfying the constraint $\mathbb{E}[\Psi_n(X_1, X_2, \dots, X_n)] = \mu_{n+1}$ is obtained for

$$\begin{aligned}\Psi_n^*(X_1, \dots, X_n) &= \mu_{n+1} \\ &+ \frac{1}{c} \mathbb{E}[\ln \mathbb{E}[\exp(-c\mu_{n+1}(\Theta)) | X_1, \dots, X_n]] \\ &- \frac{1}{c} \ln \mathbb{E}[\exp(-c\mu_{n+1}(\Theta)) | X_1, \dots, X_n].\end{aligned}$$

(This constraint is made in order to guarantee financial equilibrium.)

Proof: (i) This is a classic result, and its proof can be found in many statistical textbooks. An easy way to see it consists in noting that

$$\begin{aligned} & \mathbb{E}\left[\left(\mu_{n+1}(\Theta) - \Psi_n(X_1, \dots, X_n)\right)^2\right] \\ &= \mathbb{E}\left[\left(\mu_{n+1}(\Theta) - \Psi_n^*(X_1, \dots, X_n) \right. \right. \\ &\quad \left. \left. + \Psi_n^*(X_1, \dots, X_n) - \Psi_n(X_1, \dots, X_n)\right)^2\right] \\ &= \mathbb{E}\left[\left(\mu_{n+1}(\Theta) - \Psi_n^*(X_1, \dots, X_n)\right)^2\right] \\ &\quad + \mathbb{E}\left[\left(\Psi_n^*(X_1, \dots, X_n) - \Psi_n(X_1, \dots, X_n)\right)^2\right], \end{aligned}$$

which is clearly minimal for $\Psi_n \equiv \Psi_n^*$.

(ii) Starting from

$$\begin{aligned} & \mathbb{E}\left[\exp\left[-c\left(\mu_{n+1}(\Theta) - \Psi_n(X_1, \dots, X_n)\right)\right]\right] \\ &= \mathbb{E}\left[\left[\exp\left[c\Psi_n(X_1, \dots, X_n)\right]\mathbb{E}\left[\exp\left[-c\mu_{n+1}(\Theta)\right]|X_1, \dots, X_n\right]\right]\right] \\ &= \mathbb{E}\left[\exp\left[c\left(\Psi_n(X_1, \dots, X_n) - \Psi_n^*(X_1, \dots, X_n)\right)\right]\right. \\ &\quad \left.\exp\left[c\mu_{n+1}\right]\exp\left[\mathbb{E}\ln\mathbb{E}\left[\exp\left[-c\mu_{n+1}(\Theta)\right]|X_1, \dots, X_n\right]\right]\right]. \end{aligned}$$

Now, let us apply Jensen's inequality to get

$$\begin{aligned} & \mathbb{E}\exp\left[-c\left(\mu_{n+1}(\Theta) - \Psi_n(X_1, \dots, X_n)\right)\right] \\ &\geq \exp\left[c\mathbb{E}\left[\Psi_n(X_1, \dots, X_n) - \Psi_n^*(X_1, \dots, X_n)\right]\right] \\ &\quad \exp\left[c\mu_{n+1}\right]\exp\left[\mathbb{E}\ln\mathbb{E}\left[\exp\left[-c\mu_{n+1}(\Theta)\right]|X_1, \dots, X_n\right]\right]. \end{aligned}$$

Because of the constraint on the expectation of the Ψ_n s, the first exponential is 1, thus completing the proof. \square

Fitting Loss Distributions in the Presence of Rating Variables

Farrokh Guiahi*

Abstract[†]

This paper focuses on issues and methodologies for fitting alternative statistical models—parametric probability distributions—to samples of insurance loss data. The interactions of loss distributions, deductibles, policy limits, and rating variables in the context of fitting distributions to losses are discussed. Fitted loss distributions serve an important function in pricing insurance products. The methodology developed in this paper is applied to a sample of insurance loss data that has the lognormal as the underlying loss distribution.

Key words and phrases: *generalized linear models, curve fitting, right-censored and left-truncated data, rating variables, maximum likelihood estimation, iteratively re-weighted least squares, parametric distribution*

1 Introduction

The price of an insurance product, i.e., the gross premium charged, consists of the pure premium, expenses, and a profit margin. The determination of pure premium is dependent on the knowledge of frequency and severity distributions of the potential claims. For pricing

*Farrokh Guiahi, Ph.D., F.C.A.S., A.S.A., is an associate professor of business computer information systems and quantitative methods at the Zarb School of Business, Hofstra University. His research interests include developing statistical models for pricing insurance products and applications of financial economics to insurance.

Dr. Guiahi's address is: Department of Business Computer Information Systems and Quantitative Methods, Zarb School of Business, Hofstra University, Hempstead, NY 11549, USA. Internet address: acsfzg@hofstra.edu

[†]The author is grateful to the anonymous referees and the editor for their many helpful comments and suggestions that have improved the quality and the scope of this paper.

some insurance products, only the mean of the frequency and severity distributions are sufficient. To price excess covers, however, the entire severity distribution must be known because one is interested in means of the form $\mathbb{E}[\max(0, X - M)]$ where X is the loss and $M > 0$ is a suitable retention.

To determine a severity distribution, the actuary fits a pre-selected parametric distribution to historical losses. So fitting distributions to losses is an integral component of pricing many insurance products. Hogg and Klugman (1984) provide a good introduction to the subject of fitting distributions to losses.

This paper supplements Hogg and Klugman (1984) by focusing on certain related topics. First, more emphasis is placed on the procedures for fitting loss distributions to individual loss data rather than grouped data. Second, methodologies required to incorporate rating variables in the process of fitting distributions to losses are presented. Finally, readers may find the computer program (codes), given to compute maximum likelihood estimates of parameters of the model used to be of some value.

The paper is organized as follows: Section 2 describes the types (complete or incomplete) of insurance data available and specifies the proper form for the likelihood function for each type. A procedure to incorporate rating factors into a curve-fitting process and assessing the effect of rating factors on loss distributions are discussed in Section 3. Two methods to compute maximum likelihood estimate of model parameters, and the notion of generalized residuals are given in Section 4. Section 5 illustrates how the methodology presented in this paper can be applied by using a sample of commercial fire loss data (Table A1 of the appendix). Some concluding statements are made in Section 6.

2 Complete and Incomplete Data

Some preliminaries regarding losses, deductibles, policy limits, and rating variables as inputs for fitting distributions to losses are presented. Then the proper form of the likelihood function is defined.

2.1 The Nature of Insurance Data

Insurance data considered here have the following characteristics:

- Losses are specified individually;

- For each individual loss, the information about its deductible and policy limit is furnished; and
- For each loss, we have auxiliary policy information regarding the rating variables.

These three items are further discussed below.

Losses are given on an **individual basis** and have not been grouped by loss size. The methodologies to fit distributions to data differ, depending on whether losses are grouped or individually specified. Losses may be closed or open. The amount recorded for each loss is the incurred value at the latest available evaluation period. If some losses in the sample data are still open as of the latest evaluation period, then those losses should be properly adjusted for further development. For more on loss development and reserving, see, for example, Brown and Gottlieb (2001, Chapter 4) or Wisner (1990, Chapter 4). Unfortunately, most of the methodologies for developing losses to their ultimate values are only available for grouped data. Further research is needed in the area of developing individual losses to their individual ultimate values. These individual losses should be suitably trended to reflect values expected in the future.

Deductibles are used to exclude certain losses. Usually deductibles are relatively small—for example, a few hundred or a few thousand dollars. For a large insured, however, deductibles may be sizable due to the existence of self-insured retention or other underlying coverages. Only dollar deductibles are considered here. Time deductibles such as waiting periods are not treated. A reported loss with a value in excess of its deductible is defined as *left-truncated*. If a loss arises from a policy with no underlying deductible, then for the purpose of the computation, a value of zero is imputed as the deductible amount. It is not required that the deductible amounts be the same for each loss.

Policy limits serve to restrict the amount of payment on a given loss or a loss occurrence. When the loss amount is at least as large as its policy limit, the loss is said to have been *right-censored*. If a loss arises from a policy where there is no underlying policy limit, then any amount greater than the loss amount may be imputed as the policy limit. In these instances, those losses have not been censored. Varying policy limits are allowed. No grouping of losses based upon deductible or policy limit amounts is required.

Samples of insurance loss data are said to be *incomplete*. This is due to inclusion of left-truncated (losses in excess of deductibles) and right-censored (some losses capped by their respective policy limits) data in the sample. Due to this incompleteness of data, it becomes

more difficult to estimate the parameters of a loss distribution and to assess the goodness of fit. Many traditional approaches for estimation of parameters of a loss distribution or assessing the goodness of fit of a distribution are valid only if the sample of observations is complete; that is, when there are neither left-truncated nor right-censored observations in the sample.

Rating variables in insurance depends upon the line of business, the degree of competition present in the market, and regulation. The effect of these rating variables upon loss distributions has important implications for underwriting selection. It also provides for a more differentiated rating system. How to incorporate the information provided by rating variables into the process of fitting distributions to losses is discussed below.

2.2 Likelihood Function

The standard approach to analyzing losses is to assume that losses are a realization of a probabilistic process governed by a parametric statistical distribution. Once the parametric distribution is selected to present the distribution of losses, the task of fitting a distribution to the loss data becomes one of estimating parameters of the selected distribution.

Some commonly used statistical methods to estimate parameters of a distribution are the method of moments, the least squares estimation, and the maximum likelihood estimation. This paper focuses on the maximum likelihood approach because, under certain conditions, maximum likelihood parameter estimates have many desirable properties including: uniqueness, consistency, asymptotic unbiasedness, asymptotic normality, and asymptotic efficiency; see, for example, Bain and Engelhardt (1992, Chapter 9.4, page 316).

The fact that most insurance data are incomplete (i.e., the data include left-truncated or right-censored observations) the method of the maximum likelihood estimation must be carefully applied. The likelihood function must be properly specified to reflect the presence of left-truncated or right-censored observations.

The following are necessary notations needed to write an expression for the likelihood function when the data are incomplete. Let

n = Number of losses;

y_i = Size of i^{th} loss (incurred value);

D_i = Deductible for the i^{th} loss;

PL_i = Policy limit for the i^{th} loss;

$f(y_i; \theta, \varphi)$ = Probability distribution function (pdf) of the loss amount random variable for complete data; and

$F(y_i; \theta, \varphi)$ = Cumulative distribution function (cdf) of the loss amount random variable for complete data; where

θ is the parameter of interest to the investigator and φ is the vector of incidental parameters. Note that the incidental parameters in φ are often referred to as nuisance parameters by statisticians.

The functional form of the likelihood function for a given loss depends upon whether (i) there is an applicable deductible, and (ii) whether the loss is capped by the policy limit. Hence, the contribution of a loss to the likelihood function may be one of the four mutually exclusive and exhaustive cases, written as L_{i1} , L_{i2} , L_{i3} , and L_{i4} as defined below. In addition, four indicator variables are introduced— δ_{i1} , δ_{i2} , δ_{i3} , and δ_{i4} —in order to write a succinct expression for the likelihood function of the sample. These four cases are considered next.

Case 1: No deductible and loss below policy limit (neither left-truncated nor right-censored), the complete data case:

$$L_{i1} = f(y_i; \theta, \varphi)$$

$$\delta_{i1} = \begin{cases} 1, & \text{if } D_i = 0 \text{ and } y_i < PL_i \\ 0, & \text{otherwise.} \end{cases}$$

Case 2: A deductible and loss below policy limit (left-truncated) data:

$$L_{i2} = \frac{f(D_i + y_i; \theta, \varphi)}{1 - F(D_i; \theta, \varphi)}$$

$$\delta_{i2} = \begin{cases} 1, & \text{if } D_i > 0 \text{ and } y_i < PL_i \\ 0, & \text{otherwise.} \end{cases}$$

Case 3: No deductible and loss capped by policy limit (right-censored) data:

$$L_{i3} = 1 - F(PL_i; \theta, \varphi)$$

$$\delta_{i3} = \begin{cases} 1, & \text{if } D_i = 0 \text{ and } y_i \geq PL_i \\ 0, & \text{otherwise.} \end{cases}$$

Case 4: A deductible and loss capped by policy limit (left-truncated and right-censored) data:

$$L_{i4} = \frac{1 - F(D_i + PL_i; \theta, \varphi)}{1 - F(D_i; \theta, \varphi)}$$

$$\delta_{i4} = \begin{cases} 1, & \text{if } D_i > 0 \text{ and } y_i \geq PL_i \\ 0, & \text{otherwise.} \end{cases}$$

In each of these four cases, the contributions of the i^{th} loss to the likelihood function (L_i) and the log-likelihood function (l_i) are

$$L_i = \prod_{j=1}^4 L_{ij}^{\delta_{ij}} \quad \text{and} \quad l_i = \ln L_i = \sum_{j=1}^4 \delta_{ij} l_{ij}$$

respectively, where $l_{ij} = \ln L_{ij}$. The likelihood and log-likelihood functions for the sample are given by:

$$L = \prod_i^n L_i \quad \text{and} \quad l = \sum_{i=1}^n l_i.$$

3 Using Rating Variables in Curve-Fitting

In statistical data analysis, one commonly assumes that sample data are a realization of random variables that are independent and identically distributed. The assumption of identically distributed random variables is usually not tenable with insurance data. Insurance risks are normally heterogeneous. Each risk has its own characteristics and its own propensity to produce a potential loss. Thus, we expect the loss distribution for fire for a small, unprotected frame building to be different from a large, highly protected, and fire-resistant building. It is desirable to have loss distributions that account for these differences. To a certain extent, underwriting rating factors reflect risk characteristics.

For this reason, risks with same values pertaining to their underwriting attributes are grouped together to form “homogeneous” classes for the purpose of rating.

The traditional approach for obtaining loss distributions dependent upon risk attributes is to segment losses into subgroups. Then, for each subgroup, a separate fitted loss distribution is obtained. For instance, in fire insurance, losses may be classified broadly by construction as frame, masonry, and fire-resistant. Three fitted loss distributions can be obtained according to the types of construction. When we utilize more than one rating factor (say, the three rating factors construction, protection, and occupancy) and allocate losses to cells formed by common values of rating factors, then the problem of fitting separate distributions to many cells becomes more complicated. This is due to the fact that if we use a separate parametric loss distribution for each cell, then the total number of parameters used may be too large in comparison to the number of observations. One principal advantage of using a statistical modeling approach to fitting distributions to losses, in presence of rating factors, is that fewer parameters in total are used and all of the losses are utilized to estimate model parameters simultaneously. This is the approach used in this paper.

Our approach to incorporate rating variables into the curve-fitting process is an extension of the generalized linear models (GLM) methodology. McCullagh and Nelder (1989) provide an excellent account of the theory and applications of GLM. The GLM approach, as originally developed, was intended only for complete data.

Loss distributions dependent upon rating variables have important implications for underwriting selection and determination of rates. By including the rating variables, one generally improves the fit to the data. A statistical modeling approach to curve fitting enables one to assess the effect of rating variables on loss distributions by performing statistical tests of hypotheses.

The GLM methodology consists of three components: the random component, the systematic component, and the link function.

- **The random component** pertains to the distribution of the random variable of interest, Y (e.g., loss or a transformation of the loss). We assume Y has a distribution belonging to the exponential family of distributions. The general form for the density of the exponential family is

$$f(y; \theta, \varphi) = \exp\left[\frac{(\theta y - b(\theta))}{a(\varphi)} + c(y, \varphi)\right]$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are real valued functions, θ is the primary parameter of interest, and φ is a vector of parameters often referred to as nuisance parameters. Some distributions belonging to the exponential family are normal, gamma, and inverse Gaussian.

- The **systematic component** of a GLM specifies the explanatory variables, x_1, \dots, x_p (e.g., rating variables). The explanatory variables may only influence the distribution of the Y through a single linear function called the *linear predictor*, η ,

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

- The **link function**, $g(\cdot)$, specifies how the mean of Y is related to the linear predictor, i.e.,

$$g(\mathbb{E}[Y]) = \eta = \sum_{j=0}^p \beta_j x_j.$$

The form of the link function varies by the type of distribution within the exponential family of distributions. Thus, the GLM method has a formal approach for relating the explanatory variables to a parameter of a distribution.

This paper uses a lognormal with parameters μ and σ^2 to represent the underlying loss distribution. There are several reasons for this selection. First, it is easy to interpret the parameters of a lognormal distribution. By taking the logarithm of the losses, the μ parameter represents the location parameter (mean), and the σ parameter is the scale (standard deviation). Second, lognormal distribution has been previously used to describe the distribution of fire losses (Benckert and Jung 1974). Third, by transforming lognormal into normal, a member of the exponential family of distributions, the methodology developed for GLM can be applied to the problem. For the normal distribution, the appropriate link function is the identity map, i.e., $g(y) \equiv y$, which leads to

$$\mathbb{E}[\ln(Y)] = \mu = \eta = \sum_j \beta_j x_j.$$

In GLM, each explanatory variable is considered either as a factor (categorical) or as a covariate (quantitative). For example, gender, construction type, and protection may be considered as categorical in nature, while age and the amount of insurance are considered to be quantitative. For the i^{th} loss, the linear predictor, η_i , can be written as

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \sum_{j=0}^p x_{ij} \beta_j = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad (1)$$

where $\boldsymbol{\beta}$ is a $(p + 1) \times 1$ vector of unknown parameters, and \mathbf{x}_i is a $(p + 1) \times 1$ vector of known constants, $x_{i0}, x_{i1}, \dots, x_{ip}$ with $x_{i0} = 1$. Hence the constant term β_0 is the intercept in the expression for the linear predictor. The other x_{ij} s components, $1 \leq j \leq p$, are used to represent rating variables.

The value of p is partially dependent upon the number of categorical rating factors included in the model, as well as their respective number of levels (values). In addition, p depends upon the number of quantitative rating variables in the model. When rating variables are not taken into consideration or when the information about them is not available, then $p = 0$.

Following are examples of the types of linear predictors, η_i , discussed throughout this paper. In fire insurance, some commonly used categorical rating factors are construction, protection, and occupancy. The amount of insurance (i.e., the value of the insured building) is taken as a measure of exposure and is quantitative. Here, for illustrative purposes only, the focus is mostly on construction and building value.

Assume there are three possible construction types (or levels): frame, masonry, and fire-resistant. In regression analysis, as well as in GLM, the contribution of a categorical variable to a linear predictor comes from specifying dummy variables. For the construction rating factor, two dummy variables C_{i1} and C_{i2} are introduced defined as follows:

$$C_{i1} = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ risk is a frame;} \\ 0, & \text{otherwise,} \end{cases}$$

$$C_{i2} = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ risk is a masonry;} \\ 0, & \text{otherwise.} \end{cases}$$

For the i^{th} loss, let BV_i denote the amount of insurance purchased by the policyholder to cover damages arising from peril of fire to the

building. For a fire policy, the policy limit for the building cover is consonant with the building value. Because there is a wide range of variability among building values, one can use the logarithm of the building value instead of building value as our covariate in the linear predictor. For these two variables (construction and building value), six linear predictors are defined yielding six statistical models:

$$\text{Model A: } \eta_i = \beta_0$$

$$\text{Model B: } \eta_i = \beta_0 + \beta_1 C_{i1} + \beta_2 C_{i2}$$

$$\text{Model C: } \eta_i = \beta_0 + \beta_1 \ln(BV_i)$$

$$\text{Model D: } \eta_i = \beta_0 + \beta_1 \ln(BV_i) + \beta_2 C_{i1} + \beta_3 C_{i2}$$

$$\text{Model E: } \eta_i = \beta_0 + \beta_1 \ln(D_i) + \beta_2 \ln(BV_i) + \beta_3 C_{i1} + \beta_4 C_{i2}$$

$$\text{Model F: } \eta_i = \beta_0 + \beta_1 \ln(BV_i) + \beta_2 C_{i1} + \beta_3 C_{i2} + \beta_4 C_{i1} \ln(BV_i) \\ + \beta_5 C_{i1} \ln(BV_i)^2$$

The linear predictor given by Model A is used when either one does not consider the information given by rating variables or when no information on rating variables is available. In these instances, one fits a distribution to the data that does not account for rating variables. Model A is our base model (distribution). The base distribution is used as a benchmark to gauge the relative improvement in fit by including rating variables.

Model B is appropriate if construction is the only rating factor used. Using the statistical methodology developed here, all data are used to estimate the values of the parameters β_0 , β_1 , β_2 simultaneously. This approach is different from the one in which the data are segmented into three groups according to types of construction.

Model C is used when one wishes to examine only the effect of exposure size (building value) on loss distribution. Model D accounts for both construction and building value. In this case, the vector

$$\mathbf{x}_i^T = (1, \ln(BV_i), C_{i1}, C_{i2})$$

represents the contribution of the i^{th} risk's attributes to the linear predictor, and p has the value of three.

Model E is an extension of Model D. Here, one wishes to determine whether, in the presence of construction and building value, the deductible affects the distribution of losses. Finally, Model F is another extension of Model D that includes interaction terms for construction and building value.

To each linear predictor, there corresponds a statistical model with parameters β , a vector of regression coefficients, and σ^2 . Procedures for estimating model parameters are discussed below.

A criterion used to compare alternative statistical models is Akaike's information criterion, AIC (Akaike 1973), which is defined as

$$\text{AIC} = -2 \times (\text{Maximized Log-Likelihood} - \text{No. Estimated Parameters}).$$

When two models are compared, the model with a smaller AIC value is the more desirable one. The AIC is based on log-likelihood, and it penalizes the log-likelihood by subtracting for the number of parameters estimated. Two other model selection criteria used in statistics are Schwarz's Bayesian information criterion (BIC) (Schwarz 1978) and deviance as used in generalized linear models (McCullagh and Nelder 1989). These three criteria are based on the value of maximized log-likelihood function.

The linear predictors given by Models A through D provide examples of nested models. For nested models, some models are a special case of a more general model. The linear predictors Models A, B, and C are special cases of the linear predictor Model D. For the linear predictor Model D, one can entertain the following statistical tests of hypotheses in order to assess the effect of rating variables:

$$H_0^{(1)} : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_0^{(2)} : \beta_2 = \beta_3 = 0$$

$$H_0^{(3)} : \beta_1 = 0.$$

The null hypothesis $H_0^{(1)}$ is used to test if either construction or building value (exposure size) has any effect on loss distribution. The failure to reject $H_0^{(1)}$, subject to the usual interpretation of errors probability type, suggests that the rating variables have no appreciable influence on the loss distribution. The rejection of hypothesis $H_0^{(1)}$ implies that the inclusion of building value or construction in the linear predictor gives a superior model as compared to the fit by the base distribution, Model A. The failure to reject the null hypothesis $H_0^{(2)}$ suggests that in the presence of building value, the addition of the construction factor does not improve the fit. The null hypothesis $H_0^{(3)}$ can be similarly interpreted.

For Models A and D, the null hypothesis $H_0^{(1)}$, can be tested using the likelihood ratio test. The asymptotic distribution of these test statistics

is χ^2 with degrees of freedom being equal to the number of β_i s set to zero, which, in this case is three. In Section 5, this test and similar tests are conducted, based on the data given in Table A1 of the appendix.

4 Maximum Likelihood Estimation of Parameters

4.1 Software Considerations

The estimation of parameters of the underlying distribution serves two purposes. First, the complete specification of a fitted distribution requires replacing the model parameters with their respective parameter estimates. Second, the effect of rating variables on the loss distribution can be assessed by using a likelihood ratio test statistic whose value is dependent upon maximum likelihood estimates of parameters.

Two methods are provided to estimate the maximum likelihood estimate (MLE) of parameters. The conventional approach to computation of the MLE of parameters is based on writing an expression for the likelihood or log-likelihood function. The partial derivatives of the likelihood or log-likelihood function with respect to parameters are computed and equated to zero. The solution of the system of nonlinear equations is achieved by using iterative numerical procedures.

An alternative method for computing the MLE of parameters is to use a solver, i.e., a black box approach. There are several software packages that are capable of computing MLEs including SAS[®], SYSTAT[®], and S-Plus[®]. In addition Microsoft Excel[®] has a solver that can be used as an optimizer to compute MLE of parameters. We have relied on a standard function, `ms`, available in the S-Plus[®] program, to compute the MLE of parameters. The codes for a program to determine the MLE of parameters of Model D, using lognormal as the underlying distribution, are given in Exhibit A of the appendix.

The S-Plus[®] solver program requires as input the specification of initial values for model parameters β and σ^2 . The program outputs consist of the MLE of parameters as well as the value of negative maximized log-likelihood function. Obtaining the MLE of parameters using a solver does not require the calculation of partial derivatives of the likelihood function. Thus, the MLEs of parameters are obtained with little computational effort on the part of the user. Another advantage of relying upon the S-Plus[®] solver, based on the author's experience, is that the algorithm used is not sensitive to the specification of initial values of the model parameters.

The solver (black box) approach for computing the MLE, however, has some limitations. First, to determine large sample (asymptotic) confidence intervals for the β_j s, it is necessary to compute Fisher's matrix (Tanner 1993). An estimate of Fisher's matrix involves the computation of first and second order partial derivatives of the likelihood function, which is not normally available if one uses a solver to compute the MLE. Second, the conventional approach for calculating the MLE can be further extended (as described in Section 4.2) to develop the notion of generalized residuals.

4.2 Direct Calculation of MLEs

An iterative procedure for calculating the MLE of parameters is now provided; it does not require the use of a solver.

The probability distribution functions (pdf), f , and the cumulative distribution functions (cdf), F , of the lognormal are

$$\begin{aligned} f(y_i; \mu_i, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{y_i} \exp \left\{ \frac{1}{2\sigma^2} [\ln(y_i) - \mu_i]^2 \right\} \\ &= \frac{1}{\sigma y_i} \phi \left(\frac{\ln(y_i) - \mu_i}{\sigma} \right) \\ F(y_i; \mu_i, \sigma^2) &= \Phi \left(\frac{\ln(y_i) - \mu_i}{\sigma} \right) \end{aligned}$$

where

$$\phi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \quad \text{and} \quad \Phi(y) = \int_{-\infty}^y \phi(t) dt$$

i.e., $\phi(y)$ and $\Phi(y)$ are the pdf and cdf, respectively, of the standard normal random variable. The parameter μ is the same as the θ parameter of the density function, as defined in Section 2. It is the mean of random variable $\ln(Y)$. The nuisance parameter σ corresponds to the φ parameter.

The statistical modeling approach used here relates the rating variables (explanatory variables) to the μ parameter of the lognormal. The link function is the identity map in the case of normal distribution, and η_i is defined in equation (1).

The basis log-likelihood functions are:

$$\begin{aligned}
 l_{i1} &= \text{constant} - \ln(\sigma) - \frac{(\ln(\gamma_i) - \mu_i)^2}{2\sigma^2} \\
 l_{i2} &= \text{constant} - \ln(\sigma) - \frac{(\ln(D_i + \gamma_i) - \mu_i)^2}{2\sigma^2} - \ln\left(1 - \Phi\left(\frac{\ln(D_i) - \mu_i}{\sigma}\right)\right) \\
 l_{i3} &= \ln\left(1 - \Phi\left(\frac{\ln(\text{PL}_i) - \mu_i}{\sigma}\right)\right) \\
 l_{i4} &= \ln\left(1 - \Phi\left(\frac{\ln(D_i + \text{PL}_i) - \mu_i}{\sigma}\right)\right) - \ln\left(1 - \Phi\left(\frac{\ln(D_i) - \mu_i}{\sigma}\right)\right).
 \end{aligned}$$

Using the fact that

$$\frac{\partial \mu_i}{\partial \beta_j} = x_{ij} \quad \text{and} \quad \frac{\partial l_{ir}}{\partial \beta_j} = x_{ij} \frac{\partial l_{ir}}{\partial \mu_i}$$

for $r = 1, 2, 3, 4$, it follows that

$$\frac{\partial l_{i1}}{\partial \beta_j} = \frac{s_i x_{ij}}{\sigma} \tag{2}$$

$$\frac{\partial l_{i1}}{\partial \sigma} = \frac{1}{\sigma} (s_i^2 - 1) \tag{3}$$

$$\frac{\partial l_{i2}}{\partial \beta_j} = \frac{x_{ij}}{\sigma} (t_i - h(u_i)) \tag{4}$$

$$\frac{\partial l_{i2}}{\partial \sigma} = \frac{1}{\sigma} (t_i^2 - 1 - u_i h(u_i)) \tag{5}$$

$$\frac{\partial l_{i3}}{\partial \beta_j} = \frac{x_{ij} h(v_i)}{\sigma} \tag{6}$$

$$\frac{\partial l_{i3}}{\partial \sigma} = \frac{v_i h(v_i)}{\sigma} \tag{7}$$

$$\frac{\partial l_{i4}}{\partial \beta_j} = \frac{x_{ij}}{\sigma} (h(w_i) - h(u_i)) \tag{8}$$

$$\frac{\partial l_{i4}}{\partial \sigma} = \frac{1}{\sigma} (w_i h(w_i) - u_i h(u_i)) \tag{9}$$

where

$$h(y) = \frac{\phi(y)}{1 - \Phi(y)} \tag{10}$$

$$s_i = \frac{\ln(y_i) - \mu_i}{\sigma} \tag{11}$$

$$t_i = \frac{\ln(D_i + y_i) - \mu_i}{\sigma} \tag{12}$$

$$u_i = \frac{\ln(D_i) - \mu_i}{\sigma} \tag{13}$$

$$v_i = \frac{\ln(PL_i) - \mu_i}{\sigma} \tag{14}$$

$$w_i = \frac{\ln(D_i + PL_i) - \mu_i}{\sigma} \tag{15}$$

The function $h(y)$ is referred to as *the hazard rate* as used in the analysis of survival data; see, for example Cox and Oakes (1984, Chapter 2, page 14). It follows that

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \sum_i^n \sum_{r=1}^4 \delta_{ir} \frac{\partial l_{ir}}{\partial \beta_j} \\ &= \sum_i^n \frac{x_{ij}}{\sigma} \left[\delta_{i1} s_i + \delta_{i2} (t_i - h(u_i)) + \delta_{i3} h(v_i) \right. \\ &\quad \left. + \delta_{i4} (h(w_i) - h(u_i)) \right] \end{aligned} \tag{16}$$

for $j = 0, 1, \dots, p$.

To solve for the maximum likelihood estimates, $\hat{\beta}_j$ s, one has to equate each of the $(p + 1)$ equations in equation (16) to zero. This involves the daunting task of solving simultaneously a system of nonlinear equations for the $\hat{\beta}_j$ s.

One approach to circumvent this difficulty is to define a new variable, z_i , and replace the problem of solving a system of nonlinear equations by that of regressing the z_i s on the x_{ij} s iteratively. As software for performing multiple regression is readily available in many computing environments, these iterations should be easy to perform. Thus equation (16) can be rewritten as

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\sigma} \sum_i \left(\frac{z_i - \mu_i}{\sigma} \right) x_{ij} \tag{17}$$

for $j = 0, 1, 2, \dots, p$. This suggests that we define

$$\begin{aligned} z_i &= \mu_i + \frac{\sigma^2}{x_{ij}} \sum_{r=1}^4 \delta_{ir} \frac{\partial l_{ir}}{\partial \beta_j} \\ &= \mu_i + \sigma [\delta_{i1}s_i + \delta_{i2}(t_i - h(u_i)) + \delta_{i3}h(v_i) + \delta_{i4}(h(w_i) - h(u_i))] \\ &= \delta_{i1} \ln(y_i) + \delta_{i2}[\ln(D_i + y_i) - \sigma h(u_i)] + \delta_{i3}[\mu_i + \sigma h(v_i)] \\ &\quad + \delta_{i4}[\mu_i + \sigma(h(w_i) - h(u_i))] \end{aligned} \quad (18)$$

for $i = 1, 2, \dots, n$.

Setting the $(p + 1)$ partial derivatives in equation (17) equal to zero yields

$$\sum_i (z_i - \hat{\mu}_i) x_{ij} = 0 \quad (19)$$

for $j = 0, 1, 2, \dots, p$. We can write equation (19) in the matrix form as

$$\mathbf{X}^T \mathbf{Z} - \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = 0, \quad (20)$$

which resembles the normal equations in ordinary regression analysis. This equation yields the solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Z} \quad (21)$$

provided the design matrix \mathbf{X} has full rank, i.e., provided $(\mathbf{X}^T \mathbf{X})^{-1}$ exists. If the vector \mathbf{Z} does not depend on the parameters $\boldsymbol{\beta}$ and σ^2 , then equation (20) [or, equivalently, equation (21)] gives the least squares estimate of $\boldsymbol{\beta}$. It is worth noting, when the data are complete, we have:

$$\delta_{i1} = 1, \quad \delta_{i2} = \delta_{i3} = \delta_{i4} = 0, \quad \text{and } z_i = \ln(y_i)$$

and \mathbf{Z} does not depend on model parameters. The least squares estimate of $\boldsymbol{\beta}$ can be obtained by regressing \mathbf{Z} only once on \mathbf{X} by solving the normal equation (20).

When the data are incomplete, however, \mathbf{Z} is dependent on $\boldsymbol{\beta}$ and σ^2 . The procedure to estimate $\boldsymbol{\beta}$, equation (20), is an application of the method known as iteratively re-weighted least squares (IRLS). The IRLS method has been applied to derive maximum likelihood estimates and robust regression coefficients; see Green (1984).

The essence of the IRLS procedure is as follows:

1. Select initial values for β and σ^2 .
2. Use these initial values of β and σ^2 to compute Z .
3. Regress Z on X to obtain an updated estimate of β .
4. Update the value of σ^2 and use the updated values of β and σ^2 to re-compute Z .
5. Repeat this procedure, i.e., updating β and σ^2 until there is no appreciable change in the values of updated parameters.

To proceed formally, a procedure is needed to re-compute or update the values of σ^2 . To update the initial estimate of σ , we note that

$$\begin{aligned} (z_i - \mu_i)^2 = & \delta_{i1}(\ln(y_i) - \mu_i)^2 \\ & + \delta_{i2}[(\ln(D_i + y_i) - \mu_i) - \sigma h(u_i)]^2 \\ & + \delta_{i3}[\sigma^2 h^2(v_i)] + \delta_{i4}[\sigma^2(h(w_i) - h(u_i))^2] \end{aligned} \quad (22)$$

which leads to

$$\frac{\partial l}{\partial \sigma} = \sum_i \frac{1}{\sigma^2} (z_i - \mu_i)^2 - DF \quad (23)$$

where

$$\begin{aligned} DF = & \sum_i \delta_{i1} + \delta_{i2}(1 + h(u_i)(u_i + h(u_i) - 2(\frac{\ln(D_i + y_i) - \mu_i}{\sigma}))) \\ & + \delta_{i3}[h(v_i)(h(v_i) - v_i)] \\ & + \delta_{i4}[(h(w_i) - h(u_i))^2 - w_i h(w_i) + u_i h(u_i)]. \end{aligned} \quad (24)$$

Setting equation (23) to zero gives

$$\hat{\sigma} = \sqrt{\frac{\sum_i (z_i - \hat{\mu}_i)^2}{DF}} \quad (25)$$

where $\hat{\mu}_i = \mathbf{x}_i^T \hat{\beta}$. This expression for $\hat{\sigma}$ (in equation (25)), as found in many multiple regression texts, is derived from an expression similar to equation (25). The differences are:

- Instead of z_i s, the observed values of the dependent variables are used; and

- In place of DF, as defined by equation (24), the degrees of freedom used is $(n - p - 1)$.

Let $\hat{\beta}^{(k)}$ and $\hat{\sigma}^{(k)}$ denote the estimates of β and σ obtained at the k^{th} iteration. The steps needed to compute the maximum likelihood estimates of model parameters based on the IRLS procedure are as follows:

Step 1: Initially, regress $\ln(D_i + y_i)$ on the rating variables x_{ij} s. Compute $\hat{\beta}^{(0)}$ as the initial estimate for β derived from the multiple regression coefficients. The square root of mean sums of squares (MSE) from the multiple regression output is used to determine $\hat{\sigma}^{(0)}$, the initial estimate for σ . We refer to these initial values, $\hat{\beta}^{(0)}$ and $\hat{\sigma}^{(0)}$, as *naïve* estimates of β and σ^2 . These estimates are ordinary least squares estimates, which do not account for the incompleteness of the data.

Step 2: Use $\hat{\beta}^{(0)}$ and $\hat{\sigma}^{(0)}$ to compute z_{is} .

Step 3: Use $\hat{\beta}^{(0)}$ and $\hat{\sigma}^{(0)}$ from Step 2 to compute z_{is} [equation (18)]. Regress z_{is} on x_{ij} s and compute a new estimate $\hat{\beta}^{(1)}$ based upon regression coefficients. Use equation (25) to calculate $\hat{\sigma}^{(1)}$, a new estimate for σ .

Step 4: If $\hat{\beta}^{(1)} = \hat{\beta}^{(0)}$ and $\hat{\sigma}^{(1)} = \hat{\sigma}^{(0)}$, then stop. Otherwise, replace $\hat{\beta}^{(0)}$ and $\hat{\sigma}^{(0)}$ by $\hat{\beta}^{(1)}$ and $\hat{\sigma}^{(1)}$ and return to Step 2.

In order for the iterations to stop, we use the following rule. Stop at the k^{th} iteration, if

$$\max_{1 \leq j \leq p} \left| \beta_j^{(k)} - \beta_j^{(k-1)} \right| \leq \epsilon, \quad \text{and} \quad \left| \sigma^{(k)} - \sigma^{(k-1)} \right| \leq \epsilon \quad (26)$$

for, say $\epsilon < 0.0005$. A numeric application of this procedure is given in Section 5.

A few remarks should be made about the convergence procedure. First, the above algorithm may not always converge for initial values of the parameters $\hat{\beta}^{(0)}$ and $\hat{\sigma}^{(0)}$ as determined from Step 1. Second, the maximum likelihood estimates $\hat{\beta}$ and $\hat{\sigma}$ may not be unique. These problems occur when the maximum likelihood estimates are derived from an iterative numerical method (Tanner 1993). To prove the uniqueness of the MLE requires further research and is not within the scope of the present paper.

4.3 Generalized Residuals Defined

Akaike's information criterion, as defined earlier, is a measure of overall fit of distribution and is used to compare alternative statistical models. In regression analysis, the residuals are used to check model assumptions and detect outliers. The usual definition of residuals as used in ordinary regression is not applicable here as we have a regression problem with incomplete data. The notion of generalized residuals is developed to deal with regressions using truncated and censored data.

In the case of complete data, Z does not depend upon the value of β and σ . In this case let e denote the vector of residuals. When the data are incomplete, let e^* denote vector of generalized residuals, i.e.,

$$e = Z - X\hat{\beta} \quad (27)$$

and

$$e^* = \tilde{Z} - X\hat{\beta} \quad (28)$$

where \tilde{Z} is defined by equation (18) with β and σ replaced by their respective maximum likelihood estimates. The generalized residuals concept is an extension of the notion of adjusted residuals as defined by Lawless (1982). Lawless defined adjusted residuals for regression models with right-censored observations in the case of lognormal. The notion of adjusted residuals, as defined by Lawless (1982), is extended to regression problems subject to left-truncated as well as right-censored observations. The generalized residual, as defined here, is an exploratory data analysis tool for an informal assessment of fit.

5 Numerical Illustrations

Examples are now provided on the following:

1. MLE of model parameters using a "solver";
2. An example to illustrate the use of iteratively re-weighted least squares (IRLS) method to estimate MLE of parameters;
3. Assessing the effect of rating variables on loss distributions; and
4. Plotting the generalized residuals as an explanatory data analysis tool.

The data used in numerical examples 1), 3), and 4) above are based on a sample of commercial fire losses as given in Table A1 of the appendix. The example 2) above, the application of IRLS procedure, uses the data in Table A2 of the appendix. For both sets of data, the lognormal is used as the underlying loss distribution

5.1 MLE of Model Parameters

For the data in Table A1 and the Model D, we have the naïve estimate of β and σ , i.e., the least squares estimate of β and σ as

$$\beta_0^{LS} = 4.568, \beta_1^{LS} = 0.238, \beta_2^{LS} = 1.068, \beta_3^{LS} = 0.040, \text{ and } \sigma^{LS} = 1.322.$$

The superscript LS is used to emphasis that these are least squares estimates of parameters. The least squares estimates do not account for some observations being subject to either truncation or censoring. The least square estimates are also the initial estimates of parameters for the S-Plus[®] program. The S-Plus[®] program is given in Exhibit A of the appendix.

The MLE for Model D parameters (see the S-Plus[®] program) are given as:

$$\hat{\beta}_0 = 1.715, \hat{\beta}_1 = 0.332, \hat{\beta}_2 = 2.155, \hat{\beta}_3 = 0.411, \text{ and } \hat{\sigma} = 1.899.$$

In addition, the negative of the maximized log-likelihood function has a value of 892.710. The above MLE of parameters were determined using a solver function `ms` of S-Plus[®].

Consider an insured risk (building) valued at \$1,000,000 and a construction type that is masonry. The average severity (ground up) value based on the MLE of parameters is \$4993. If one uses the naïve estimate of the parameters (least squares), the average severity for the same risk is \$6440. Thus, the average severity based on the naïve estimate is 29 percent larger than the true estimate (based on the MLE of parameters). Such a difference has practical implications for pricing insurance products.

An example is provided where the MLE of the model parameters is computed by the IRLS method. This procedure requires regressing the vector Z on the design matrix X a number of times, as outlined in Section 4. Although this procedure is theoretically sound, based on the author's experience the method is sensitive to the selection of initial value of parameters. For illustrative purposes only, this method

Table 1
MLE of Model D Parameters Using Table A2
And the Linear Predictor $\eta_i = \beta_0 + \beta_1 \ln(\text{BV}_i) + \beta_2 C_{i1} + \beta_3 C_{i2}$

Iteration	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}$
0	5.9645	0.1435	0.4677	-0.0602	1.7147
1	5.6912	0.1313	0.7006	-0.2138	1.8738
2	5.5884	0.1205	0.8420	-0.2591	1.8967
3	5.5659	0.1136	0.9160	-0.2762	1.9475
4	5.5573	0.1081	0.9668	-0.2778	1.9672
5	5.5570	0.1044	0.9984	-0.2762	1.9858
6	5.5570	0.1017	1.0201	-0.2731	1.9946
7	5.5579	0.1000	1.0341	-0.2705	2.0020
8	5.5583	0.0989	1.0435	-0.2684	2.0059
9	5.5587	0.0982	1.0497	-0.2669	2.0089
10	5.5589	0.0977	1.0538	-0.2657	2.0106
11	5.5590	0.0973	1.0565	-0.2650	2.0119
12	5.5591	0.0971	1.0582	-0.2645	2.0126
13	5.5591	0.0970	1.0594	-0.2641	2.0132
14	5.5591	0.0969	1.0602	-0.2639	2.0135
15	5.5591	0.0968	1.0607	-0.2637	2.0137
16	5.5591	0.0968	1.0610	-0.2636	2.0139

is applied to the data in Table A2 of the appendix based on Model D. Table 1 illustrates the intermediate value of estimates of β and σ^2 at different iterations before the convergence to values $\hat{\beta}$ and $\hat{\sigma}$ occurs.

5.2 Assessing the Effect of Rating Variables

Section 3 defined six linear predictors that corresponded to six statistical models. For the data in Table A1 of the appendix, based on lognormal, the estimates of linear predictors and the negatives of maximized log-likelihood functions for each model are presented in Table 2.

Nested models can be compared based upon the values of likelihood ratio statistics. The only difference between Model E and Model D is the inclusion of the deductible term in the linear predictor equation. Using Model E, we can consider the null hypothesis

Table 2
Summary of Model Fitting Information

Model A	$\hat{\mu}_i = 5.887,$ $\hat{\sigma} = 2.302, -\ln L = 897.8$
Model B	$\hat{\mu}_i = 5.537 + 1.938C_{i1} + 0.161C_{i2},$ $\hat{\sigma} = 2.142, -\ln L = 894.8$
Model C	$\hat{\mu}_i = 3.162 + 0.252 \ln(BV_i),$ $\hat{\sigma} = 2.126, -\ln L = 896.8$
Model D	$\hat{\mu}_i = 1.715 + 0.332 \ln(BV_i) + 2.155C_{i1} + 0.411C_{i2},$ $\hat{\sigma} = 1.899, -\ln L = 892.7$
Model E	$\hat{\mu}_i = 2.215 - 0.483 \ln(D_i) + 0.478 \ln(BV_i) + 2.604C_{i1}$ $+ 0.416C_{i2},$ $\hat{\sigma} = 2.088, -\ln L = 891.8$
Model F	$\hat{\mu}_i = 2.350 + 0.284 \ln(BV_i) + 0.758C_{i1} + 0.0707C_{i2}$ $+ 0.112 \ln(BV_i)C_{i1} + 0.0238 \ln(BV_i)C_{i2}$ $\hat{\sigma} = 1.899, -\ln L = 892.7$

$$H_0^{(4)} : \beta_1 = 0$$

To test $H_0^{(4)}$ the following likelihood ratio statistic is used: $-2(\ln L_D - \ln L_E)$, where L_D and L_E correspond to the values of maximized likelihood function for Models D and E, respectively. The asymptotic distribution of this test statistic is a χ^2 distribution with 1 degree of freedom. At the 5 percent significance level we cannot reject the null hypothesis $H_0^{(4)}$. Thus, we can drop the deductible term and use the simpler Model D instead of Model E.

Next, we compare Model D with Model F. The difference between the two models is the inclusion of interaction terms between exposure size and construction. We can test for the effect of interaction terms based on Model F by considering the following test of hypothesis:

$$H_0^{(5)} : \beta_4 = \beta_5 = 0$$

The likelihood ratio test statistic used is $-2(\ln L_D - \ln L_F)$ where L_D and L_F correspond to the values of maximized likelihood function for Models D and F, respectively. The asymptotic distribution of this test

statistic is χ^2 with 2 degrees of freedom. This observed value does not fall in the reject region when the significance level is 5 percent. Again, we can drop the interaction terms and use the simpler Model D.

Finally for nested Models A, B, C, and D, Table 3 provides useful statistics.

Table 3
Nested Hypotheses Based On Model D

Hypothesis	Likelihood Ratio Test Statistic	χ^2_d Distribution	
		d	95 th Percentile
H_0			
$\beta_1 = \beta_2 = \beta_3 = 0$	$-2(\ln L_A - \ln L_D) = 10.11$	3	7.81
$\beta_2 = \beta_3 = 0$	$-2(\ln L_C - \ln L_D) = 8.24$	2	5.99
$\beta_1 = 0$	$-2(\ln L_B - \ln L_D) = 4.25$	1	3.84

Notes: d denotes the number of degrees of freedom for the χ^2 . In addition L_A, L_B, L_C and L_D correspond to the values of maximized likelihood function for Models A, B, C, and D respectively.

The results in Table 3 should be interpreted carefully. First, the distribution of test statistics for performing tests of hypotheses is not exact. The large sample (asymptotic) distribution of the likelihood ratio statistic, i.e., χ^2 distribution, is used. Second, the sample is relatively small in size. With these qualifications in mind, let us interpret the results of Table 3.

First, the observed test statistics are larger than 95th percentiles of respective χ^2 distributions. The implications are that each null hypothesis should be rejected at a 5 percent significance level. Hence, rating variables are useful in the description of loss distributions. Second, Model D has relatively the largest value of likelihood function, representing the best fit among the four models.

5.3 Generalized Residuals: A Diagnostic Tool

By examining the various plots of generalized residuals against fitted values and explanatory variables, an informal assessment of fit is made. Also, the plot of these residuals is helpful in determining extreme observations. If the plots of generalized residuals exhibit a systematic pattern, then the implication is that some assumption about the regression model is violated. In these instances, one has to make appropriate corrections to the regression model. We can write equation (18) as

$$\begin{aligned}
 e_i^* &= \tilde{z}_i - \hat{\mu}_i \\
 &= \delta_{i1}(\ln(y_i) - \hat{\mu}_i) + \delta_{i2}[\ln(D_i + y_i) - (\hat{\mu}_i + \sigma h(\tilde{u}_i))] \\
 &\quad + \delta_{i3}[(\hat{\mu}_i + \hat{\sigma} h(v_i)) - \hat{\mu}_i] \\
 &\quad + \delta_{i4}[(\hat{\mu}_i + \hat{\sigma} h(w_i)) - (\hat{\mu}_i + \hat{\sigma} h(u_i))]
 \end{aligned} \tag{29}$$

\tilde{u}_i , \tilde{v}_i , and \tilde{w}_i are obtained from equations (13), (14), and (15) by replacing β and σ^2 by their MLE $\hat{\beta}$ and $\hat{\sigma}^2$.

For a normal random variable, X , with mean μ and variance σ^2 , the conditional mean of X is

$$\begin{aligned}
 \mathbb{E}[X|X > a] &= \int x f(x|X > a) dx \\
 &= \frac{1}{1 - F(a)} \int_a^\infty x f(x) dx \\
 &= \mu + \sigma h\left(\frac{a - \mu}{\sigma}\right).
 \end{aligned} \tag{30}$$

where a is a constant and h is the hazard function in equation (10).

We re-interpret equation (29), in light of the properties of normal distribution given by equation (30), in Table 4 below. The four cases defined in Table 4 correspond to cases as defined for the likelihood function in Section 2.

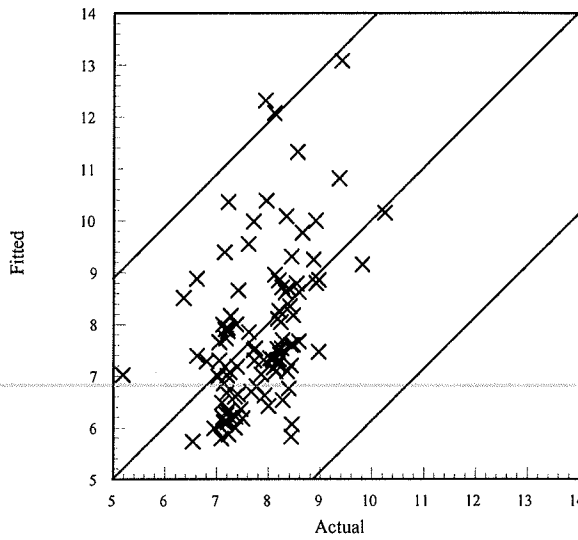
Table 4
Interpretation of Residuals for Various Cases

Case	Description	Actual	Fitted
1	$D_i = 0, y_i < PL_i$	$x_i = \ln(y_i)$	$\hat{\mu}_i = \mathbb{E}[X_i]$
2	$D_i > 0, y_i < PL_i$	$x_i = \ln(D_i + y_i)$	$\mathbb{E}[X_i X_i > \ln(D_i)]$
3	$D_i = 0, y_i \geq PL_i$	$\mathbb{E}[X_i X_i > \ln(PL_i)]$	$\hat{\mu}_i = \mathbb{E}[X_i]$
2	$D_i > 0, y_i \geq PL_i$	$\mathbb{E}[X_i X_i > \ln(D_i + PL_i)]$	$\mathbb{E}[X_i X_i > \ln(D_i)]$

Notes: $X_i \sim N(\hat{\mu}_i, \hat{\sigma}^2)$ where $\hat{\mu}_i$ and $\hat{\sigma}^2$ are maximum likelihood estimates of μ_i and σ^2 .

Figure 1 shows three straight lines superimposed on the scatter plot of fitted against actual values as defined in Table 4. The middle line is the 45-degree line. If a point is on the line, then its fitted and actual value will be the same. For the points off the 45-degree line, the vertical

Figure 1
A Scatter Plot of Actual Vs. Fitted in Model D



distance from any point to the 45-degree line represents the generalized residual value. The other two parallel lines in Figure 1 represent lines that are a distance of $\pm 2 \hat{\sigma}_{gres}$ above or below the 45-degree line, where $\hat{\sigma}_{gres}$ is sample standard deviation of generalized residuals. By analogy with ordinary regression theory, we would expect that the majority (95 percent) of scatter points to lie between the two lines. Finally, no systematic pattern is observed in Figure 1 when the actual values are plotted against the fitted values.

6 Summary and Conclusions

This paper addresses issues that are germane to fitting parametric loss distributions to insurance data. The presence of deductibles and policy limits renders the insurance data incomplete and complicates both fitting and assessing the fit of these distributions. Two procedures are stated for determining the parameters' MLEs.

A new methodology is introduced for incorporating rating factors into the curve fitting process. It is shown that the likelihood ratio test

statistic can be used to assess the effect of rating factors on loss distributions. The concept of generalized residuals is developed as a vehicle for informally assessing the quality of the fit. This new methodology is illustrated via a numerical example.

We conclude that improper estimation of the parameters of a loss distribution can result in substantial errors in pricing the underlying insurance product. Also, the inclusion of rating factors can provide a better fit to insurance loss data.

References

- Bain, L.J. and Engelhardt, M. *An Introduction to Probability and Mathematical Statistics*, Second Edition. Belmont, Calif.: Duxbury Press, 1992.
- Benckert, L.G. and Jung, J. "Statistical Models of Claim Distributions in Fire Insurance." *ASTIN Bulletin* 8 (1974): 1-25.
- Brown, R.L. and Gottlieb, L.R. *Introduction to Ratemaking and Loss Reserving for Property and Casualty Insurance*, Second Edition. Winsted, Conn.: Actex, 2001.
- Cox, D.R. and Oakes, D. *Analysis of Survival Data*. London, England: Chapman and Hall, 1984.
- Green, P.J. "Iteratively Re-weighted Least Squares for Maximum Likelihood Estimation and Robust and Resistant Alternatives (with discussion)." *Journal of Royal Statistical Society B* 46 (1984): 149-192.
- Hogg, R.V. and Klugman, S.A. *Loss Distributions*. New York, N.Y.: John Wiley & Sons, 1984.
- Lawless, J.F. *Statistical Models and Methods for Lifetime Data*. New York, N.Y.: John Wiley & Sons, 1982.
- McCullagh, P. and Nelder, J.A. *Generalized Models*, Second Edition. New York, N.Y.: Chapman and Hall, 1989.
- Tanner, M.A. *Tools for Statistical Inference*. New York: Springer-Verlag, 1993.
- Wiser, R.F. "Loss Reserving." In *Foundations of Casualty Actuarial Science*, Second Edition. Arlington, Va.: Casualty Actuarial Society, 1990.

Appendix

Exhibit A An S-Plus® Program to Compute MLEs for Model D

```
lognormal.model.D<- function(b0,b1,b2,b3,sigma,
                             data.matrix)
{ D <- data.matrix[,1]
  PL <- data.matrix[,2]
  y <- data.matrix[,3]
  cnst <- data.matrix[,4]
  z <- D+(y*(y<PL)+PL*(y>=PL))
  C1 <- cnst == 1
  C2 <- cnst == 2
  d <-D+(D == 0)*1
  mu <- b0+b1*log(PL)+b2*C1+b3*C2
  delta1 <- (D == 0)*(y < PL)
  delta2 <- (D > 0)*(y < PL)
  delta3 <- (D == 0)*(y >= PL)
  delta4 <- (D > 0)*(y >= PL)
  L1 <- dlnorm(z,mu,sigma)
  L2 <- dlnorm(z,mu,sigma)/(1-plnorm(d,mu,sigma))
  L3 <- 1-plnorm(z,mu,sigma)
  L4 <- (1-plnorm(z,mu,sigma))/
        (1-plnorm(d,mu,sigma))
  logL <-delta1*log(L1)+delta2*log(L2)
        +delta3*log(L3)+delta4*log(L4) -logL }
min.model.D<-ms(~lognormal.model.D(b0,b1,b2,b3,
                                   sigma,TableA),
start=list(b0=4.568,+b1=0.238, b2=1.068,b3=0.0403,
           sigma=1.322))
min.model.D
value: 892.7099
parameters:
      b0      b1      b2      b3      sigma
1.715296 0.3317345 2.154994 0.4105021 1.898501
formula: ~ lognormal.model.D(b0, b1, b2, b3,
                             sigma, TableA)
100 observations
call: ms(formula = ~ lognormal.model.D(b0, b1, b2, b3,
                                       sigma, TableA),
start = +list(b0 = 4.568, b1 = 0.238, b2 =1.068,
              b3 = 0.0403, sigma = 1.322))
```

Table A1
Insurance Company Data (in Dollars)

i	D_i	PL_i	y_i	Construction Code
1	1,000	57,000	502	2
2	250	41,000	31,971	1
3	1,000	1,000	367	1
4	250	60,000	698	2
5	100	10,000	4,863	2
6	250	24,000	834	2
7	250	16,000	646	1
8	250	60,000	198	2
9	1,000	66,000	275	2
10	250	36,000	500	1
11	100	53,000	1,518	2
12	250	70,000	2,430	2
13	250	51,000	357	1
14	250	79,000	2,008	2
15	500	139,000	3,044	1
16	250	155,000	238	2
17	250	150,000	3,244	2
18	250	98,000	850	2
19	250	100,000	198	2
20	100	110,000	110,000*	1
21	250	115,000	1,191	1
22	250	100,000	1,852	3
23	5,000	153,000	4,433	1
24	250	120,000	100	2
25	250	100,000	2,501	2
26	250	350,000	1,057	2
27	250	373,000	180	1
28	1,000	208,000	9,385	1
29	1,000	600,000	2,300	3
30	1,000	284,000	5,589	1
31	1,000	263,000	652	2
32	250	312,000	3,975	1
33	250	280,000	485	2
34	1,000	312,000	2,092	2
35	2,500	250,000	250,000*	1

Notes: * Denotes a censored observation.

Table A1 (Continued)
Insurance Company Data (in Dollars)

i	D_i	PL_i	y_i	Construction
				Code
36	250	300,000	250	2
37	500	625,000	1,305	3
38	1,000	319,000	6,729	3
39	500	9,214,000	185	2
40	250	43,000	75	2
41	1,000	1,000	865	3
42	100	33,000	206	2
43	250	7,000	2,303	1
44	250	64,000	11,760	2
45	250	45,000	402	2
46	500	30,000	3,352	1
47	250	2,000	511	1
48	0	10,000	1,115	2
49	250	52,000	237	2
50	250	3,000	1,197	2
51	100	50,000	7,107	2
52	250	89,000	535	2
53	1,000	200,000	5,959	2
54	250	100,000	1,224	3
55	250	85,000	85,000*	1
56	250	103,000	2,358	2
57	250	110,000	31,243	2
58	500	110,000	1,488	1
59	250	175,000	2,702	3
60	1,000	154,000	850	2
61	250	100,000	300	2
62	250	134,000	930	2
63	500	125,000	305	2
64	1,000	115,000	190	2
65	250	630,000	1,875	1
66	1,000	402,000	5,075	2
67	500	204,000	972	2
68	250	300,000	271	3
69	250	350,000	87	1
70	500	595,000	625	2

Notes: * Denotes a censored observation.

Table A1 (Continued)
Insurance Company Data (in Dollars)

i	D_i	PL_i	y_i	Construction Code
71	1,000	275,000	20,934	1
72	250	290,000	609	1
73	250	560,000	325	2
74	1,000	371,000	6,012	1
75	1,000	362,000	860	2
76	250	317,000	2,720	2
77	500	6,817,000	1,040	3
78	1,000	3,010,000	48,762	1
79	1,000	3,000,000	22,930	3
80	1,000	800,000	498	3
81	500	838,000	990	2
82	250	1,400,000	5,491	3
83	1,000	1,500,000	1,185	3
84	500	36,819,000	6,032	2
85	250	1,282,000	13,775	2
86	250	1,000,000	150	3
87	1,000	6,127,000	4,536	2
88	100	1,140,000	298	3
89	1,000	1,910,000	335	2
90	5,000	6,023,000	20,576	1
91	250	700,000	230	2
92	1,000	1,000,000	200	2
93	500	1,442,000	1,247	1
94	1,000	2,000,000	10,000	2
94	1,000	2,526,000	4,525	3
96	500	65,065,000	16,981	2
97	1,000	1,236,000	4,911	2
98	1,000	5,000,000	81,692	2
99	250	2,275,000	21,447	2
100	1,000	2,700,000	992	2

Notes: * Denotes a censored observation.

Table A2
Insurance Company Data (in Dollars)

i	BV_i	y_i	D_i	Construction Code
1	250,000	1,809	250	2
2	84,000	614	0	2
3	10,000	10,000*	0	1
4	4,798,000	676	1,000	2
5	125,000	346	1,000	2
6	100,000	95,542	250	2
7	350,000	801	1,000	2
8	14,000	14,000*	0	2
9	28,000	255	0	1
10	2,320,000	145	100	1
11	250,000	2,988	5,000	2
12	1,800,000	2,725	0	1
13	123,000	2,288	1,000	1
14	350,000	3,648	1,000	1
15	750,000	2,803	0	1
16	100,000	1,451	1,000	2
17	150,000	538	0	2
18	212,000	8,559	100	1
19	16,000	913	1,000	2
20	155,000	424	250	2
21	360,000	270	500	3
22	4,500,000	42,797	1,000	3
23	700,000	294	1,000	2
24	25,000	25,000*	1,000	1
25	162,000	5,115	5,000	2
26	650,000	2,249	250	1
27	10,000	3,600	100	2
28	7,222,000	12,338	1,000	2
29	150,000	156	250	2
30	347,000	28,380	25,000	2
31	200,000	1,703	250	2
32	10,885,000	1,636	1,000	2
33	1,848,000	1,658	0	3
34	950,000	166	500	3
35	598,000	126	0	2

Notes: * Denotes a censored observation.

Table A2 (Continued)
Insurance Company Data (in Dollars)

i	BV_i	y_i	D_i	Construction
				Code
36	72,000	328	100	1
37	47,000	41,128	100	3
38	185,000	350	0	2
39	600,000	2,295	1,000	1
40	3,500,000	3,529	250	3
41	125,000	107	250	2
42	1,320,000	378	100	2
43	135,000	3,197	250	3
44	30,000	572	0	1
45	240,000	4,067	500	3
46	50,000	79	1,000	1
47	270,000	6,413	1,000	3
48	250,000	610	0	1
49	67,000	67,000*	250	1
50	10,000	9,364	1,000	2
51	500,000	1,323	100	2
52	572,000	980	1,000	2
53	700,000	632,003	50	1
54	416,000	5,366	500	2
55	22,000	1,854	0	2
56	350,000	2,131	500	2
57	20,000	447	1,000	2
58	650,000	5,974	1,000	2
59	4,000,000	3,591	250	3
60	2,200,000	1,584	500	3
61	550,000	1,066	500	2
62	5,000	1,902	1,000	1
63	270,000	490	1,000	1
64	3,652,000	950	250	3
65	875,000	5,090	0	2
66	120,000	2,171	500	2
67	50,000	282	0	3
68	1,636,000	7,352	100	2
69	700,000	1,424	0	2
70	50,000	488	100	1

Notes: * Denotes a censored observation.

Table A2 (Continued)
Insurance Company Data (in Dollars)

i	BV_i	y_i	D_i	Construction Code
71	47,000	730	0	2
72	170,000	451	250	3
73	18,000	251	0	2
74	23,000	3,490	2,500	2
75	2,200,000	525	1,000	3
76	1,000	195	250	3
77	50,000	19,572	1,000	2
78	24,000	599	500	3
79	450,000 *	450,000	1,000	1
80	150,000	670	1,000	3
81	500,000	163,704	250	3
82	250,000	2,632	250	2
83	1,000	887	250	3
84	747,000	902	0	1
85	15,000	336	0	2
86	350,000	51	0	2
87	1,401,000	4,750	1,000	2
88	1,556,000	484	0	3
89	160,000	838	0	2
90	750,000	3,368	1,000	1
91	2,103,000	1,844	1,000	1
92	135,000	847	250	3
93	624,000	113,749	10,000	2
94	186,000	153	0	2
95	1,756,000	12,867	0	2
96	75,000	4,144	0	1
97	550,000	6,664	1,000	2
98	79,000	258	500	2
99	5,000	114	0	2
100	17,139,000	1,952	1,000	2

Notes: * Denotes a censored observation.

Linear Empirical Bayes Estimation of Survival Probabilities with Partial Data

Mostafa Mashayekhi*

Abstract[†]

In this paper we consider linear empirical Bayes estimation of survival probabilities with partial data from right-censored and possibly left-truncated observations. Such data are produced by studies in which the exact times of death are not recorded and the length of time that each subject may be under observation cannot exceed one unit of time. We obtain asymptotically optimal linear empirical Bayes estimators, with respect to the squared error loss function, under the assumption that the probability of death under observation in a unit time interval is proportional to the length of observation. This assumption is sometimes implied by Balducci's assumption and sometimes is implied by the assumption of uniform distribution of deaths.

Key words and phrases: *asymptotically optimal, credibility theory*

*Mostafa Mashayekhi, Ph.D., A.S.A., M.A.A.A., is an assistant professor of actuarial science at the University of Nebraska-Lincoln. He received his bachelor's degree in economics from the University of London's Queen Mary College and his Master's degree in econometrics and mathematical economics from the London School of Economics and Political Science. He obtained his Ph.D. in statistics from Michigan State University. His research interests include statistical decision theory (especially the compound decision theory and empirical Bayes methods), survival models, and applications of stochastic calculus in actuarial mathematics.

Dr. Mashayekhi's address is: Actuarial Science Program, Department of Finance, University of Nebraska-Lincoln, Lincoln NE 68588-0426, USA. Internet address: mmashaye@unlnotes.unl.edu

[†]The author would like to thank the anonymous referee for carefully reading the manuscript and for helpful suggestions that led to a considerable improvement in the presentation of this paper.

1 Introduction

Consider the problem of estimating the mortality rate q_x or p_x with partial data from right-censored and possibly left-truncated observations¹ from a study of n individuals. Suppose the i^{th} individual comes under observation at age $x + r_i$ and is scheduled to be under observation for u_i years until age $x + s_i$, where $u_i = s_i - r_i$ and $0 \leq r_i < s_i \leq 1$. The data are partial in the sense that the exact times of death are not recorded. For each i , the data only show whether the i^{th} individual did or did not die under observation. Here the observable random variables are $\delta_1, \dots, \delta_n$ where

$$\delta_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ person dies under observation; and} \\ 0 & \text{otherwise.} \end{cases}$$

Thus a typical record of data would contain i, x, r_i, u_i , and δ_i .

Because the times of death are not known, one cannot find the product-limit estimator with these data. Even when the exact times of death are known, the product limit estimator based on left-truncated observations (Klein and Moeschberger, 1997, pp. 114-115) can produce an unreasonable estimate of p_x .

The maximum likelihood method does not provide a compelling solution in this case either. The maximum likelihood method requires a distributional assumption that makes it possible to write $u_i q_{x+r_i}$ in terms of q_x . The three well-known assumptions that actuaries use for $0 \leq t \leq 1$ are: (i) the Balducci assumption, i.e., ${}_{1-t}q_{x+t} = (1-t)q_x$; (ii) the assumption of uniform distribution of deaths, i.e., ${}_tq_x = tq_x$; and (iii) the constant force of mortality, i.e., ${}_tq_x = 1 - (1 - q_x)^t$. Under each of these assumptions, except for trivial cases, the likelihood equation $d\mathcal{L}/dq_x = 0$, where

$$\mathcal{L} = \prod_{i=1}^n (1 - u_i q_{x+r_i})^{1-\delta_i} (u_i q_{x+r_i})^{\delta_i}$$

does not have a closed form solution unless n is small. When there is no closed form solution, one may find a solution by numerical methods. As the likelihood equation $d\mathcal{L}/dq_x = 0$ may have multiple roots, it is difficult to determine, however, if the solution obtained by numerical methods is the value of the root that has optimal large sample properties.

¹An observation is said to be right-censored if the individual being observed is alive when the study ends. An observation is said to be left-truncated if the individual entered the study after age x .

Because maximum likelihood estimators are justified mainly by their desired large sample properties, the maximum likelihood approach in this case may not be appealing.

Another method of estimation is the method of moments. This method is one of the oldest statistical estimation methods. One of its biggest advantages over other statistical estimation methods is that it produces easy-to-compute estimates. One of its disadvantages is that it may produce an estimate that is outside the possible range of the parameter. Another disadvantage of the method of moments is that it may produce multiple estimators for the same parameter.

To demonstrate this, consider, for example, estimation of q_x with partial data as described above under the assumption that

$$u_i q_{x+r_i} = u_i \times q_x \tag{1}$$

for each i . The assumed equality in equation (1) is the exact form of the approximation given in equation (6.3) of London (1988). Note that equation (1) cannot be satisfied without restrictions on r_i and u_i . Specifically, equation (1) without restrictions on r_i and u_i gives $0.5q_x = 0.5q_{x+0.5} = 0.5q_x$, which, for $q_x > 0$, contradicts the identity

$$q_x = 0.5q_x + (1 - 0.5q_x) \times 1 - 0.5q_{x+0.5}.$$

The equality in equation (1) is practically plausible in three cases only: (i) with $s_i = 1$ and $r_i = 0$ for all i in which case the equality is trivially true; (ii) under Balducci's assumption with $s_i = 1$ for all i ; and (iii) under the uniform distribution of deaths assumption with $r_i = 0$ for all i . Under these three cases London (1988) (equations (6.7), (6.10), (6.13)) proposes the method of moments estimator given by

$$\hat{q}_x^{(a)} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n u_i} \tag{2}$$

which is obtained by setting the random variable $\sum_{i=1}^n \delta_i$ equal to its expected value and solving for q_x . Another observable random variable that one can equate to its expected value to yield a method of moments estimator is $\sum_{i=1}^n u_i^{-1} \delta_i$, which has expected value equal to ${}_n q_x$. This method of moments estimator is given by

$$\hat{q}_x^{(b)} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{u_i}. \quad (3)$$

Note that $\hat{q}_x^{(a)}$ and $\hat{q}_x^{(b)}$ are linear estimators of q_x . In general let $\omega_1, \dots, \omega_n$ be non-negative weights such that $\sum_{i=1}^n \omega_i = 1$. Because

$$\mathbb{E} \left[\sum_{i=1}^n \omega_i \frac{\delta_i}{u_i} \right] = q_x$$

the method of moments estimator is given by

$$\hat{q}_x^{(\omega)} = \sum_{i=1}^n \omega_i \frac{\delta_i}{u_i}. \quad (4)$$

Clearly $\hat{q}_x^{(a)}$ and $\hat{q}_x^{(b)}$ are special cases of $\hat{q}_x^{(\omega)}$.

Because $\hat{q}_x^{(\omega)}$ is linear in the δ_i/u_i s, it is natural to ask if there are better linear estimators than $\hat{q}_x^{(a)}$ and $\hat{q}_x^{(b)}$. From a Bayesian perspective, one can achieve a better result using the linear Bayes estimator, which is presented in Section 2. As will be seen, the linear Bayes estimator depends on the first two moments of the prior distribution. When these moments are known the linear Bayes estimator is available. If these two moments are unknown, however, they must be estimated and one can use the linear empirical Bayes estimator described in Section 3, which also contains a discussion of the asymptotic optimality of linear empirical Bayes estimators of q_x .

2 The Linear Bayes Estimator

In a Bayes estimation problem, one is faced with a data set consisting of n observable k -dimensional random vectors (k can be 1), $\mathbf{X}_1, \dots, \mathbf{X}_n$, and an unobservable random variable or vector θ . Given $\theta, \mathbf{X}_1, \dots, \mathbf{X}_n$ are mutually independent.

The loss function $L(t, \theta)$ specifies the loss of estimating (predicting) θ by $t = t(\mathbf{X}_1, \dots, \mathbf{X}_n)$. Bayesians are interested in estimators that minimize the expected loss in some sense.

Definition 1. An estimator $\hat{\theta} = \hat{\theta}(\mathbf{X}_1, \dots, \mathbf{X}_n)$ is called a Bayes estimator if

$$\mathbb{E} [L(\hat{\theta}, \theta)] = \min_t \mathbb{E} [L(t(\mathbf{X}_1, \dots, \mathbf{X}_n), \theta)]$$

where $\mathbb{E} []$ denotes the expectation with respect to the joint distribution of all of the random variables involved.

In other words, a Bayes estimator for a given loss function is an estimator that minimizes the expected loss over all estimators. As the basic method of moments estimators are linear (see equation (4)), we will consider linear Bayes estimators.

Definition 2. An estimator $\hat{\theta}^*$ is called linear Bayes if

$$\mathbb{E} [L(\hat{\theta}^*, \theta)] = \min_{a_0, \dots, a_n} \mathbb{E} [L(t(\mathbf{X}_1, \dots, \mathbf{X}_n), \theta)]$$

for t a linear function of the data, i.e., $t = a_0 + \sum_{i=1}^n a_i X_i$.

Observe that for the squared error loss function given by $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, we have $L(1 - \hat{\theta}, 1 - \theta) = (\hat{\theta} - \theta)^2 = L(\hat{\theta}, \theta)$. Hence an estimator $\hat{\theta}$ is a Bayes (linear Bayes) estimator of θ if, and only if, $(1 - \hat{\theta})$ is a Bayes (linear Bayes) estimator of $(1 - \theta)$. Therefore, the linear Bayes (linear empirical Bayes) estimator of p_x is automatically found when we find the linear Bayes (linear empirical Bayes) estimator of q_x .

The following assumption gives a formal description of the model for our estimation problem.

Assumption 1. Let $\theta = q_x$ and $X_i = \delta_i/u_i$, then θ, X_1, \dots, X_n are random variables such that

- 1.1 $\mathbb{P}[0 \leq \theta \leq 1] = 1, \mathbb{P}[\theta = 1] < 1, \text{ and } \mathbb{P}[\theta = 0] < 1;$
- 1.2 Given θ , the random variables X_1, \dots, X_n are uncorrelated; and
- 1.3 $u_i X_i$ is a Bernoulli random variable taking the values 0 or 1 such that

$$\mathbb{P}[u_i X_i = 1] = u_i \theta,$$

where $0 < u_i \leq 1$ is a known constant for $i = 1, \dots, n$.

Assumption 1.3 corresponds to the assumed equality given in equation (1). Under Assumption 1, $\mathbb{E}[X_i|\theta] = \theta$, and $\text{Var}[X_i|\theta] = u_i\theta(1 - u_i\theta)/u_i^2$.

Let $\mu = \mathbb{E}[\theta]$ and $\sigma^2 = \text{Var}[\theta]$. Then we have $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \mathbb{E}[\text{Var}[X_i|\theta]] + \text{Var}[\mathbb{E}[X_i|\theta]] = u_i^{-1}\mu - \mu^2$. Therefore

$$\mathbb{E}[X_i^2] = \text{Var}[X_i] + \mu^2 = u_i^{-1}\mu \quad (5)$$

and, for $i \neq j$,

$$\mathbb{E}[X_iX_j] = \mathbb{E}[\mathbb{E}[X_iX_j|\theta]] = \mathbb{E}[\theta^2] = \mu^2 + \sigma^2. \quad (6)$$

The following theorem gives the linear Bayes estimator of θ , i.e., of q_x . Its proof is given in the appendix.

Theorem 1. *Under Assumption 1 the linear Bayes estimator $\hat{\theta}^*$ of θ under the squared error loss is given by*

$$\hat{\theta}^* = q_x^* = b_0\mu + \sum_{i=1}^n b_iX_i \quad (7)$$

where

$$\alpha_i = [u_i^{-1}\mu - (\mu^2 + \sigma^2)]^{-1}, \quad (8)$$

$$b_i = (1 + \sigma^2 \sum_{i=1}^n \alpha_i)^{-1} \sigma^2 \alpha_i, \quad (9)$$

for $i = 1, \dots, n$, and $b_0 = 1 - \sum_{i=1}^n b_i$.

The next question is the determination of μ and σ^2 . To a purely Bayesian actuary, the prior density of θ , $\pi(\theta)$, is completely known; hence, μ and σ^2 are known so that $\hat{\theta}^*$ can be determined easily from equation (7). An actuary who is not a pure Bayesian, however, would not have an explicitly known prior distribution. In this case the actuary may use either the uniform distribution as a non-informative prior for θ or use the empirical Bayes approach to estimate μ , σ^2 , α_i , and b_i in equation (7). The empirical Bayes approach is described in the next section.

Examples of priors for θ (i.e., for q_x) are:

- $\pi(\theta) = 1$ for $0 < \theta < 1$. This is a non-informative prior because it reflects the actuary's complete ignorance of any prior information on q_x . This is an extreme case.
- Suppose a mortality study is done every three years on a block of policies. In the year 2000 study the actuary feels that mortality has dropped between, say, five and 25 percent from its previous level of $q_x^{(1997)}$ in 1997. In the absence of further information the actuary's prior would be

$$\pi(\theta) = \begin{cases} \frac{1}{0.20q_x^{(1997)}} & \text{for } 0.75q_x^{(1997)} < \theta < 0.95q_x^{(1997)} \\ 0 & \text{otherwise.} \end{cases}$$

The model described in Assumption 1 is similar to the credibility theory model of Bühlmann (1967); it reduces to the Bühlmann (1967) model when $u_i = 1$ for $i = 1, 2, \dots, n$.

3 Linear Empirical Bayes Estimators

In the empirical Bayes approach pioneered by Robbins (1955), one is faced with m independent copies of the same decision problem. In the i^{th} problem there is a random pair (X_i, θ_i) where X_i is observable and θ_i is not observable. Conditional on $\theta_i = \theta$, X_i has a specified density $f(\cdot, \theta)$ for every i . In some of the variations of the empirical Bayes estimation that were later developed (e.g., Bühlmann and Straub (1970) and its generalization in Sundt (1983), or Ghosh and Meeden (1986)) in the i^{th} problem there is an observable random vector $\mathbf{X}_i = (X_{i1}, \dots, X_{in_i})$ where n_i s are not necessarily equal. There is a non-negative loss function $L(t, \theta)$. The unobservable θ_i s are assumed to be i.i.d. with unknown common distribution function $G(\cdot)$.

To put this in the context of a mortality study, suppose there are m similar portfolios of insured lives, and the i^{th} portfolio consists of n_i lives. The j^{th} individual in the i^{th} portfolio comes under observation at age $x + r_{ij}$ and is scheduled to be under observation for u_{ij} years until age $x + s_{ij}$, where $u_{ij} = s_{ij} - r_{ij}$ and $0 \leq r_{ij} < s_{ij} \leq 1$. For each j , the data only show whether the j^{th} individual in the i^{th} portfolio did or did not die under observation. Here the observable random variables are $\delta_{i1}, \dots, \delta_{in_i}$ where

$$\delta_{ij} = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ person in } i^{\text{th}} \text{ portfolio dies under observation; and} \\ 0 & \text{otherwise.} \end{cases}$$

Each individual in the i^{th} portfolio is characterized by an unobservable random mortality rate $\theta_i = q_x^{(i)}$ and the θ_i s are values of an unobservable random sample from the same distribution. The data consist of the available observations as shown in Table 1. The random variables X_{ij} are defined by

$$X_{ij} = \frac{\delta_{ij}}{u_{ij}}$$

for $j = 1, 2, \dots, n_i$ and $i = 1, 2, \dots, m$. The problem is the simultaneous estimation of the θ_i s.

Table 1
Illustration of the Empirical Bayes Problem

Portfolio	Mortality	Outcome			Death or Survival		
	Rate	Observations			Period		
1	θ_1	δ_{11}	...	δ_{1n_1}	u_{11}	...	u_{1n_1}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
i	θ_i	δ_{i1}	...	δ_{in_i}	u_{i1}	...	u_{in_i}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
m	θ_m	δ_{m1}	...	δ_{mn_m}	u_{m1}	...	u_{mn_m}

To avoid needless complications, Robbins assumes the existence of a Bayes decision function t_G such that

$$\mathbb{E}[L(t_G(X_m), \theta_m)] = \min_t \mathbb{E}[L(t(X_m), \theta_m)].$$

Robbins shows that when G is not known (and, hence, t_G is not directly available) for each problem, one may use asymptotically optimal decision rules that use the data from all of the m decision problems. These decision rules asymptotically give us the same risk that we would have with the knowledge of t_G . According to Robbins' definition, a sequence of decision rules $t_m(\cdot) = t_m(X_1, \dots, X_m; \cdot)$ is asymptotically optimal relative to G as $m \rightarrow \infty$ if

$$\mathbb{E}[L(t_m(X_m), \theta_m)] - \mathbb{E}[L(t_G(X_m), \theta_m)] \rightarrow 0 \text{ as } m \rightarrow \infty,$$

where $\mathbb{E}[\cdot]$ denotes the expectation over all random variables. Though $t_m(\cdot)$ is a decision function and not an estimator, its value $t_m(X_m) = t_m(X_1, \dots, X_m; X_m)$ is an empirical Bayes estimator for the m^{th} estimation problem, and, in the context of this paper, its value $t_m(X_k) = t_m(X_1, \dots, X_m; X_k)$ is the empirical Bayes estimator for the k^{th} problem, $k = 1, 2, \dots, m$.²

In the linear empirical Bayes estimation problem considered by Robbins (1983), the minimizing rule is the linear Bayes rule in the sense that it minimizes the Bayes risk for the i^{th} problem within the class of all estimators of the form $aX_i + b$. Thus, t_m is asymptotically optimal if the excess of risk of using t_m over the risk of using the linear Bayes rule converges to zero as the number of problems m increases.

Many variations of the linear empirical Bayes approach have been used by statisticians; see, for example, Morris (1983) for a list of some remarkable examples. These variations usually occur in cases where there are many similar independent estimation problems and the number of observations in each problem is small. In such cases one can do significantly better by borrowing strength from data from other problems. The strength is obtained through estimation of the prior distribution (in unrestricted empirical Bayes) or estimation of the necessary moments of the prior distribution (in the case of linear empirical Bayes) by using similar data. A notable example of linear Bayes (linear empirical Bayes approach) well known to actuaries is the Bühlmann (1967) approach in credibility theory.

The variation that we are considering is slightly different from Robbins' empirical Bayes or linear empirical Bayes in the sense that our m problems are not identical when the sample sizes are different or when the durations of time that different subjects are under observation are

²It must be emphasized that although $t_m(X_k) = t_m(X_1, \dots, X_m; X_k)$ is an estimator for the k^{th} problem, $k = 1, 2, \dots, m$ in the context of this paper, it is not true for what Robbins does. Robbins (1955) uses so-called *delete bootstrap rules* because he has posed his problem in a non-parametric unrestricted empirical Bayes context. Non-delete bootstrap rules, although desirable, are difficult to use in the non-parametric unrestricted empirical Bayes context. In this paper, however, we consider a linear empirical Bayes estimation problem, which can be solved through the estimation of only the first two moments of the prior distribution. This has allowed us to use the more desirable non-delete rules. Specifically, we have used all of the observations to find estimators for the first two moments of the prior distribution and hence the shape of the decision rule. We then have used observations from each problem to find the linear empirical Bayes estimator for that problem. This is not what Robbins (1955) has done. He considers empirical Bayes estimators for the m^{th} problem only.

not equal. Still, we may define the linear empirical Bayes estimators $\hat{\theta}_1^{\text{EB}}, \dots, \hat{\theta}_m^{\text{EB}}$ to be asymptotically optimal if, with $\hat{\theta}_i^*$ denoting the linear Bayes estimator for the i^{th} problem, for each $i = 1, \dots, m$ we have

$$\mathbb{E} [(\hat{\theta}_i^{\text{EB}} - \theta_i)^2] - \mathbb{E} [(\hat{\theta}_i^* - \theta_i)^2] \rightarrow 0 \text{ as } m \rightarrow \infty.$$

The model we are considering is formalized in the following assumption.

Assumption 2. $(X_{11}, \dots, X_{1n_1}, \theta_1), \dots, (X_{m1}, \dots, X_{mn_m}, \theta_m)$ are independent random vectors such that

2.1 $\theta_1, \dots, \theta_m$ are identically distributed random variables with $\mathbb{P}[0 \leq \theta_i \leq 1] = 1$, $\mathbb{P}[\theta_i = 1] < 1$, and $\mathbb{P}[\theta_i = 0] < 1$;

2.2 Conditional on θ_i , the X_{i1}, \dots, X_{in_i} are uncorrelated and;

2.3 $u_{ij}X_{ij}$ is Bernoulli with parameter $u_{ij}\theta_i$ where $0 < u_* \leq u_{ij} \leq 1$ are known numbers; and

2.4 There exists a K such that $2 \leq n_i \leq K < \infty$ for all i .

Assumption 2 is similar to Bühlmann and Straub (1970). In the Bühlmann and Straub model $(\theta_1, X_{11}, \dots, X_{1n_1}), \dots, (\theta_m, X_{m1}, \dots, X_{mn_m})$ are m independent random vectors such that the θ_i s are unobservable and X_{ij} is observable for $i = 1, \dots, m$ and $j = 1, \dots, n_i$. There are functions μ_1 and ν such that

$$\mathbb{E}[X_{it}|\theta_i] = \mu_1(\theta_i)$$

and

$$\text{Cov}[X_{ir}, X_{is}|\theta_i] = \begin{cases} \frac{\nu(\theta_i)}{p_{ir}} & \text{if } r = s \\ 0 & \text{otherwise,} \end{cases}$$

where the p_{ir} s are known constants. In Bühlmann-Straub the n_i s are equal. In later variations, however, n_i s are not necessarily equal. Observe that when u_{ij} s are all equal our model satisfies the above assumptions by choosing $\mu_1(\theta) = \theta$, and $\nu(\theta) = u^{-1}\theta(1 - u\theta)$, and $p_{ij} = 1$, with u being the common value of the u_{ij} s. Also note that in the Bühlmann (1967) model the conditional distributions are not

completely specified. In our model the conditional distributions are completely specified to be Bernoulli.

Assumption 2 is used throughout the rest of this paper and therefore we will not mention it in the statement of every lemma or theorem. In the remainder of this paper all incompletely described limits are as $m \rightarrow \infty$ through positive integers.

Let μ and σ^2 denote the mean and variance of θ_i , respectively. Observe that under Assumption 2 we have

$$\mathbb{E} [X_{ij}] = \mu \quad \text{and} \quad \text{Var} [X_{ij}] = \frac{\mu}{u_{ij}} - \mu^2.$$

Similar to equations (5) and (6), we have

$$\mathbb{E} [X_{ij}^2] = u_{ij}^{-1} \mu$$

and for $k \neq j$

$$\mathbb{E} [X_{ij}X_{ik}] = \mu^2 + \sigma^2. \tag{10}$$

Let

$$\begin{aligned} \bar{X}_{i\bullet} &= \sum_{j=1}^{n_i} \omega_{ij} X_{ij}, \quad N = \sum_{i=1}^m n_i, \quad \bar{X}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^m n_i \bar{X}_{i\bullet}, \quad \text{and} \\ Y_i &= \frac{1}{\binom{n_i}{2}} \sum_{1 \leq j < k \leq n_i} X_{ij} X_{ik}, \end{aligned}$$

where the ω_{ij} s are non-negative weights such that $\sum_{j=1}^{n_i} \omega_{ij} = 1$. We propose using the following estimates for μ and σ^2

$$\hat{\mu} = \bar{X}_{\bullet\bullet} \tag{11}$$

and

$$\hat{\sigma}^2 = \max(0, \bar{Y} - \hat{\mu}^2) \tag{12}$$

respectively, where

$$\bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_i. \tag{13}$$

The linear empirical Bayes estimator of θ_i , based on these estimators of μ and σ^2 is given by

$$\hat{\theta}_i^{\text{EB}} = \hat{q}_x^{(i)} = \hat{b}_{i0}\hat{\mu} + \sum_{j=1}^{n_i} \hat{b}_{ij}X_{ij} \quad (14)$$

where

$$\hat{\alpha}_{ij} = \begin{cases} \frac{1}{u_{ij}^{-1}\hat{\mu} - (\hat{\mu}^2 + \hat{\sigma}^2)} & \text{if } u_{ij}^{-1} - (\hat{\mu}^2 + \hat{\sigma}^2) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

and

$$\hat{b}_{ij} = (1 + \hat{\sigma}^2 \sum_{j=1}^{n_i} \hat{\alpha}_{ij})^{-1} \hat{\sigma}^2 \hat{\alpha}_{ij}, \quad (16)$$

for $i = 1, \dots, m$ and $j = 1, \dots, n_i$, and

$$\hat{b}_{i0} = 1 - \sum_{j=1}^{n_i} \hat{b}_{ij}. \quad (17)$$

It can be proved (see Theorem 2 in the appendix) that the $\hat{\theta}_i^{\text{EB}}$ s are asymptotically optimal linear empirical Bayes estimators in the sense that for every $i = 1, \dots, m$

$$\mathbb{E} [(\hat{\theta}_i^{\text{EB}} - \theta_i)^2] - \mathbb{E} [(\hat{\theta}_i^* - \theta_i)^2] \rightarrow 0. \quad (18)$$

where $\hat{\theta}_i^*$ is the linear Bayes estimator of θ_i .

If we choose $\sigma = 0$, so that the class of prior distributions under consideration reduces to the class of point priors (the traditional frequentist approach) then with $m = 1$, the linear empirical Bayes estimators in equation (14) will be the same as the estimator $\hat{q}_x^{(w)}$ in equation (4).

A natural question to ask now is how do we choose the w_{ij} s? Observe that according to Theorem 2 every choice of w_{ij} s provides an asymptotically optimal estimator. However a smaller variance of $\bar{X}_{..}$ means a better speed of convergence. Because the variance of $\bar{X}_{..}$ is minimized when variance of each $\bar{X}_{i.}$ is minimized and

$$\begin{aligned} \text{Var} [\bar{X}_{i\bullet}] &= \mathbb{E} [\text{Var} [\bar{X}_{i\bullet} | \theta_i]] + \text{Var} [\mathbb{E} [\bar{X}_{i\bullet} | \theta_i]] \\ &= \sum_{j=1}^{n_i} \omega_{ij}^2 \left(\frac{\mu - u_{ij}(\mu^2 + \sigma^2)}{u_{ij}} \right) + \sigma^2, \end{aligned}$$

we need to minimize

$$\sum_{j=1}^{n_i} \omega_{ij}^2 \left(\frac{\mu - u_{ij}(\mu^2 + \sigma^2)}{u_{ij}} \right)$$

subject to the constraint $\sum_{j=1}^{n_i} \omega_{ij} = 1$. Writing the Lagrangian

$$\sum_{j=1}^{n_i} \omega_{ij}^2 \left(\frac{\mu - u_{ij}(\mu^2 + \sigma^2)}{u_{ij}} \right) - \lambda \left(\sum_{j=1}^{n_i} \omega_{ij} - 1 \right)$$

and setting the partial derivatives equal to zero yields the minimizer

$$\omega_{ij}^* = \frac{c_{ij}}{\sum_{j=1}^{n_i} c_{ij}}, \tag{19}$$

where

$$c_{ij} = \frac{u_{ij}}{\mu - u_{ij}(\mu^2 + \sigma^2)}. \tag{20}$$

As ω_{ij}^* depends on the unknown parameters, it is not available. Note, however, that $\mu^2 + \sigma^2 = \mathbb{E}[\theta^2]$. Therefore in cases when θ^2 is so small that its expectation becomes negligible we have

$$\omega_{ij}^* \cong \frac{u_{ij}}{\sum_{j=1}^{n_i} u_{ij}}.$$

The above argument also shows that the choice of weights in the moment estimator of equation (2) is a reasonable choice when q_x is small. One can also use the Chebychev's inequality, similar to the proof of

consistency of $\bar{X}_{\bullet\bullet}$, to show that $\hat{q}_x^{(a)}$ of equation (2) converges in probability to q_x as $n \rightarrow \infty$. Thus when there is a large homogeneous sample available for estimation of q_x there is not much to gain by using the linear empirical Bayes method. The problem, however, is that it is not always feasible to have a large sample of homogeneous subjects. When there is a large sample of subjects that can be broken into many homogeneous groups, one can show by using a variation of the weak law of large numbers (Hannan and Fabian (1985), Theorem 2.3.9) that using the estimator of equation (2) will provide a weighted average of the failure probabilities of the homogeneous groups that are in the large sample. An actuary who uses such a weighted average in the determination of premiums can expect to face some anti-selection by those who feel the premium is unfair to them. Breaking the large sample into many homogeneous groups on the other hand may leave a small number of subjects in each homogeneous group. In such a case one can gain by using a linear empirical Bayes estimator instead of using the moment estimator of equation (2) for each homogeneous sample separately.

4 Concluding Remarks

In this paper we obtain an asymptotically optimal linear empirical Bayes estimator of θ_i , with the yardstick of performance being the risk of the linear Bayes estimator. The main reason for using linear empirical Bayes estimators instead of the empirical Bayes estimators is that linear empirical Bayes estimators exist under milder conditions and are usually much easier to compute. When it is possible to reduce the risk of an asymptotically optimal linear Bayes estimator with a simple adjustment, one should not hesitate to do so.

It is easy to see that by construction we have $\hat{\theta}_i^{\text{EB}} \geq 0$. It is possible, however, that the value of $\hat{\theta}_i^{\text{EB}}$ could become more than 1. Let $\hat{\theta}_i^{**}$ be equal to $\hat{\theta}_i^{\text{EB}}$ when $\hat{\theta}_i^{\text{EB}} \leq 1$ and let $\hat{\theta}_i^{**} = 1$ otherwise. The θ_i s are known to be in $[0,1]$; therefore, we have $\mathbb{E} [(\hat{\theta}_i^{**} - \theta_i)^2] \leq \mathbb{E} [(\hat{\theta}_i^{\text{EB}} - \theta_i)^2]$ because $|\hat{\theta}_i^{**} - \theta_i| \leq |\hat{\theta}_i^{\text{EB}} - \theta_i|$.

We started this paper by considering the survival probabilities as related to life insurance. The method of estimation that we present, however, may find more applications in the casualty insurance. Consider, for example, the case when an insurer who has insured a large number N of drivers is interested in assessing the risk due to severe accidents that cannot happen to a person more than once. Examples

of such accidents include fatal accidents and accidents resulting in a severe disability so that the person will not be able to drive again.

Suppose that the insurer is able to classify the N policy holders according to factors such as age, area, etc. into m homogeneous groups with n_i drivers in the i^{th} group for $i = 1, \dots, m$ such that m is large and each n_i is small. Also suppose that it is reasonable to assume the probability of an accident for the j^{th} driver in the i^{th} class during the policy period is equal to $u_{ij}\theta_i$ where u_{ij} is the duration of time the person is insured and θ_i is the probability of an accident by a typical member of the i^{th} class in a unit interval of time. Let B_{ij} denote the amount of loss the insurer will suffer if the j^{th} driver in the i^{th} class faces an accident.

In this case because each n_i is small and also because when the u_{ij} s are not equal the probabilities of accident during the policy period for different drivers are not equal, the Poisson distribution or the negative binomial distribution will not give a good approximation for the distribution of the number of accidents in each group. Therefore, using a compound Poisson model or compound negative binomial model for each class will not be accurate. In such a case, using the individual risk model (Bowers et al., 1986) for each class can produce more accurate results. In order to use the individual risk model, however, the insurer would need an estimate of θ_i for $i = 1, \dots, m$. In such a case, the method presented in this paper can be used to obtain the desired estimates when the insurer has experience data for these m classes from a past year.

A very important question that every practitioner may ask before using any variations of the empirical Bayes approach is how large should m be? Because answering this question accurately requires knowledge of the rate of convergence of the risk of the empirical Bayes estimator, this question is often a good cause for further research when asymptotic results are obtained through application of convergence theorems such as the Lebesgue Dominated Convergence Theorem. For some results that provide a step for further research in this direction, see Hesselager (1992).

Appendix: The Proofs

In order to prove Theorem 1, we note the following: Suppose that (i) θ, X_1, \dots, X_n are random variables with finite second moments (so that they all belong to the L_2 space, and (ii) the loss function is the squared error loss function given by $L(t, \theta) = (t - \theta)^2$). Then, from the

definition of the L_2 projection (see, for example, Brockwell and Davis 1987, Chapter 2), the Bayes estimator of θ is the L_2 projection of θ on the set of all functions of X_1, \dots, X_n that belong to the L_2 space. The linear Bayes estimator of θ is the L_2 projection of θ on the closed span of $\{1, X_1, \dots, X_n\}$.

Proof: Because $\mathbb{P}[0 \leq \theta \leq 1] = 1$, $\mathbb{P}[\theta = 1] < 1$, and $\mathbb{P}[\theta = 0] < 1$, we have $\mu = \mathbb{E}[\theta] > \mathbb{E}[\theta^2] = \mu^2 + \sigma^2$. Because $0 < u_i \leq 1$, it follows that each α_i is well defined and greater than zero. We must show that $\hat{\theta}^*$ is a version of the L_2 projection of θ on the closure of the linear span of $\{1, X_1, \dots, X_n\}$. Thus it is enough to check that $(\hat{\theta}^* - \theta)$ is L_2 perpendicular to 1 and to X_i for $i = 1, \dots, n$ because, if $\mathbb{E}[\hat{\theta}^* - \theta] = 0$ and $\mathbb{E}[(\hat{\theta}^* - \theta)X_i] = 0$, then for all a_0, \dots, a_n

$$\mathbb{E}\left[(\hat{\theta}^* - \theta)\left(a_0 + \sum_{i=1}^n a_i X_i\right)\right] = 0$$

so that $\hat{\theta}^* - \theta$ is perpendicular to every element of the closed span of $\{1, X_1, \dots, X_n\}$. We have

$$\mathbb{E}[\hat{\theta}^* - \theta] = \left(1 - \sum_{i=1}^n b_i\right)\mu + \sum_{i=1}^n b_i\mu - \mu = 0.$$

So it remains to show that $\mathbb{E}[(\hat{\theta}^* - \theta)X_i] = 0$ for each $i = 1, \dots, n$. Because $\mathbb{E}[\theta X_i] = \mathbb{E}[\mathbb{E}[\theta X_i | \theta]] = \mathbb{E}[\theta \mathbb{E}[X_i | \theta]] = \mathbb{E}[\theta^2] = \mu^2 + \sigma^2$, it is enough to show that $\mathbb{E}[\hat{\theta}^* X_i] = \mu^2 + \sigma^2$. We have

$$\mathbb{E}[\hat{\theta}^* X_i] = b_0 \mu^2 + \sum_{j \neq i} b_j \mathbb{E}[X_j X_i] + b_i \mathbb{E}[X_i^2]. \quad (21)$$

Thus, from equations (5) and (6) and by definition of α_i , it easily follows that the right side of equation (21) is equal to

$$\frac{\mu^2 + \sigma^2(\mu^2 + \sigma^2) \left(\sum_{i=1}^n \alpha_i\right) + \sigma^2 \alpha_i \times \alpha_i^{-1}}{1 + \sigma^2 \sum_{i=1}^n \alpha_i} = \mu^2 + \sigma^2,$$

and Theorem 1 is proved. \square

Lemma 1. Let $\hat{\mu}$ be as defined in equation (11) and $\hat{\sigma}^2$ be as defined in equation (12). Then

$$\hat{\mu} \xrightarrow{P} \mu \tag{22}$$

and

$$\hat{\sigma}^2 \xrightarrow{P} \sigma^2. \tag{23}$$

Proof: Because $u_{ij}X_{ij}$ is Bernoulli and $u_{ij} \geq u_*$, we have $0 \leq X_{ij} < u_*^{-1}$. This gives $0 \leq \bar{X}_{i\bullet} \leq u_*^{-1}$ and hence $\text{Var}[\bar{X}_{i\bullet}] \leq \mathbb{E}[\bar{X}_{i\bullet}^2] \leq u_*^{-2}$. Therefore

$$\text{Var}[\bar{X}_{\bullet\bullet}] = \left(\sum_{i=1}^m n_i \right)^{-2} \sum_{i=1}^m n_i^2 \text{Var}[\bar{X}_{i\bullet}] \leq m^{-1} K u_*^{-2} \xrightarrow{P} 0. \tag{24}$$

Hence, equation (22) follows from equation (24), from Chebychev's inequality, and from the fact that $\mathbb{E}[\bar{X}_{\bullet\bullet}] = \mu$.

From equation (10), it follows that $\mathbb{E}[\bar{Y}] = \mu^2 + \sigma^2$. Because $0 \leq X_{ij} \leq u_*^{-1}$, we have $Y_i \leq u_*^{-2}$ and, hence, $\text{Var}[Y_i] \leq u_*^{-4}$. Therefore $\text{Var}[\bar{Y}] \leq m^{-1} u_*^{-4} \xrightarrow{P} 0$. By Chebychev's inequality it follows that $\bar{Y} \xrightarrow{P} \mu^2 + \sigma^2$. Because $\bar{X}_{\bullet\bullet} \xrightarrow{P} \mu$, it follows that $\bar{X}_{\bullet\bullet}^2 \xrightarrow{P} \mu^2$ and, hence, $\bar{Y} - \bar{X}_{\bullet\bullet}^2 \xrightarrow{P} \sigma^2$. Because $\sigma^2 \geq 0$, continuity of the function $g(x) = \max(0, x)$ gives equation (23). \square

Lemma 2. Suppose $\hat{\mu} \xrightarrow{P} \mu$ and $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$. Let $\hat{\alpha}_{ij}$ be given by equation (15) and $\alpha_{ij} = [u_{ij}^{-1}\mu - (\mu^2 + \sigma^2)]^{-1}$. Let \hat{b}_{ij} be as in equation (16) and

$$b_{ij} = (1 + \sigma^2 \sum_{j=1}^{n_i} \alpha_{ij})^{-1} \sigma^2 \alpha_{ij}.$$

Let $\hat{b}_{i0} = 1 - \sum_{j=1}^{n_i} \hat{b}_{ij}$ and $b_{i0} = 1 - \sum_{j=1}^{n_i} b_{ij}$. Then for each $i = 1, \dots, m$ and $k = 0, 1, \dots, n_i$,

$$\hat{b}_{ik} - b_{ik} \xrightarrow{P} 0. \tag{25}$$

Proof: We prove the lemma by first showing that

$$\hat{\alpha}_{ij} - \alpha_{ij} \xrightarrow{P} 0. \tag{26}$$

Because $0 < u_{ij}^{-1} \leq u_*^{-1}$ we have

$$\left(\frac{\hat{\mu}}{u_{ij}} - (\hat{\mu}^2 + \hat{\sigma}^2) \right) - \frac{1}{\alpha_{ij}} = \frac{\hat{\mu} - \mu}{u_{ij}} - (\hat{\mu}^2 - \mu^2) - (\hat{\sigma}^2 - \sigma^2) \xrightarrow{P} 0. \tag{27}$$

If a_m and a'_m are two sequences such that $a_m \geq a > 0$ and $a_m - a'_m \rightarrow 0$, then eventually $a'_m \geq a/2 > 0$. Hence, eventually

$$\frac{1}{a_m} - \frac{1}{a'_m} = \frac{a'_m - a_m}{a_m a'_m} \rightarrow 0.$$

Therefore, because $\alpha_{ij}^{-1} \geq \mu - (\mu^2 + \sigma^2) > 0$, equation (26) follows from equation (27) by the fact (Billingsley, 1986, p. 274; Royden, 1968, p. 93) that a sequence a_m converges in probability to zero if and only if every subsequence of a_m has a further subsequence that converges to zero with probability 1.

Let $\varepsilon > 0$ and $i \in \{1, \dots, m\}$. Observe that $\sum_{j=1}^{n_i} |\hat{\alpha}_{ij} - \alpha_{ij}| > \varepsilon$ only if for some $j \in \{1, \dots, n_i\}$,

$$|\hat{\alpha}_{ij} - \alpha_{ij}| > n_i^{-1} \varepsilon > K^{-1} \varepsilon.$$

Thus we have

$$\mathbb{P} \left[\left| \sum_{j=1}^{n_i} \hat{\alpha}_{ij} - \sum_{j=1}^{n_i} \alpha_{ij} \right| > \varepsilon \right] \leq \mathbb{P} \left[\sum_{j=1}^{n_i} |\hat{\alpha}_{ij} - \alpha_{ij}| > \varepsilon \right] \tag{28}$$

$$\leq \sum_{j=1}^{n_i} \mathbb{P} [|\hat{\alpha}_{ij} - \alpha_{ij}| > K^{-1} \varepsilon] \rightarrow 0 \tag{29}$$

by equation (26) and the assumption that $n_i \leq K$. This means that $\sum_{j=1}^{n_i} \hat{\alpha}_{ij} - \sum_{j=1}^{n_i} \alpha_{ij} \xrightarrow{P} 0$ and, hence,

$$(1 + \hat{\sigma}^2 \sum_{j=1}^{n_i} \hat{\alpha}_{ij}) - (1 + \sigma^2 \sum_{j=1}^{n_i} \alpha_{ij}) \xrightarrow{P} 0. \tag{30}$$

Because $1 + \sigma^2 \sum_{j=1}^{n_i} \alpha_{ij} \geq 1$ it follows from equation (30) that

$$(1 + \hat{\sigma}^2 \sum_{j=1}^{n_i} \hat{\alpha}_{ij})^{-1} - (1 + \sigma^2 \sum_{j=1}^{n_i} \alpha_{ij})^{-1} \xrightarrow{P} 0. \tag{31}$$

It follows from equations (26) and (31) that for $j = 1, \dots, n_i$, we have $\hat{b}_{ij} - b_{ij} \xrightarrow{P} 0$. Because $n_i \leq K$, it follows that

$$\sum_{j=1}^{n_i} \hat{b}_{ij} - \sum_{j=1}^{n_i} b_{ij} \xrightarrow{P} 0$$

which means we also have $\hat{b}_{i0} - b_{i0} \xrightarrow{P} 0$, and the proof is complete. \square

Theorem 2. . Let $\hat{\mu}$ be as in Lemma 1. For $j = 0, 1, \dots, n_i$ let \hat{b}_{ij} be as defined in Lemma 2. Let $\hat{\theta}_i^{EB} = \hat{b}_{i0}\hat{\mu} + \sum_{j=1}^{n_i} \hat{b}_{ij}X_{ij}$. Then $\hat{\theta}_i^{EB}$ is an asymptotically optimal linear empirical Bayes estimator in the sense that for every $i = 1, \dots, m$ with $\hat{\theta}_i^*$ denoting the linear Bayes estimator of θ_i ,

$$\mathbb{E} [(\hat{\theta}_i^{EB} - \theta_i)^2] - \mathbb{E} [(\hat{\theta}_i^* - \theta_i)^2] \rightarrow 0. \tag{32}$$

Proof: From Lemma 2, it easily follows that $\hat{\theta}_i^{EB} - \hat{\theta}_i^* \xrightarrow{P} 0$. Because $0 \leq X_{ij} < u_*^{-1}$, we obtain that $\hat{\theta}_i^{EB}$ and $\hat{\theta}_i^*$ are both bounded. We also have $0 \leq \theta_i \leq 1$. Therefore

$$|(\hat{\theta}_i^{EB} - \theta_i)^2 - (\hat{\theta}_i^* - \theta_i)^2| = |\hat{\theta}_i^{EB} + \hat{\theta}_i^* - 2\theta_i| \cdot |\hat{\theta}_i^{EB} - \hat{\theta}_i^*| \xrightarrow{P} 0.$$

Because $(\hat{\theta}_i^{EB} - \theta_i)^2 - (\hat{\theta}_i^* - \theta_i)^2$ is bounded, the assertion of the theorem follows by the bounded convergence theorem. \square

References

- Billingsley, P. *Probability and Measure*, Second Edition. New York, N.Y.: Wiley, 1986.
- Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A., and Nesbitt, C.J. *Actuarial Mathematics*, Schaumburg, Ill.: Society of Actuaries, 1986.
- Brockwell, P.J. and Davis, R.A. *Time Series: Theory and Methods*. New York, N.Y.: Springer-Verlag, 1987.
- Bühlmann, H. "Experience Rating and Credibility." *ASTIN Bulletin* 4 (1967): 199-207.
- Bühlmann, H. and Straub, E. "Glaubwürdigkeit für Schadensätze." *Bulletin of the Swiss Association of Actuaries* 70 (1970): 111-133. English translation by C.E. Brooks.
- Fabian, V. and Hannan, J. *Introduction to Probability and Mathematical Statistics*. New York, N.Y.: John Willey & Sons, 1985.
- Ghosh, M. and Meeden, G. "Empirical Bayes Estimation in Finite Population Sampling." *Journal of the American Statistical Association* 81 (1986): 1058-1062
- Klein, J.P. and Moeschberger, M.L. *Survival Analysis*. New York, N.Y.: Springer-Verlag, 1997.
- London, D. *Survival Models and Their Estimation*, Third Edition. Winsted, Conn.: Actex, 1997.
- Hesselager, O. "Rates of Risk Convergence of Empirical Linear Bayes Estimators." *Scandinavian Actuarial Journal* (1992): 88-94.
- Morris, C.N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (1983): 47-59.
- Robbins, H. "An Empirical Bayes Approach to Statistics." *Proceedings of the Third Berkeley Symposium of Mathematical Statistics Probability* 1 (1955): 157-164.
- Robbins, H. "The Empirical Bayes Approach to Statistical Decision Problems." *Annals of Mathematics and Statistics* 35 (1964): 1-20.
- Robbins, H. "Some Thoughts on Empirical Bayes Estimation." *Annals of Statistics* 11 (1983): 713-723.
- Royden, H.L. *Real Analysis*, Second Edition. New York, N.Y.: Macmillan, 1968.
- Sundt, B. "Parameter Estimation in Some Credibility Models." *Scandinavian Actuarial Journal* (1983): 239-255.

Controlling the Solvency Interaction Among a Group of Insurance Companies

Alexandros Zimbidis* and Steven Haberman†

Abstract

Pooling of risks is an efficient risk management technique used by large employee benefit schemes of multinational companies to self-insure their retirement and other benefit obligations. This technique forms a basis for formulating a general control theoretic model for the interaction between insurance companies within a pooling network. The objective of these insurance companies is to avoid insolvency yet maintain stable premium and surplus processes. A general control system of equations that is used as a model for the interaction of m insurance companies within the network is first analyzed. An analytic solution is provided. Questions concerning the stability and optimal parameter design for the system are investigated. The special case of two identical companies is analyzed in detail.

Key words and phrases: control theory, self-insurance, pooling, stability, optimal parameter design, feedback mechanism

*Alexandros Zimbidis Ph.D., F.H.A.S., is a lecturer in the statistics department at Athens University of Economics and Business, Greece. He received his Ph.D. from the City University, London and is a Fellow of the Hellenic Actuarial Society. He has been working for Allianz Life in Greece as an actuary and as Director of Group Insurance Department since 1993. His current research interests include control theory with applications to insurance systems and mathematical methods for pension funds.

Dr. Zimbidis's address is: Allianz Life, 124 Kifisias Ave, 11526 Athens, GREECE. Internet address: alexzimbidis@allianz.gr

†Steven Haberman Ph.D., D.Sc., F.I.A., A.S.A., F.S.S., is a professor of actuarial science and Dean of the School of Mathematics at the City University, London. He has published widely in actuarial and related fields. His current research interests include mortality forecasting, morbidity, modeling, premium rating, and pension funding.

Dr. Haberman's address is: City University, Northampton Square, London EC1V0HB, UNITED KINGDOM. Internet address: s.haberman@city.ac.uk

1 Introduction

Since the 1980s, the rise in numbers, size, and shape of large multinational corporations have created a demand for special insurance products. As parent corporations exerted more control over their subsidiaries, the demand for insurance to cover contingencies in different countries grew. Insurance companies have responded to this demand by constructing multinational insurance networks. These networks are established through special reinsurance agreements between affiliated insurance companies (William M. Mercer, 1988).

One of the most important products sold through these networks is the pooling arrangement. Pooling is a special kind of self-insurance established to manage risks. For example, a multinational corporation with employee benefit schemes in two or more countries may use self-insurance to cover benefits as they are needed for all of their employees (Hart et al., 1996).

Two basic problems arise with developing models of pooling arrangements:

- Specifying a model to describe the premium rating process associated with sharing the claims experience of each insurance company in the pool; and
- Specifying a model for describing the interaction of the surpluses among the insurance companies participating in the pool.

The specification of these models is used as a starting point in the formulation of an optimal control theoretic model of the overall interactions among the group of insurance companies in the network.

Optimal control theory was developed in the late 1950s by scientists and engineers to investigate the properties of dynamic systems of difference or differential equations. As it is often difficult to obtain analytic solutions for many dynamic systems, control theory is concerned with the examination of the qualitative properties of these systems. One of the important qualitative properties is the stability of the system.

The stability of a system refers to the way the system reacts to different external input signals, the way it returns to its initial state or to a designated state, and whether or not it remains within an acceptable region of this state. In the insurance context, stability is directly related to the level of the surplus. A stable insurance system can react effectively by anticipating the appropriate premium (output variable) to any claim (input variable) pattern in order to maintain (in the long run) the

surplus (state variable) and consequently maintain the insolvency risk at an acceptable level.

Since 1980, actuaries have applied the results of control theory to actuarial problems. Balzer and Benjamin (1980), Martin-Löf (1983, 1994), Vandebroek and Dhaene (1990), Loades (1998), Runggaldier (1998), Schäl (1998), Chang (2000), and Zimbidis and Haberman (2001) have produced interesting actuarial papers using control theory methods and techniques to solve practical actuarial problems. Control theory may be used in other problems in which there exists an interaction between two or more insurance companies or between different lines of insurance businesses.

We have two main objectives for this paper: (i) to provide a comprehensive and convenient model for the interaction of the surplus among a group of insurance companies within the pooling network and the associated control actions that may be necessary for the management of the network; and (ii) to analyze the resulting system of equations that arise when we consider the control theory approach to solving this insurance problem.

The paper is organized as follows: Section 2 describes the assumptions and notation used throughout the paper. Section 3 introduces the general control model with m insurance companies in the network and the resulting system of equations and its solution. Certain properties of the solution, such as stability and optimality, are discussed in Section 4. Section 5 provides a detailed study of the model and its solution in the simpler case of two identical insurance companies in the network. A summary and conclusions are provided in Section 6. The appendix provides an algorithm for computing the determinant for a key matrix used in our analysis.

2 Assumptions and Notations

Suppose there are m insurance companies participating in a multinational insurance network that operates in m countries (one insurance company per country) covering the risks associated with the benefit payments from the multinational corporation's international employee benefits scheme. A typical employee benefits scheme may include some or all of the following benefits: term life insurance, accidental death and dismemberment insurance, permanent/temporary disability insurance, and medical benefits.

At the end of each year the accumulated surplus (whether it is positive or negative) is redistributed within the network of insurance com-

panies under a specific set of rules. The course of action mandated by these rules is enforced by the holding company or by a neutral central unit that coordinates the network in order to smooth the operational result and solvency requirement of each company.

The insurance companies all use the same experience rating procedure to calculate annual premiums. The experience rating procedure has the following characteristics:

- Experience rating is based on the most recently available claims experience;
- There is a time delay of f years, i.e., it takes f years for incurred claims to be fully reported, processed, and settled. Thus the available claim information at the beginning of the n^{th} year (or at the end of $(n - 1)^{\text{th}}$ year) refers to the experience of the years $n - f - 1$, $n - f - 2$, $n - f - 3$, ..., 2, 1, 0, i.e., years prior to and inclusive of year $n - f - 1$;
- Premiums are calculated annually at the beginning of each year according to a base premium and a profit sharing scheme;
- The base premium is calculated using the most recently available claims experience and taking into account the necessary expense margins;
- The profit-sharing scheme mandates an extra modification of the base premium through a refund (charge) to the policyholder a certain percentage of the benefit scheme's total accumulated surplus (deficit). This correction is aimed at driving the accumulated surplus to zero in the long run; and
- Each company passes to the other $(m - 1)$ companies a pre-determined percentage of its accumulated surplus at the end of each year.

In general the predetermined percentages are not equally divided and are defined by a matrix Λ , called the *harmonization matrix*, that governs the surplus exchange. That is

$$\Lambda = [\lambda_{ij}] \in \mathbb{R}^{m \times m}$$

where $\lambda_{ij} > 0$ is the predetermined percentage of surplus that the i^{th} company passes to j^{th} company. This obviously implies that

$$\sum_{j=1}^m \lambda_{ij} = 1 \quad \text{for } i = 1, 2, \dots, m. \quad (1)$$

The quantity λ_{ii} , $i = 1, 2, \dots, m$ determines the percentage of surplus retained by the i^{th} company. It is further assumed that each company has its own operational parameter values for expenses, feedback, accumulation, and inflation factors.

The following notations are used throughout the paper:

m = Number of insurance companies participating in the multinational network.

f = Length of time delay (measured in years).

e_k = Expense factor for the k^{th} company, i.e., $(1 - e_k) \times$ Gross Premium is the margin for expenses. The expense factor vector is $\mathbf{e} = (e_1, e_2, \dots, e_m)$.

R_k = Accumulation factor ($R_k = 1 + j_k$), using an annual rate of investment return of j_k for the k^{th} company. The vector for the accumulation factor is $\mathbf{R} = (R_1, R_2, \dots, R_m)$.

F_k = Inflation factor ($F_k = 1 +$ inflation rate) of the k^{th} company. This factor indicates internal growth of the total annual claims, attributable to inflation or to business growth. The vector for inflation is $\mathbf{F} = (F_1, F_2, \dots, F_m)$.

λ_{ij} = Interaction factor, $i, j = 1, 2, \dots, m$, is the proportion of surplus that the i^{th} company passes to the j^{th} company and constitutes the harmonization matrix $\mathbf{\Lambda} = (\lambda_{ij})$

ε_k = Profit sharing factor (feedback factor) for the k^{th} company, which includes both the local and international premium repayments and determines the percentage of accumulated surplus repaid to the policyholders. The vector of profit sharing factors is $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m)$.

$C_{k,n}$ = Actual total amount of annual incurred claims for the k^{th} company in the n^{th} year, for $k = 1, \dots, m$.

$\hat{C}_{k,n}$ = Estimated total expected annual incurred claims in year n , i.e., in $(n - 1, n)$. There is a delay of f years in updating information. The $\hat{C}_{k,n}$ is a weighted average of the inflation-adjusted claims

over the two most recent years where data are available, i.e., for $k = 1, \dots, m$

$$\hat{C}_{k,n} = \frac{1}{M_k} (F_k^{2+f} C_{k,n-f-2} + F_k^{1+f} C_{k,n-f-1}) \quad (2)$$

where

$$M_k = F_k^{1+f} (1 + F_k) \quad (3)$$

is a normalizing constant.

$P_{k,n}$ = Gross annual premium paid at the end of the n^{th} year for the k^{th} company. The gross premium is determined as an expense-adjusted premium $P_{k,n}^{(e)}$ less the surplus adjustment, where

$$P_{k,n}^{(e)} = \hat{C}_{k,n} + (1 - e_k) P_{k,n} = \frac{\hat{C}_{k,n}}{e_k}.$$

It follows that

$$\begin{aligned} P_{k,n} &= P_{k,n}^{(e)} - \varepsilon_k S_{k,n-f-1} \\ &= \frac{1}{e_k} \hat{C}_{k,n} - \varepsilon_k S_{k,n-f-1} \end{aligned} \quad (4)$$

for $k = 1, 2, \dots, m$. Equation (4) is also called the *decision function*.

$S_{k,n}$ = Accumulated surplus at the end of the n^{th} year for the k^{th} company where

$$S_{k,n} = R_k \sum_{i=1} \lambda_{ik} S_{i,n-1} + e_k P_{k,n} - C_{k,n} \quad (5)$$

for $k = 1, 2, \dots, m$.

The quantities m , f , e_k , R_k , ε_k , and λ_{ij} are assumed to be constant over time.

This set of assumptions is used as a basis to derive a model and a system of equations and to examine the analytical solution of this system, its stability, and the optimal parameter design with respect to

the interaction arising from the surplus exchange process. The formulation of the problem is similar to that of Balzer and Benjamin (1980), Balzer (1982), Benjamin (1984), and Zimbidis and Haberman (2001). These authors have investigated the stability and parameter design of a single company, consequently without the presence of any interaction phenomenon.

3 The Model and System of Equations

From the point of view of control theory, claims may be considered as an input variable, the surplus as a state variable, and premiums as an output variable. The whole system (i.e., the multinational company's employee benefit scheme) starts from an initial value for the first year's premium, then claims data provide the input background for the development of the surplus level—the surplus represents the state of the system. Finally, using both claims (directly) and surplus information through a feedback¹ mechanism, a decision function is built for premium development. The amount of feedback action is not obviously determined. The level of the state variable and how much is fed back to the system must be evaluated carefully in order to achieve and/or maintain the required stability.

For the k^{th} company, the n^{th} year's premium and surplus are determined using the following model:

$$P_{k,n} = \frac{F_k^{2+f}}{M_k e_k} C_{k,n-f-2} + \frac{F_k^{1+f}}{M_k e_k} C_{k,n-f-1} - \varepsilon_k S_{k,n-f-1} \quad (6)$$

and

$$S_{k,n} = R_k \lambda_{11} S_{1,n-1} + \dots + R_k \lambda_{m1} S_{m,n-1} + \frac{F_k^{2+f}}{M_k} C_{k,n-f-2} + \frac{F_k^{1+f}}{M_k} C_{k,n-f-1} - e_k \varepsilon_k S_{k,n-f-1} - C_{k,n}, \quad (7)$$

for $k = 1, 2, \dots, m$.

Each of the m insurance companies generates its own system of equations. These systems, however, cannot be solved independently

¹In this context, a feedback mechanism can be used to measure the surplus level and calculate how much of this surplus (deficit) should be refunded (charged) to policyholders. In other words, through a feedback mechanism we decide how much of the state information must be fed back to the system.

where $\mathbf{x}_n \in \mathbb{R}^{m(1+f)}$, $\mathbf{y}_n \in \mathbb{R}^m$, and $\mathbf{u}_n \in \mathbb{R}^{m(3+f)}$.

It must be understood that the inputs, \mathbf{u}_n , are determined using the actual $C_{k,n-j}$ s when they are available or $\hat{C}_{k,n-j}$ s when the actual $C_{k,n-j}$ s are not available. In other words the following substitution is used:

$$C_{k,n-j} \begin{cases} \text{is replaced by } \hat{C}_{k,n-j} & \text{for } j = 0, 1, \dots, f; \\ \text{remains unchanged} & \text{for } j = f + 1, f + 2, \dots \end{cases}$$

For ease of exposition, we introduce four matrices **A**, **B**, **C**, **D** whose elements are themselves matrices.

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1m} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2m} \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_{m1} & \mathbf{A}_{m2} & \cdots & \mathbf{A}_{mm} \end{bmatrix} \in \mathbb{R}^{m(1+f)} \times \mathbb{R}^{m(1+f)},$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1m} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2m} \\ \vdots & \vdots & & \vdots \\ \mathbf{B}_{m1} & \mathbf{B}_{m2} & \cdots & \mathbf{B}_{mm} \end{bmatrix} \in \mathbb{R}^{m(1+f)} \times \mathbb{R}^{m(3+f)},$$

$$\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_m] \in \mathbb{R}^m \times \mathbb{R}^{m(1+f)},$$

$$\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_m] \in \mathbb{R}^m \times \mathbb{R}^{m(3+f)}.$$

The elements of the super-matrices **A**, **B**, **C**, and **D** are defined below:

$$\mathbf{A}_{ii} = \begin{bmatrix} R_i \lambda_{ii} & 0 & 0 & \cdots & 0 & -e_i \varepsilon_i \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{(1+f)} \times \mathbb{R}^{(1+f)},$$

$$\mathbf{A}_{i,j} = \begin{bmatrix} R_i \lambda_{ji} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(1+f)} \times \mathbb{R}^{(1+f)},$$

$(i \neq j)$

$$\mathbf{C}_i = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & -\varepsilon_i \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{R}^m \times \mathbb{R}^{(1+f)},$$

$$\mathbf{B}_{ii} = \begin{bmatrix} -1 & 0 & \cdots & 0 & \frac{F_i^{1+f}}{M_i} & \frac{F_i^{2+f}}{M_i} \\ 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{(1+f)} \times \mathbb{R}^{(3+f)},$$

with $\mathbf{B}_{ij} = \mathbf{O}$ for $i \neq j$, and

$$\mathbf{D}_i = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \frac{F_i^{1+f}}{M_i e_i} & \frac{F_i^{2+f}}{M_i e_i} \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{R}^m \times \mathbb{R}^{(3+f)}.$$

The system (equation (8)) can now be written as:

$$\left. \begin{aligned} \mathbf{x}_n &= \mathbf{A}\mathbf{x}_{n-1} + \mathbf{B}\mathbf{u}_n \\ \mathbf{y}_n &= \mathbf{C}\mathbf{x}_{n-1} + \mathbf{D}\mathbf{u}_n \end{aligned} \right\}. \quad (9)$$

Following Cadzow (1973), the analytical solution to equation (9) is given by:

$$\left. \begin{aligned} \mathbf{x}_n &= \mathbf{A}^n \mathbf{x}_0 + \sum_{k=0}^{n-1} \mathbf{A}^k \mathbf{B} \mathbf{u}_{n-k-1}, \\ \mathbf{y}_n &= \mathbf{C} \mathbf{A}^{n-1} \mathbf{x}_0 + \mathbf{C} \sum_{k=0}^{n-2} \mathbf{A}^{n-2-k} \mathbf{B} \mathbf{u}_k + \mathbf{D} \mathbf{u}_n \end{aligned} \right\} \quad (10)$$

4 Properties of the Solution

Obtaining the analytical solution, equation (10), is not always necessary in order to understand the important properties of the system. We will now explore the stability of the system. This requires extensive use of eigenvalues and the characteristic function of a matrix.

The expression $|\rho \mathbf{I} - \mathbf{A}|$, which is the determinant of matrix $(\rho \mathbf{I} - \mathbf{A})$, can be expressed as a polynomial of ρ . This polynomial, $\phi_m(\rho)$, is called the characteristic polynomial of \mathbf{A} and is written as

$$\phi_m(\rho) = \sum_{r=0}^{m(f+1)} a_r \rho^r. \quad (11)$$

It follows that ρ_r is an eigenvalue of the matrix \mathbf{A} if and only if $\phi_m(\rho_r) = |\rho_r \mathbf{I} - \mathbf{A}| = 0$.

The necessary definitions, theorems, and results of linear algebra used in the remainder of the paper may be found in Healy (1995). The appendix contains an algorithm for calculating the characteristic function of \mathbf{A} .

4.1 Stability Analysis

For a dynamic system of the form described in equation (9), a point \mathbf{x}_* is called an equilibrium point if and only if \mathbf{x}_* satisfies the equation

$$\mathbf{A} \mathbf{x}_* = \mathbf{x}_*.$$

This equation clearly has at least one solution, i.e., the zero solution. Under certain conditions, however, the zero solution is the unique solution. Specifically, if $\det(\mathbf{A}) \neq 0$, then zero solution is the unique solution.

This statement is proved by considering the determinant of matrix \mathbf{A} and confirming that $\det(\mathbf{A})$ differs from zero.

It is easy to show (see Section 5.2 for an outline of this) that

$$\det(\mathbf{A}) = (-1)^{m(1+f)} e_1 e_2 \cdots e_m \varepsilon_1 \varepsilon_2 \cdots \varepsilon_m,$$

which is different from zero for $e_r, \varepsilon_r \neq 0$ for $r = 1, 2, \dots, m$. Consequently, in most practical situations $\mathbf{0}$ is the only equilibrium point of the system.

According to Cadzow (1973, Chapter 3, page 106), a dynamic system of the form described in equation (9) is said to be stable at a state point \mathbf{x} (also called a stability point) if and only if the trajectory of the system that starts within a neighborhood of \mathbf{x} remains close to \mathbf{x} at all future times. The mathematical implication of this definition is that a state point \mathbf{x} is a stability point if and only if all the roots of the characteristic polynomial of the \mathbf{A} matrix have modulus less than unity.

It follows that the dynamic system of equation (9), is stable if the modulus of each eigenvalue of \mathbf{x} is less than unity, i.e.,

$$|\rho_r| < 1 \tag{12}$$

for $r = 1, 2, \dots, m(1+f)$ where ρ_r is the r^{th} eigenvalue of \mathbf{A} . It follows that the system is unstable if $|\rho_r| > 1$ for any k . Hence a sufficient condition for the system to be unstable is $\prod_{r=1}^{m(1+f)} |\rho_r| > 1$. But, as

$$\phi_m(\rho) = \prod_{r=1}^{m(1+f)} (\rho - \rho_r),$$

a sufficient condition for the system to be unstable is $|\phi_m(0)| > 1$.

It is easy to prove that the first and last coefficients of $\phi_m(\rho)$ are $a_{m(1+f)} = 1$ and $a_0 = e_1 e_2 \cdots e_m \varepsilon_1 \varepsilon_2 \cdots \varepsilon_m$. Applying the above criterion for instability requires

$$\prod_{r=1}^{m(1+f)} |\rho_r| = \prod_{r=1}^m |e_r \varepsilon_r| > 1. \tag{13}$$

In practice, expenses and profits will almost always be such that $0 < e_r < 1$ and $0 < \varepsilon_r < 1$ for $r = 1, 2, \dots, m$, so most practical systems will not satisfy equation (13)'s criterion for instability. This does not mean, however, that the system will automatically be stable.

4.2 Optimal Parameter Choices

The criterion for parameter optimality is defined in terms of the *fastest response time* of the system to different input signals.² A set of the parameter values is optimal if and only if the state vector moves to a desirable state (normally toward a stability point) faster than under any other choice of the parameter values, irrespective of the form, nature, or magnitude of the input vector. Below we describe a method that is useful in finding the approximate values of the optimal set of parameters.

Let $\mathbf{s} = (\mathbf{e}, \boldsymbol{\varepsilon}, \mathbf{R}, \boldsymbol{\Lambda})$ denote a particular choice for the parameter values and \mathbb{S} denote the closed set of all possible choices for \mathbf{s} . Define $\mathbf{A}(\mathbf{s})$ be the \mathbf{A} matrix derived from the choice of \mathbf{s} . If $\rho_r(\mathbf{s})$ is the r^{th} eigenvalue of $\mathbf{A}(\mathbf{s})$, let $\rho^{\max}(\mathbf{s})$ be the maximum absolute value of the eigenvalues of matrix $\mathbf{A}(\mathbf{s})$, i.e.,

$$\rho^{\max}(\mathbf{s}) = \max\{|\rho_1(\mathbf{s})|, |\rho_2(\mathbf{s})|, \dots, |\rho_{m(1+f)}(\mathbf{s})|\},$$

then the speed of the response of the system depends on the maximum absolute value of the eigenvalue, $\rho^{\max}(\mathbf{s})$. The smaller the value of $\rho^{\max}(\mathbf{s})$, the faster the response of the system.

Suppose there is an $\mathbf{s}^* \in \mathbb{S}$ such that

$$\rho^{\max}(\mathbf{s}^*) = \rho^* = \min\{\rho^{\max}(\mathbf{s}) : \mathbf{s} \in \mathbb{S}\}, \quad (14)$$

it follows that

$$\rho^* \geq \sqrt[m(1+f)]{\prod_{r=1}^m e_r \varepsilon_r}. \quad (15)$$

The minimization of the maximum root of $\phi_m(\rho)$ is easily obtained in two special cases:

Case 1: If $\phi_m(\rho)$ has a single real root with multiplicity $m(1 + f)$, i.e.,

$$\phi_m(\rho) = (\rho - \rho_0)^{m(1+f)}$$

²The response time refers to the time it takes for the system output (or state) variables to return to the initial state or move to a designated point.

where

$$\rho^* = \left(\prod_{r=1}^m e_r \varepsilon_r \right)^{\frac{1}{m(1+f)}}$$

Using a binomial expansion yields

$$\phi_m(\rho) = \sum_{r=0}^{m(1+f)} \binom{m(1+f)}{r} (-\rho^*)^{m(1+f)-r} \rho^r,$$

i.e., the coefficient of ρ^r is a_r where

$$a_r = \binom{m(1+f)}{r} (-\rho_0)^{m(1+f)-r}$$

for $r = 0, 1, \dots, m(1+f)$. This gives a system of $m(1+f) + 1$ equations for a_r that contains $m^2 + 3m$ (control) parameters, i.e., $e_1, \dots, e_m, \varepsilon_1, \dots, \varepsilon_m, R_1, \dots, R_m, \lambda_{11}, \dots, \lambda_{mm}$. Some of these may be fully controlled (the ε vector and the Λ matrix) or partially controlled (the e and R vectors). Our aim should be the optimal selection of all the controlled parameters such that the system becomes solvable.

Case 2: If $\phi_m(\rho)$ is such that all of its roots lie on the circumference of the circle in the complex plane centered at the origin and with radius ρ^* , where

$$\rho^* = \left(\prod_{r=1}^m e_r \varepsilon_r \right)^{\frac{1}{m(1+f)}}$$

In this case $\phi_m(\rho)$ has the form

$$\phi_m(\rho) = \rho^{m(1+f)} + \prod_{r=1}^m e_r \varepsilon_r \quad (16)$$

and its roots are proportional to the complex roots of $m(1+f)\sqrt{-1}$, i.e.,

$$\rho_j = \rho^* \left(\cos \left(\frac{(2r-1)\pi}{m(1+f)} \right) + i \sin \left(\frac{(2r+1)\pi}{m(1+f)} \right) \right)$$

where $j = 1, 2, \dots, m(1+f)$. Notice that this case appears when

$$a_{m(1+f)-1} = a_{m(1+f)-2} = \dots = a_2 = a_1 = 0, \quad (17)$$

i.e., all the a_r coefficients are zero except the first and last ones.

For large values of m the system of equations is rather complicated and there is no obvious choice of a choice of \mathbf{s}^* that results in a root with multiplicity $m(1+f)$. In such situations, we are forced to follow a trial and error procedure to determine \mathbf{s}^* . In other words, if $\hat{\mathbb{S}}$ is the set of all *practical parameter choices* then we can calculate $\rho^{\max}(\mathbf{s})$ for each $\mathbf{s} \in \hat{\mathbb{S}}$ then choose the \mathbf{s}^* that produces the minimum $\rho^{\max}(\mathbf{s})$.

5 The Special Case of Two Identical Companies

To further illustrate the ideas described in Section 4, let us consider a simple situation with two insurance companies ($m = 2$) in the network and a one year delay factor ($f = 1$). In order to facilitate the calculations, we assume that the companies are identical with respect to operational parameters, i.e., $e_1 = e_2 = e$, $R_1 = R_2 = R$, $F_1 = F_2 = F$, $M = F^2 + F^3$, $\epsilon_1 = \epsilon_2 = \epsilon$, and $\lambda_{12} = \lambda_{21} = \lambda$. As we assumed each company passes the same percentage λ of its surplus fund to the other company, the harmonization matrix Λ is

$$\Lambda = \begin{bmatrix} 1 - \lambda & \lambda \\ \lambda & 1 - \lambda \end{bmatrix}.$$

These operational assumptions are reasonable because multinational networks tend to be composed of similar companies with respect to operational matters. The assumption of identical companies is necessary in order to obtain closed form analytical solutions and results.

5.1 The Solution

The matrix A is given by

$$A = \begin{pmatrix} R(1-\lambda) & -e\varepsilon & R\lambda & 0 \\ 1 & 0 & 0 & 0 \\ R\lambda & 0 & R(1-\lambda) & -e\varepsilon \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

and its characteristic polynomial is

$$\phi_2(\rho) = \begin{vmatrix} \rho - R(1-\lambda) & e\varepsilon & -R\lambda & 0 \\ -1 & \rho & 0 & 0 \\ -R\lambda & 0 & \rho - R(1-\lambda) & e\varepsilon \\ 0 & 0 & -1 & \rho \end{vmatrix}.$$

Developing this determinant across the second row, yields

$$\phi_2(\rho) = (\rho^2 - R\rho + e\varepsilon)(\rho^2 - R(1-2\lambda)\rho + e\varepsilon).$$

The four roots of this quartic polynomial are

$$\rho_1, \rho_2 = \frac{R \pm \sqrt{R^2 - 4e\varepsilon}}{2} \quad (18)$$

and

$$\rho_3, \rho_4 = \frac{R(1-2\lambda) \pm \sqrt{R^2(1-2\lambda)^2 - 4e\varepsilon}}{2}. \quad (19)$$

We now examine the behavior of the system with respect to three types of inputs: spike signals, step signals, and sine signals assuming the zero initial condition $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{y}_0 = \mathbf{0}$ for any situation.

5.1.1 Spike Signals

Let us assume that a spike signal³ appears as the input of the first subsystem while the second subsystem has a zero input, i.e.,

³In practice, a spike input signal may be interpreted as the appearance of an isolated unexpected claim into the system.

$$C_{1,n} = \begin{cases} 1, & n = 0 \\ 0, & n = 1, 2, \dots \end{cases}$$

and $C_{2,n} = 0$ for $n = 0, 1, \dots$. The input vectors are

$$\mathbf{u}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{u}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{u}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{u}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and $\mathbf{u}_n = \mathbf{0}$ for $n = 4, 5, \dots$. Substituting these values of \mathbf{u}_n for $n = 0, 1, 2, \dots$ in the general solution given in equation (10) gives:

$$\begin{aligned} \mathbf{x}_n &= \mathbf{A}^n \mathbf{x}_0 + \mathbf{A}^{n-1} \mathbf{B} \mathbf{u}_0 + \mathbf{A}^{n-2} \mathbf{B} \mathbf{u}_1 + \mathbf{A}^{n-3} \mathbf{B} \mathbf{u}_2 + \mathbf{A}^{n-4} \mathbf{B} \mathbf{u}_3 \\ \mathbf{y}_n &= \mathbf{C} \mathbf{A}^{n-1} \mathbf{x}_0 + \mathbf{C} \mathbf{A}^{n-2} \mathbf{B} \mathbf{u}_0 + \mathbf{C} \mathbf{A}^{n-3} \mathbf{B} \mathbf{u}_1 \\ &\quad + \mathbf{C} \mathbf{A}^{n-4} \mathbf{B} \mathbf{u}_2 + \mathbf{C} \mathbf{A}^{n-5} \mathbf{B} \mathbf{u}_3 \end{aligned}$$

for $n = 5, 6, \dots$. Because $\mathbf{x}_0 = \mathbf{0}$, and $\mathbf{B} \mathbf{u}_1 = \mathbf{0}$, the solution takes the form

$$\begin{aligned} \mathbf{x}_n &= \mathbf{A}^{n-1} \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mathbf{A}^{n-3} \begin{bmatrix} \frac{F^2}{M} \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mathbf{A}^{n-4} \begin{bmatrix} \frac{F^3}{M} \\ 0 \\ 0 \\ 0 \end{bmatrix} \\ \mathbf{y}_n &= \mathbf{C} \mathbf{A}^{n-2} \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mathbf{C} \mathbf{A}^{n-4} \begin{bmatrix} \frac{F^2}{M} \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mathbf{C} \mathbf{A}^{n-5} \begin{bmatrix} \frac{F^3}{M} \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{aligned} \quad (20)$$

If the modulus of each of the eigenvalues of \mathbf{A} is less than unity, \mathbf{x}_n and \mathbf{y}_n in equation (20) will converge to zero as n increases to ∞ .

5.1.2 Step Signal

We assume a step signal⁴ for the first input variable while zero for the second one, i.e.,

$$C_{1,n} = \begin{cases} 0, & n < 0 \\ 1, & n \geq 0 \end{cases} \quad \text{and } C_{2,n} = 0 \quad \text{for } n = 1, 2, \dots$$

then,

$$\mathbf{u}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{u}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{u}_n = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{for } n \geq 3.$$

Thus $\mathbf{Bu}_0 = \mathbf{Bu}_1$, and $\mathbf{Bu}_n = \mathbf{0}$, $n \geq 3$. The latter equality holds as $M = F^2 + F^3$. The solution can be written as

$$\left. \begin{aligned} \mathbf{x}_n &= [\mathbf{A}^{n-1} + \mathbf{A}^{n-2}] \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mathbf{A}^{n-3} \begin{bmatrix} \frac{F^2}{M} - 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \\ \mathbf{y}_n &= \mathbf{C}[\mathbf{A}^{n-2} + \mathbf{A}^{n-3}] \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mathbf{CA}^{n-4} \begin{bmatrix} \frac{F^2}{M} - 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{aligned} \right\} \quad (21)$$

for $n = 5, 6, \dots$. If the modulus of each of the eigenvalues of \mathbf{A} is less than unity, \mathbf{x}_n and \mathbf{y}_n in equation (21) will asymptotically converge to zero as n increases.

5.1.3 Sine Signal

Let us consider the case where the input variable can be expressed as a sine signal. The assumption of a sine input signal may be more

⁴A step signal may be interpreted as a claim of size one occurring annually.

realistic in some cases as it may represent the underlying underwriting cycle that occurs in many insurance markets (Berger 1988).

For $n = 0, 1, \dots$, let $C_{1,n} = \sin(\omega_1 n + \phi_1)$ and $C_{2,n} = \sin(\omega_2 n + \phi_2)$ with $\omega_1 = \omega_2 = \pi$, $\phi_1 = -\frac{\pi}{6}$ and $\phi_2 = \frac{\pi}{2}$. This leads to the following:

n	0	1	2	3	4
$C_{1,n}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	\dots
$C_{2,n}$	1	-1	1	-1	\dots

Consequently the input vectors,

$$\mathbf{u}_0 = \begin{bmatrix} -0.5 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{u}_1 = \begin{bmatrix} 0.5 \\ -0.5 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{u}_2 = \begin{bmatrix} -0.5 \\ 0.5 \\ -0.5 \\ 0 \\ 1 \\ -1 \\ 1 \\ 0 \end{bmatrix} \text{ and}$$

$$\mathbf{u}_{2k+1} = \begin{bmatrix} 0.5 \\ -0.5 \\ 0.5 \\ -0.5 \\ -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}, \text{ and } \mathbf{u}_{2k+2} = \begin{bmatrix} -0.5 \\ 0.5 \\ -0.5 \\ 0.5 \\ 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$$

for $k = 1, 2, \dots$. It follows that

$$\mathbf{B}\mathbf{u}_0 = \begin{bmatrix} 0.5 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \mathbf{B}\mathbf{u}_1 = \begin{bmatrix} -0.5 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{B}\mathbf{u}_2 = \begin{bmatrix} 0.5 \left(1 - \frac{F^2}{M}\right) \\ 0 \\ -\left(1 - \frac{F^2}{M}\right) \\ 0 \end{bmatrix},$$

$$\mathbf{Bu}_{2k+1} = \begin{bmatrix} -0.5 \left(1 - \frac{F^2}{M} + \frac{F^3}{M} \right) \\ 0 \\ 1 - \frac{F^2}{M} + \frac{F^3}{M} \\ 0 \end{bmatrix}, \text{ and } \mathbf{Bu}_{2k+1} = \begin{bmatrix} 0.5 \left(1 - \frac{F^2}{M} + \frac{F^3}{M} \right) \\ 0 \\ - \left[1 - \frac{F^2}{M} + \frac{F^3}{M} \right] \\ 0 \end{bmatrix},$$

for $k = 1, 2, \dots$. As we observe for the vectors calculated before $\mathbf{Bu}_1 = -\mathbf{Bu}_0$ and $\mathbf{Bu}_{k+1} = -\mathbf{Bu}_k$ for $k = 2n + 1, n \geq 1$. Now assuming again the zero initial condition, i.e., $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{y}_0 = \mathbf{0}$, we obtain the general solution for the state of the system

$$\mathbf{x}_n = \mathbf{A}^{n-1}\mathbf{Bu}_0 - \mathbf{A}^{n-2}\mathbf{Bu}_0 + \mathbf{A}^{n-3}\mathbf{Bu}_2 + \mathbf{A}^{n-4}\mathbf{Bu}_3 - \mathbf{A}^{n-5}\mathbf{Bu}_3 + \dots \\ + (-1)^{n-1}\mathbf{A}\mathbf{Bu}_3 + (-1)^n\mathbf{Bu}_3.$$

Rearranging the terms of this equation we obtain

$$\mathbf{x}_n = \mathbf{A}^{n-2}(\mathbf{A} - \mathbf{I})\mathbf{Bu}_0 + \mathbf{A}^{n-3}\mathbf{Bu}_2 + (\mathbf{A}^{n-4} - \mathbf{A}^{n-3} + \dots \\ + (-1)^n\mathbf{I})\mathbf{Bu}_3.$$

It follows that

$$\mathbf{x}_n = \mathbf{A}^{n-2}(\mathbf{A} - \mathbf{I})\mathbf{Bu}_0 + \mathbf{A}^{n-3}\mathbf{Bu}_2 \\ + (\mathbf{I} + \mathbf{A}^2 + \dots + \mathbf{A}^{n-3})(\mathbf{A} - \mathbf{I})\mathbf{Bu}_3 \quad \text{if } n \text{ is odd;} \\ \mathbf{x}_n = \mathbf{A}^{n-2}(\mathbf{A} - \mathbf{I})\mathbf{Bu}_0 + \mathbf{A}^{n-3}\mathbf{Bu}_2 \\ + [(\mathbf{A} - \mathbf{I})(\mathbf{A} + \mathbf{A}^3 + \dots + \mathbf{A}^{n-3}) + \mathbf{I}]\mathbf{Bu}_3; \quad \text{if } n \text{ is even.}$$

If the modulus of the eigenvalues of \mathbf{A} are less than unity, then \mathbf{x}_n does not converge as n increases. In fact it fluctuates between two limits

$$\mathbf{x}_n \rightarrow \begin{cases} \mathbf{Q}(\mathbf{A} - \mathbf{I})\mathbf{Bu}_3 & \text{if } n = 2k + 1 \text{ and } k \rightarrow \infty, \\ [(\mathbf{A} - \mathbf{I})\mathbf{A}\mathbf{Q} + \mathbf{I}]\mathbf{Bu}_3 & \text{if } n = 2k + 1 \text{ and } k \rightarrow \infty, \end{cases}$$

where $\mathbf{Q} = \mathbf{I} + \mathbf{A}^2 + \mathbf{A}^4 + \dots$. It can be easily proved via the appropriate definition of a norm that as n goes to infinity the sequence of solutions \mathbf{x}_n is bounded.

5.2 Stability and Optimality

5.3 The Zero Stability Point

First we will establish that the system has only one equilibrium point at the zero point, $\mathbf{0}$.

$$\mathbf{A} = \left[\begin{array}{ccccc|cccc} R_1\lambda_{11} & 0 & \cdots & 0 & -e_1\varepsilon_1 & R_1\lambda_{21} & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 & & & & & \\ \vdots & \vdots & & \vdots & \vdots & & & & 0 & \\ 0 & 0 & \cdots & 0 & 0 & & & & & \\ 0 & 0 & \cdots & 1 & 0 & & & & & \\ \hline R_2\lambda_{12} & 0 & \cdots & 0 & 0 & R_2\lambda_{22} & 0 & \cdots & 0 & -e_2\varepsilon_2 \\ & & & & & 1 & 0 & \cdots & 0 & 0 \\ & & & & & \vdots & \vdots & & \vdots & \vdots \\ & & & 0 & & 0 & 0 & \cdots & 0 & 0 \\ & & & & & 0 & 0 & \cdots & 1 & 0 \end{array} \right]$$

$$= \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where $\mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{A}_{21}, \mathbf{A}_{22} \in \mathbb{R}^{(1+f) \times (1+f)}$ are defined in an obvious manner. Developing $\det(\mathbf{A})$ across the second row and again the minor across the second row continuously after f -steps we obtain

$$\det(\mathbf{A}) = (-1)^f \begin{vmatrix} -e_1\varepsilon_1 & 0 & 0 & 0 \\ 0 & & & \\ 0 & & \mathbf{A}_{22} & \\ \vdots & & & \\ 0 & & & \end{vmatrix}$$

or

$$\det(\mathbf{A}) = (-1)^{f+1} e_1\varepsilon_1 \det(\mathbf{A}_{22}).$$

Similarly, we develop $\det(\mathbf{A}_{22})$ and obtain:

$$\det(\mathbf{A}) = (-1)^{2(1+f)} e_1e_2\varepsilon_1\varepsilon_2 = e_1e_2\varepsilon_1\varepsilon_2.$$

5.3.1 Conditions on the Roots

To investigate the stability of the system, we focus on the pairs of roots (ρ_1, ρ_2) and (ρ_3, ρ_4) separately. There are two cases to consider for each pair of roots. As the analysis is similar for each pair, we provide a detailed treatment only for the pair (ρ_1, ρ_2) .

Case 1: The roots ρ_1, ρ_2 are real, i.e., $R^2 - 4e\varepsilon > 0$. Here we want $|\rho_1| < 1$ and $|\rho_2| < 1$. This implies that

$$\left| \frac{R \pm \sqrt{R^2 - 4e\varepsilon}}{2} \right| \leq 1,$$

i.e., $R < 1 + e\varepsilon$. Hence the pair of roots are real with absolute value less than one if and only if

$$4e\varepsilon < R^2 < (1 + e\varepsilon)^2. \quad (22)$$

Case 2: When ρ_1, ρ_2 are complex roots, i.e., $R^2 - 4e\varepsilon < 0$, the complex conjugate roots are

$$\rho_1, \rho_2 = \frac{R}{2} \pm i \frac{\sqrt{4e\varepsilon - R^2}}{2}$$

and consequently

$$|\rho_1|, |\rho_2| \leq 1 \Leftrightarrow \left[\left(\frac{R}{2} \right)^2 + \left(\frac{\sqrt{4e\varepsilon - R^2}}{2} \right)^2 \right]^{\frac{1}{2}} \leq 1 \Leftrightarrow$$

which implies that

$$\frac{R^2}{4} + \frac{4e\varepsilon - R^2}{4} \leq 1 \Leftrightarrow e\varepsilon \leq 1$$

and

$$\frac{R^2}{4} < e\varepsilon < 1. \quad (23)$$

For the second pair of roots ρ_3 and ρ_4 , we follow the same procedure and replace R with $R(1 - 2\lambda)$ to give

$$4e\varepsilon < (R(1 - 2\lambda))^2 < (1 + e\varepsilon)^2 \quad (24)$$

for real roots, and

$$\frac{R^2(1 - 2\lambda)^2}{4} < e\varepsilon < 1 \quad (25)$$

for complex roots.

5.3.2 Fastest Response Solution

Next we turn our attention to the determination of the optimal parameter values according to the fastest response criteria. Suppose we require a solution to the system such that the solution has no oscillations.

The speed at which the state variables respond to the different input signals depends on the maximum modulus of the eigenvalues of A : the smaller the maximum modulus of the eigenvalues, the faster the response. We note that the minimum value of maximum modulus of the eigenvalues is obtained when the quadratic polynomials have double roots, i.e., when

$$R^2 - 4e\varepsilon = 0 \text{ and } R^2(1 - 2\lambda)^2 - 4e\varepsilon = 0.$$

This may occur if and only if $\lambda = 0$ or $\lambda = 1$, i.e., when there is either no interaction or full interaction between the two insurance companies, and in either case we have a root of multiplicity four:

$$\rho_1 = \rho_2 = \rho_3 = \rho_4 = \frac{R}{2}. \quad (26)$$

In practical situations we cannot choose $\lambda = 0$ or $\lambda = 1$, yet we still have to minimize the maximum modulus of the roots. This means that we must choose which of the equations $R^2 - 4e\varepsilon = 0$ and $R^2(1 - 2\lambda)^2 - 4e\varepsilon = 0$ is more important and minimize the maximum modulus of the roots according to the chosen equation.

If we choose the equation $R^2 - 4e\varepsilon = 0$ then $\rho_1 = \rho_2 = R/2$ while ρ_3, ρ_4 are complex numbers such that $|\rho_3|, |\rho_4| < R/2$. The root with

the maximum absolute value is the real double root at $R/2$. As there are two complex roots, there will be oscillations in the solution.

The other option of choosing $R^2(1 - 2\lambda)^2 - 4e\varepsilon = 0$ produces four real roots, a double root

$$\rho_1, \rho_2 = \frac{R(1 - 2\lambda)}{2},$$

and two different roots

$$\rho_3, \rho_4 = \frac{R \pm \sqrt{R^2 - 4e\varepsilon}}{2}.$$

The root with the maximum absolute value is

$$\rho_3 = \frac{R + \sqrt{R^2 - 4e\varepsilon}}{2}.$$

As there are no complex roots, there will be no oscillations in the solution.

Thus we can conclude that:

1. The fastest response with oscillations occurs if we choose $R^2 - 4e\varepsilon = 0$. In this case the maximum modulus is $R/2$; and
2. The fastest response with no oscillations occurs if we choose $(1 - 2\lambda)^2 R^2 - 4e\varepsilon = 0$. In this case the maximum modulus is $(R + \sqrt{R^2 - 4e\varepsilon})/2 > R/2$.

Note that the overall fastest response with or without oscillations occurs when $R^2 - 4e\varepsilon = 0$, i.e., it yields oscillations.

A compromise is thus needed to reduce the oscillations to an acceptable level, but without unduly reducing the speed. The approach suggested is to choose the fastest overall response (i.e., $R^2 - 4e\varepsilon = 0$) and then choose λ to reduce the amplitude of the oscillations caused by the two complex conjugate roots

$$\begin{aligned} \rho_3, \rho_4 &= \frac{R}{2} \left[(1 - 2\lambda) \pm i\sqrt{4(\lambda - \lambda^2)} \right] \\ &= \frac{R}{2} (\cos \theta \pm i \sin \theta) \end{aligned} \quad (27)$$

where

$$\tan \theta = \frac{\sqrt{\lambda - \lambda^2}}{0.5 - \lambda}. \quad (28)$$

The choice of λ affects both the frequency and amplitude of the oscillations.

At this point we digress in order to discuss the importance of θ in connection with the general solution of the system. If z_1, z_2 are complex eigenvalues of matrix A of a dynamic system, then the general solution y_n contains a linear combination of the powers of z_1 and z_2 , i.e.,

$$y_n = \mu_1 z_1^n + \mu_2 z_2^n$$

where

$$\begin{aligned} \mu_1 &= a (\cos \beta + i \sin \beta) \\ z_1 &= z (\cos \theta + i \sin \theta) \\ \mu_2 &= a (\cos \beta - i \sin \beta) \\ z_2 &= z (\cos \theta - i \sin \theta), \end{aligned}$$

and

$$y_n = 2az^n \cos(n\theta + \beta).$$

5.4 Numerical Example

This example illustrates the methodology for the special case of two identical companies described in Section 5.1. The system of difference equations is:

$$\begin{aligned} S_{1,n} &= R(1 - \lambda)S_{1,n-1} + R\lambda S_{2,n-1} \\ &\quad + \frac{F^3}{M}C_{1,n-3} + \frac{F^2}{M}C_{1,n-2} - e\epsilon S_{1,n-2} - C_{1,n} \\ S_{2,n} &= R\lambda S_{1,n-1} + R(1 - \lambda)S_{2,n-1} \\ &\quad + \frac{F^3}{M}C_{2,n-3} + \frac{F^2}{M}C_{2,n-2} - e\epsilon S_{2,n-2} - C_{2,n} \\ P_{1,n} &= \frac{F^3}{Me}C_{1,n-3} + \frac{F^2}{Me}C_{1,n-2} - \epsilon S_{1,n-2} \\ P_{2,n} &= \frac{F^3}{Me}C_{2,n-3} + \frac{F^2}{Me}C_{2,n-2} - \epsilon S_{2,n-2} \end{aligned}$$

The parameters used are $R_1 = R_2 = R = 1.04$, $e_1 = e_2 = e = 0.8$, $F_1 = F_2 = F = 1$, and feedback factor $\varepsilon_1 = \varepsilon_2 = \varepsilon = 0.34$, in order to obtain the fastest response as indicated in Sections 5.1. The interaction parameter λ is allowed to vary and the time horizon is $n = 20$ years.

A spike input of 1 is used to model the claim variable of the first company. The mean and variance of the surplus variables of the two companies are calculated over the twenty years for several values of λ . The question of interest is: How does the system (i.e., the surplus variables) react to the occurrence of an unexpected claim in the first company?

Tables 1 and 2 show the development of $S_{1,n}$ and $S_{2,n}$. Observe that both surplus variables return to the stability point $\mathbf{0}$. Finally, we also observe that the summation of the variances in Tables 1 and 2 ($\text{Var}[S_{1,n}] + \text{Var}[S_{2,n}]$) is minimized for $\lambda = 0.8$. The last result means that for $\lambda = 0.8$ the system has the optimal behavior with regard to solvency requirements.

6 Summary and Conclusions

We construct an input/output control model of multinational pooling arrangements. A key aspect of these types of arrangements is the interaction of their premium, claims, and surplus processes. The objectives of this interaction are to:

- Smooth, as far as possible, the fluctuations of the surplus fund of each company participating in the pool; and to
- Spread each company's premium income and claims experiences to the block of the other companies.

The specific modeling also may be used generally for subsidiary insurance companies whose parent company wants to smooth the solvency requirement of each individual company. It can also be used for capital allocation between different lines of business.

We have derived several important results:

- Given $e_1 \neq 0$, $e_2 \neq 0$, ..., $e_m \neq 0$ and $\varepsilon_1 \neq 0$, $\varepsilon_2 \neq 0$, ..., $\varepsilon_m \neq 0$, which is normally the case in practice, the general model (for any value of f and m) has one equilibrium point, the zero point, and consequently one potential point of stability. If

Table 1
 $S_{1,n}$ for Various Values of λ

n	0	0.1	0.2	0.3	0.4
0	-1.0000	-1.0000	-1.0000	-1.0000	-1.0000
1	-1.0400	-0.9360	-0.8320	-0.7280	-0.6240
2	-0.3112	-0.1165	0.0349	0.1431	0.2080
3	0.4576	0.6238	0.6820	0.6593	0.5825
4	0.5600	0.5981	0.5113	0.3783	0.2553
5	0.4587	0.3872	0.2511	0.1538	0.1286
6	0.3256	0.2081	0.1139	0.1048	0.1485
7	0.2146	0.1023	0.0709	0.1036	0.1316
8	0.1351	0.0512	0.0581	0.0817	0.0765
9	0.0825	0.0290	0.0452	0.0481	0.0365
10	0.0493	0.0188	0.0297	0.0237	0.0212
11	0.0289	0.0130	0.0165	0.0122	0.0150
12	0.0168	0.0087	0.0083	0.0077	0.0094
13	0.0096	0.0055	0.0042	0.0051	0.0049
14	0.0055	0.0032	0.0024	0.0031	0.0025
15	0.0031	0.0018	0.0015	0.0016	0.0015
16	0.0017	0.0009	0.0009	0.0008	0.0009
17	0.0010	0.0005	0.0005	0.0004	0.0005
18	0.0005	0.0002	0.0003	0.0003	0.0003
19	0.0003	0.0001	0.0001	0.0002	0.0001
20	0.0002	0.0001	0.0001	0.0001	0.0001
$\mathbb{E}[S_{1,n}]$	0.0000	0.0000	0.0000	0.0000	0.0000
$\text{Var}[S_{1,n}]$	0.1546	0.1422	0.1254	0.1092	0.0951

Table 1 (continued)
 $S_{1,n}$ for Various Values of λ

n	0.5	0.6	0.7	0.8	0.9	1.0
0	-1.0000	-1.0000	-1.0000	-1.0000	-1.0000	-1.0000
1	-0.5200	-0.4160	-0.3120	-0.2080	-0.1040	0.0000
2	0.2296	0.2080	0.1431	0.0349	-0.1165	-0.3112
3	0.4788	0.3750	0.2983	0.2755	0.3338	0.5000
4	0.1759	0.1513	0.1703	0.1993	0.1821	0.0400
5	0.1617	0.2166	0.2562	0.2671	0.2824	0.4056
6	0.1910	0.2003	0.1813	0.1611	0.1451	0.0444
7	0.1256	0.1030	0.0924	0.0982	0.1077	0.1828
8	0.0600	0.0583	0.0688	0.0737	0.0721	0.0211
9	0.0363	0.0444	0.0448	0.0399	0.0374	0.0692
10	0.0267	0.0265	0.0228	0.0238	0.0266	0.0082
11	0.0158	0.0132	0.0143	0.0153	0.0139	0.0241
12	0.0078	0.0081	0.0090	0.0081	0.0084	0.0029
13	0.0044	0.0052	0.0046	0.0048	0.0050	0.0080
14	0.0029	0.0027	0.0027	0.0028	0.0026	0.0010
15	0.0016	0.0014	0.0016	0.0015	0.0016	0.0025
16	0.0008	0.0009	0.0009	0.0009	0.0008	0.0003
17	0.0005	0.0005	0.0005	0.0005	0.0005	0.0008
18	0.0003	0.0003	0.0003	0.0003	0.0003	0.0001
19	0.0002	0.0001	0.0001	0.0002	0.0001	0.0002
20	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
$\mathbb{E} [S_{1,n}]$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\text{Var} [S_{1,n}]$	0.0834	0.0742	0.0675	0.0637	0.0644	0.0777

Table 2
 $S_{2,n}$ for Various Values of λ

n	0	0.1	0.2	0.3	0.4
1	0.0000	-0.1040	-0.2080	-0.3120	-0.4160
2	0.0000	-0.1947	-0.3461	-0.4543	-0.5192
3	0.0000	-0.1662	-0.2245	-0.2017	-0.1250
4	0.0000	-0.0381	0.0487	0.1817	0.3047
5	0.0000	0.0715	0.2076	0.3048	0.3301
6	0.0000	0.1175	0.2118	0.2208	0.1771
7	0.0000	0.1123	0.1437	0.1110	0.0830
8	0.0000	0.0840	0.0771	0.0534	0.0587
9	0.0000	0.0536	0.0373	0.0344	0.0460
10	0.0000	0.0304	0.0196	0.0256	0.0280
11	0.0000	0.0160	0.0124	0.0167	0.0139
12	0.0000	0.0081	0.0084	0.0091	0.0073
13	0.0000	0.0041	0.0054	0.0045	0.0047
14	0.0000	0.0023	0.0031	0.0024	0.0030
15	0.0000	0.0013	0.0016	0.0015	0.0016
16	0.0000	0.0008	0.0008	0.0009	0.0008
17	0.0000	0.0005	0.0004	0.0005	0.0005
18	0.0000	0.0003	0.0003	0.0003	0.0003
19	0.0000	0.0002	0.0002	0.0001	0.0002
20	0.0000	0.0001	0.0001	0.0001	0.0001
$E[S_{2,n}]$	0.0000	0.0000	0.0000	0.0000	0.0000
$\text{Var}[S_{2,n}]$	0.0000	0.0060	0.0166	0.0268	0.0352

Table 2 (continued)
 $S_{2,n}$ for Various Values of λ

n	0.5	0.6	0.7	0.8	0.9	1.0
1	-0.5200	-0.6240	-0.7280	-0.8320	-0.9360	-1.0400
2	-0.5408	-0.5192	-0.4543	-0.3461	-0.1947	0.0000
3	-0.0212	0.0825	0.1593	0.1820	0.1238	-0.0424
4	0.3842	0.4087	0.3897	0.3607	0.3779	0.5200
5	0.2969	0.2421	0.2025	0.1916	0.1763	0.0531
6	0.1346	0.1253	0.1443	0.1645	0.1805	0.2812
7	0.0890	0.1116	0.1223	0.1164	0.1069	0.0318
8	0.0752	0.0768	0.0664	0.0614	0.0631	0.1141
9	0.0462	0.0382	0.0377	0.0426	0.0451	0.0133
10	0.0226	0.0228	0.0264	0.0254	0.0227	0.0411
11	0.0131	0.0157	0.0147	0.0136	0.0151	0.0049
12	0.0089	0.0086	0.0078	0.0087	0.0084	0.0139
13	0.0052	0.0044	0.0050	0.0048	0.0046	0.0017
14	0.0026	0.0027	0.0028	0.0026	0.0029	0.0045
15	0.0014	0.0016	0.0015	0.0016	0.0015	0.0005
16	0.0009	0.0008	0.0009	0.0009	0.0009	0.0014
17	0.0005	0.0005	0.0005	0.0005	0.0005	0.0002
18	0.0003	0.0003	0.0003	0.0003	0.0003	0.0004
19	0.0001	0.0002	0.0001	0.0001	0.0002	0.0001
20	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
$E[S_{2,n}]$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$Var[S_{2,n}]$	0.0417	0.0464	0.0499	0.0530	0.0577	0.0726

$$\prod_{r=1}^m e_r \varepsilon_r > 1$$

then the system is unstable regardless of the other parameter values, i.e., the surplus and premium levels fluctuate, with the surplus diverging to infinity.

- For the special case of two identical companies with time delay of one year $f = 1$ the exact condition for stability (assuming typical values for R and λ , i.e., $0 < R < 2$ and $0 < \lambda < 1$, and considering equations (22) through (25)) is $R - 1 < e\varepsilon < 1$.
- For the case of the two identical companies ($m = 2$), we show that the ultimate surplus level converges to zero under each of the spike and step input signals. This is a highly desirable result because it means that the system reacts properly and returns to its initial state. For the sine signal we show that the ultimate surplus fund fluctuates between two levels.
- For the special case ($m = 2$), full investigation has been done with respect to the fastest response and oscillatory form of the solution. It has been shown that the fastest response is obtained when $\varepsilon^* = \frac{R^2}{4e}$.
- For the special case of the two identical companies and considering the optimal choice for the feedback factor ε^* , we also have shown that amplitude and frequency of the oscillations depend on the interaction factor λ .

References

- Balzer L.A. and Benjamin S. "Dynamic Response of Insurance Systems With Delayed Profit/Loss Sharing Feedback to Isolated Unpredicted Claims." *Journal of the Institute of Actuaries* 107 (1980): 513-528.
- Balzer L.A. "Control of Insurance Systems With Delayed Profit/Loss Sharing Feedback and Persisting Unpredicted Claims." *Journal of the Institute of Actuaries* 109 (1982): 285-316.
- Benjamin S. "An Actuarial Layman Looks at Control Theory." *Transactions of the 22nd International Congress of Actuaries* 4 (1984): 295-310.

- Berger, L.A. "A Model of the Underwriting Cycle in Property/Liability Insurance Industry." *Journal of Risk and Insurance* 55 (1988): 298-306.
- Cadzow, J.A. *Discrete-Time Systems*. Englewood Cliffs, N.J.: Prentice-Hall, Inc, 1973.
- Chang, S-C. "Realistic Pension Funding: A Stochastic Approach." *Journal of Actuarial Practice* 8 (2000): 5-42.
- Hart, D.G., Buchanan, R.A. and Houre, B.A. *The Actuarial Practice of General Insurance*. Sydney, Australia: Institute of Actuaries of Australia: 1996.
- Healy, M.J.R. *Matrices for Statistics*. Oxford, England: Oxford University Press, 1995.
- Loades. D.H. "Elementary Engineering Control and Pension Funding." *Transactions of the 26th International Congress of Actuaries* 5 (1998): 239-266.
- Martin-Löf, A. "Premium Control in an Insurance System, an Approach Using Linear Control Theory." *Scandinavian Actuarial Journal* (1983): 1-27
- Martin-Löf, A. "Lectures on the Use of Control Theory in Insurance." *Scandinavian Actuarial Journal* (1994): 1-25
- Runggaldier, W.J. "Concept and Methods for Discrete and Continuous Time Control under Uncertainty." *Insurance: Mathematics and Economics* 22 (1998): 25-39.
- Schäl, M. "On Piecewise Deterministic Markov Control Processes: Control of Jumps and of Risk Processes in Insurance." *Insurance: Mathematics and Economics* 22 (1998): 75-91.
- Vandebroek M. and Dhaene, J. "Optimal Premium Control in a Non-life Insurance Business." *Scandinavian Actuarial Journal* (1995): 3-13
- William M. Mercer International. *Multinational Pooling. The Networks*. London, England: William M. Mercer International, 1998.
- Zimbidis, A. and Haberman, S. "The Combined Effect of Delay and Feedback in the Insurance Pricing Process." *Insurance Mathematics and Economics* 28 (2001): 263-280.

Appendix: The Characteristic Polynomial $\phi_m(\rho)$

First we calculate the form of the $\phi_2(\rho)$ then we may generalize our result for any value of m .

$$\phi_2(\rho) = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} = 0$$

where

$$\mathbf{H}_{11} = \begin{bmatrix} \rho - R_1\lambda_{11} & 0 & \cdots & 0 & e_1\varepsilon_1 \\ -1 & \rho & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & \rho & 0 \\ 0 & 0 & \cdots & -1 & \rho \end{bmatrix}$$

$$\mathbf{H}_{12} = \begin{bmatrix} -R_1\lambda_{21} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$\mathbf{H}_{21} = \begin{bmatrix} -R_2\lambda_{12} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$\mathbf{H}_{22} = \begin{bmatrix} \rho - R_2\lambda_{22} & 0 & \cdots & 0 & e_2\varepsilon_2 \\ -1 & \rho & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & \rho & 0 \\ 0 & 0 & \cdots & -1 & \rho \end{bmatrix}.$$

The analytical development of $\phi_2(\rho)$ is difficult, but we can find the final form if we follow a simple rule and a recursive procedure. The simple rule is to develop the major determinant $\phi_2(\rho)$ or the minor ones (which are produced by deleting rows and columns) across the (first) row or column that has the greatest number of zero elements. The recursive procedure is described with the following steps.

Step 1: Develop the $\phi_2(\rho)$ across the second row that has only two non-zero elements the -1 and ρ ,

$$\phi_2(\rho) = (-1)(-1)\Xi_1^{(2)} + \rho\Psi_1^{(2)}$$

where $\Xi_1^{(2)}$ is the minor determinant of $\phi_2(\rho)$, produced by deleting the first column and the second row of $\phi_2(\rho)$ and $\Psi_1^{(2)}$ is the minor determinant of $\phi_2(\rho)$, produced by deleting the second column and the second row of $\phi_2(\rho)$. [The (2) superscript of Ξ_1 and Ψ_1 refers to the case $m = 2$.]

Step 2: Develop the minor determinant $\Xi_1^{(2)}$ across the first column (having one non-zero element the -1).

$$\Xi_1^{(2)} = (-1)(-1)\Xi_2^{(2)}.$$

Step 3: Continue the development of $\Xi_i^{(2)}$

$$\Xi_i^{(2)} = (-1)(-1)\Xi_{i+1}^{(2)}$$

for $i = 2, 3, \dots, f-1$ with Ξ_i being the minor determinant of Ξ_{i-1} , produced by deleting the first column and second row of Ξ_i and

$$\Xi_f^{(2)} = \begin{vmatrix} e_1\varepsilon_1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \phi_1(\rho) & \\ 0 & & & \end{vmatrix}$$

or

$$\Xi_f^{(2)} = e_1\varepsilon_1\phi_1(\rho).$$

Step 4: Combine equations in Steps 1, 2, and 3 to obtain

$$\phi_2(\rho) = e_1\varepsilon_1\phi_1(\rho) + \rho\Psi_1^{(2)}.$$

Step 5: Develop the minor determinant $\Psi_1^{(2)}$ across the second row (having only one non-zero element, ρ).

$$\Psi_1^{(2)} = \rho \Psi_2^{(2)}.$$

Step 6: Continue the development of the determinants across the second row (similarly with $\Psi_1^{(2)}$)

$$\Psi_i^{(2)} = \rho \Psi_{i+1}^{(2)}$$

for $i = 2, 3, \dots, f-1$, where $\Psi_i^{(2)}$ is the minor determinant of $\Psi_{i-1}^{(2)}$, produced by deleting the second column and the second row and

$$\Psi_f^{(2)} = \begin{vmatrix} \rho - R_1 \lambda_{11} & -R_1 \lambda_{21} & 0 & \cdots & 0 \\ -R_2 \lambda_{12} & & & & \\ 0 & & \phi_1(\rho) & & \\ \vdots & & & & \\ 0 & & & & \end{vmatrix}.$$

Step 7: Combining from Steps 5 and 6, we obtain

$$\phi_2(\rho) = e_1 \varepsilon_1 \phi_1(\rho) + \rho^f \Psi_f^{(2)}.$$

Step 8: Develop $\Psi_f^{(2)}$ across the third row that has two non-zero elements -1 and ρ .

$$\Psi_f^{(2)} = (-1)^{(2)} \begin{vmatrix} \rho - R_1\lambda_{11} & -R_1\lambda_{21} & 0 & \cdots & 0 \\ -R_2\lambda_{12} & & & & \\ 0 & & \Xi_1^{(1)} & & \\ \vdots & & & & \\ 0 & & & & \end{vmatrix} \\ + \rho \begin{vmatrix} \rho - R_1\lambda_{11} & -R_1\lambda_{21} & 0 & \cdots & 0 \\ -R_2\lambda_{12} & & & & \\ 0 & & \Psi_1^{(1)} & & \\ \vdots & & & & \\ 0 & & & & \end{vmatrix}$$

(The (1) superscripts of Ξ_1 and Ψ_1 refers to $\phi_1(\rho)$.) $\Xi_1^{(1)}$ and $\Psi_1^{(1)}$ are produced similarly to $\Xi_1^{(2)}$ and $\Psi_1^{(2)}$ from $\phi_2(\rho)$. So we follow similar steps and finally,

$$\Psi_f^{(2)} = \begin{vmatrix} \rho - R_1\lambda_{11} & 0 & 0 \\ -R_2\lambda_{12} & 0 & e_2\varepsilon_2 \\ 0 & -1 & \rho \end{vmatrix} \\ + \rho^{f-1} \begin{vmatrix} \rho - R_1\lambda_{11} & -R_1\lambda_{21} & 0 \\ -R_2\lambda_{12} & \rho - R_2\lambda_{22} & e_2\varepsilon_2 \\ 0 & 0 & \rho \end{vmatrix}$$

which implies

$$\Psi_f^{(2)} = e_2\varepsilon_2 (\rho - R_1\lambda_{11}) \\ + \rho^f [(\rho - R_1\lambda_{11})(\rho - R_2\lambda_{22}) - R_1R_2\lambda_{12}\lambda_{21}].$$

Step 9: Develop $\phi_1(\rho)$ similarly with $\phi_2(\rho)$ and obtain

$$\phi_1(\rho) = e_2\varepsilon_2 + \rho^f (\rho - R_2\lambda_{22}).$$

Step 10: Combining the equations from Steps 7, 8, and 9 so that we finally obtain the following equation

$$\begin{aligned} \phi_2(\rho) = & e_1 e_2 \varepsilon_1 \varepsilon_2 + e_1 \varepsilon_1 \rho^f (\rho - R_2 \lambda_{22}) + \rho^f e_2 \varepsilon_2 (\rho - R_1 \lambda_{11}) \\ & + \rho^{2f} (\rho - R_1 \lambda_{11}) (\rho - R_2 \lambda_{22}) - \rho^{2f} R_1 R_2 \lambda_{12} \lambda_{21}. \end{aligned} \quad (29)$$

We observe that $\phi_2(\rho)$ is a polynomial with $a_{2(1+f)} = 1$ (coefficient of $\rho^{2(1+f)}$) and $a_0 = e_1 e_2 \varepsilon_1 \varepsilon_2$ (constant term). It is straightforward to generalize the above procedure and obtain from the equation that appears in Step 4:

$$\phi_m(\rho) = e_1 \varepsilon_1 \phi_{m-1}(\rho) + \rho \Psi_1^m.$$

Finally, $\phi_m(\rho)$ is a polynomial with $a_{m(1+f)} = 1$ (coefficient of $\rho_{m(1+f)}$) and $a_0 = e_1 e_2 \dots e_m \varepsilon_1 \varepsilon_2 \dots \varepsilon_m$ (constant term).

A Sensitivity Analysis of the Premiums for a Permanent Health Insurance (PHI) Model

Ben D. Rickayzen*

Abstract[†]

This paper presents an analysis of the parameters used in a multi-state model for permanent health insurance (PHI). The model is a simplification of that used in the United Kingdom. To avoid using duration dependent probabilities, the model splits the sick state into several sub-states to act as a proxy for duration spent in a particular state. This enables a Markov approach to be adopted. Lapses are incorporated within the model, and the net premium for a particular policy is tested for sensitivity to the various parameters used, including their interaction with the lapse rate. One of our conclusions is that the net premium is insensitive to changes in the lapse rate.

Key words and phrases: *PHI benefits, force of transition, Markov chain, lapses*

*Ben Rickayzen, B.Sc., F.I.A., is a senior lecturer and deputy head of the Department of Actuarial Science and Statistics at City University, England. He graduated with a first class honours degree in Mathematics from the University of Nottingham in 1984. He then joined the actuarial consultants, Bacon and Woodrow, and worked in their pensions consultancy until 1994, qualifying as a Fellow of the Institute of Actuaries in 1990. He joined City University as a lecturer in 1994. His main research interest is in health insurance and, in particular, income protection (formerly known as permanent health insurance) and long-term care for the elderly. He is currently a member of the Institute and Faculty of Actuaries' Health and Care ECPD Sub-Committee.

Mr. Rickayzen's address is: Department of Actuarial Science and Statistics, City University, Northampton Square, London EC1V 0HB, UNITED KINGDOM. Internet address: b.d.rickayzen@city.ac.uk

[†]The author would like to thank the reviewers and colleagues who read and commented on early versions of this paper and R.H. Plumb for comments on the history of PHI in the U.K.

1 Introduction

1.1 Overview of the U.K. PHI Business

Permanent health insurance (PHI) has been written in the U.K. for over 100 years. The business was a natural extension of the fraternal (Friendly Society) weekly sickness benefit paid to its members. The rise of the welfare state in the early part of the twentieth century saw the state assume some of the responsibilities of the fraternal societies. Consequently, the amount of business written by private insurers was limited.

The PHI business has increased since World War II, with individual and group business being written by a number of insurers. The market consists of a few specialist direct insurers and reinsurers to support their operations.

The U.K. government still provides a small long-term disability benefit. Recovery rates of state claimants are low; the benefit is a substitute for unemployment benefits. Anyone earning more than national average earnings needs to insure, but there is considerable underinsurance. Increasingly the PHI business is being referred to as *income protection insurance*.

PHI benefits are built around the U.K. pension system and are often expressed in amounts per week or per month. These benefits cease at state pension age, which is currently age 65 for males and age 60 for females. Some limited benefit period business also is written.

The contracts are similar to those issued in North America, but the terminology differs. For example, the *elimination period* is referred to as the *deferred period* in the U.K. There are similar exclusions, but benefits are paid in full for behavioral health problems. In addition, benefits are paid whether the cause of disability is due to an accident or to sickness. The major change in the last 20 years has been the switch in individual business from non-cancelable individual business to guaranteed renewable.

The primary difference between group and individual PHI business is the impact of tax on premiums and benefits paid. Under group business, the employer generally pays the premiums, which are tax deductible, and the benefits paid to the employees are taxed as salary. Under the individual business, there is no tax relief on the premiums paid, but the benefits are paid free of income tax. Waiver of premium is included as a benefit provision. The most common deferred periods are one week, four weeks, 13 weeks, 26 weeks, and 52 weeks.

Benefit limitations apply related to pre-disability income. Benefits from all sources are taken into account, including other group and individual insurances and pensions received. Various disability definitions are offered, including inability to follow any occupation.

1.2 Objectives

The objective of this paper is to introduce a practical mathematical model of a U.K. style PHI system. Specifically, the PHI system is modeled using a multi-state process in which, as a healthy individual ages, he or she may become sick then recover, become sick again, etc., until death.¹ Thus the individual's health fluctuates between two states (sickness and health) until death. If healthy, sick, and dead are viewed as separate states, the probability that a policyholder moves from the sick state to the dead state or to the healthy state depends on the time spent in the sick state. In other words, the transition probabilities depend on duration in a particular state as well as the age of the policyholder.

It is possible to incorporate the duration-dependence aspect in the model, which leads to a much more complicated model. This is the approach used in the 1991 *Continuous Mortality Investigation Report No. 12* (CMIR 12). To obtain numerical values for the transition forces within the PHI model, CMIR 12 splits the sick states into 781 sub-states, each relating to a different duration of sickness. CMIR 12 then calculates probabilities at every $1/156^{\text{th}}$ of a year of age for duration of sickness up to 5 years in all (making 780 sub-states) and all sickness periods beyond 5 years are aggregated. CMIR 12 (Part D) shows how it is possible to obtain numerical values for probabilities, annuities, etc. Clearly, CMIR 12 provides a thorough and complex model.

The approach taken in this paper is to develop a simpler model, one with only three (healthy, sick, and dead) states, then split the sick state into a small number of sub-states. We adopt the approach based on Jones (1994). Though the CMIR 12 technique of splitting the sick states into sub-states pre-dates Jones, Jones' approach is simpler because it uses constant forces of transition assumption for transition from state to state. This maintains the Markov property of the model. Increasing the number of states makes the state space more complicated, but maintaining the Markov process keeps the calculations tractable.

One advantage of using the simpler model described in this paper is that it can easily be used by actuaries who do not have access to complex models such as CMIR 12 or the detailed data required to use such

¹For a detailed discussion on the use of multi-state models in disability insurance, see, for example, Haberman and Pitacco (1999).

models. It also can be used as an initial practical model for actuaries who are interested in rough estimates for net premiums for PHI models.

The paper is organized as follows: Section 2 introduces the model of the various transition probabilities. Expressions are derived for the transition probabilities required to obtain actuarial present values. Section 3 explains the connection between the parameters used in the model and those that are derived using data contained in CMIR 12. The data contained in CMIR 12 are used to test the sensitivity of the net premium to some of the parameters involved in the transition probabilities. Section 4 describes the results, while Section 5 provides a summary and conclusions.

2 The Model

2.1 The States and Transition Probabilities

The PHI model has six states labeled one to six.

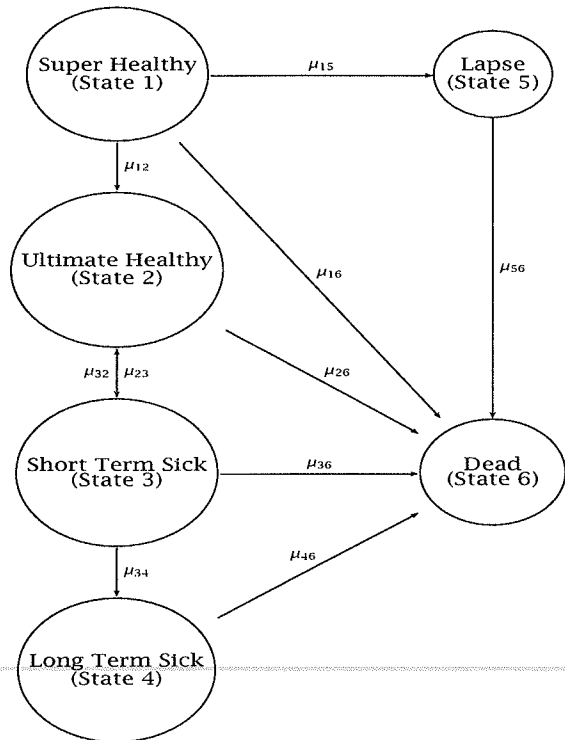
- State 1 (Super Healthy): This is the state in which new policyholders enter the model when their policy commences. Because they have provided satisfactory medical evidence, new policyholders are deemed to be select lives and therefore healthier than other insured lives of the same age. We describe these lives as *super healthy*.
- State 2 (Ultimate Healthy): It is likely that, in time, the selection effect will disappear and that the super healthy lives will move to the ultimate form of the healthy state from which they may become sick enough to make a claim under the PHI policy.
- State 3 (Short-Term Sick): It is possible to recover from the short-term sick state 3 and, therefore, to return to state 2.
- State 4 (Long-Term Sick): It is not possible to recover from the long-term sick state. Death is the only mode of exit from this state.
- State 5 (Lapse): We assume that only super healthy policyholders will lapse their policy because policyholders in any other state would find it worthwhile to continue their PHI policy.
- State 6: Death.

A diagrammatic representation of the multi-state model adopted in this paper is displayed in Figure 1.

It is possible to introduce more sickness states as a proxy to a greater number of durations of sickness. This has not been done, however, because it is difficult to choose parameter values for the transition forces between the different sick states. In addition, having more states would increase the computational problems, albeit not insurmountably.

The forces of transition between states in PHI are continuous functions that depend on many factors including such factors as age, sex, income, and the time spent in a state. Though the exact mathematical form of these functions is unknown, we are sure that they are not constant.

Figure 1
Outline of PHI Model



Due to the mathematical difficulties inherent in using continuously varying forces, however, we will adopt the general methodology described in Jones (1994), i.e., we assume that the forces of transition are piecewise constant over each age interval instead.

Suppose there are n states labeled $1, 2, \dots, n$. Let $\mu_{ij}(x+t)$ denote the force of transition from state i to state j at age $x+t$, for $i, j = 1, 2, 3, \dots, n$, $x = 0, 1, 2, \dots$, and $0 \leq t \leq 1$. If state j is not linked directly to state i then $\mu_{ij}(x+t) \equiv 0$. It is convenient also to define, for each i ,

$$\mu_{ii}(x+t) = - \sum_{\substack{j=1 \\ j \neq i}}^n \mu_{ij}(x+t), \quad (1)$$

where $i = 1, 2, 3, \dots, n$, $x = 0, 1, 2, \dots$, and $0 \leq t \leq 1$.

The piecewise constant force of transition implies that

$$\mu_{ij}(x+t) = \mu_{ij}(x) \quad \text{for } x = 0, 1, 2, \dots \text{ and } 0 \leq t < 1. \quad (2)$$

One implication of the piecewise constant transition intensities assumption is that the length of time already spent in the current state has no effect on the future length of time that the policyholder will remain in the state, i.e., a memoryless property exists. [See Haberman (1992) for more on the memoryless property of multi-state processes with constant transition intensities.]

Next, let $p_{ij}(t, x)$ be the probability that a life currently exact age x in state i will be in state j in t years time. The common approach² to deriving an expression for $p_{ij}(t, x)$ is to use the Chapman-Kolmogorov backward system of difference-differential equations as contained in Cox and Miller (1965, Chapter 4). The backward system of equations is derived by considering the interval $(0, t+h]$ as comprising subintervals $(0, h]$ and $(h, t+h]$ and letting $h \rightarrow 0$.

$$\frac{d}{dt} p_{ij}(t, x) = \sum_{k=1}^n \mu_{ik}(x) p_{kj}(t, x) \quad (3)$$

for $i, j = 1, \dots, n$, $x = 0, 1, \dots$, and $0 \leq t \leq 1$. These equations lead to a set of difference-differential equations. For illustration purposes, some of the differential equations are presented below:

²See, for example, Ramsay (1989), Jones (1994), and Haberman (1995).

$$\left. \begin{aligned}
 \frac{d}{dt} p_{11}(t) &= -(\mu_{12} + \mu_{15} + \mu_{16}) p_{11}(t) \\
 \frac{d}{dt} p_{12}(t) &= -(\mu_{12} + \mu_{15} + \mu_{16}) p_{12}(t) + \mu_{12} p_{22}(t) \\
 \frac{d}{dt} p_{22}(t) &= -(\mu_{23} + \mu_{26}) p_{22}(t) + \mu_{23} p_{32}(t) \\
 \frac{d}{dt} p_{23}(t) &= -(\mu_{23} + \mu_{26}) p_{23}(t) + \mu_{23} p_{33}(t) \\
 \frac{d}{dt} p_{33}(t) &= \mu_{32} p_{23}(t) - (\mu_{32} + \mu_{34} + \mu_{36}) p_{33}(t) \\
 \frac{d}{dt} p_{32}(t) &= \mu_{32} p_{22}(t) - (\mu_{32} + \mu_{34} + \mu_{36}) p_{32}(t) \\
 \frac{d}{dt} p_{44}(t) &= -\mu_{46} p_{44}(t) \\
 \frac{d}{dt} p_{46}(t) &= -\mu_{46} p_{46}(t) + \mu_{46} p_{66}(t) \\
 \frac{d}{dt} p_{55}(t) &= -\mu_{56} p_{55}(t) \\
 \frac{d}{dt} p_{56}(t) &= -\mu_{56} p_{56}(t) + \mu_{56} p_{66}(t) \\
 \frac{d}{dt} p_{66}(t) &= 0
 \end{aligned} \right\} \quad (4)$$

The easiest way to solve the system of differential equations given in equation (3) is to follow the method outlined by Cox and Miller (1965), which involves matrix manipulation. First define the following $n \times n$ matrices

$$\begin{aligned}
 \mathbf{M}(x) &= \{\mu_{ij}(x)\}_{i,j=1}^n = \text{The forces of transition matrix;} \\
 \mathbf{P}(t, x) &= \{p_{ij}(t, x)\}_{i,j=1}^n = \text{The transition probability matrix; and} \\
 \mathbf{P}'(t, x) &= \left\{ \frac{d}{dt} p_{ij}(t, x) \right\}_{i,j=1}^n.
 \end{aligned}$$

The Chapman-Kolmogorov backward system of equations may be written as

$$\mathbf{P}'(t, x) = \mathbf{M}(x)\mathbf{P}(t, x) \tag{5}$$

for $x = 0, 1, \dots$, and $0 \leq t \leq 1$, with boundary condition $\mathbf{P}(0, x) = \mathbf{I}$ (where \mathbf{I} is the identity matrix).

It is easily seen that equation (5) has the solution

$$\mathbf{P}(t, x) = e^{t\mathbf{M}(x)} = \mathbf{I} + \sum_{k=1}^{\infty} \frac{t^k}{k!} (\mathbf{M}(x))^k. \tag{6}$$

If it is known that $\mathbf{M}(x)$ has distinct eigenvalues $d_1(x), d_2(x), \dots, d_n(x)$, then

$$\mathbf{M}(x) = \mathbf{A}(x)\mathbf{D}(x)\mathbf{A}(x)^{-1} \quad (7)$$

where \mathbf{D} is the diagonal matrix

$$\mathbf{D} = \text{diag}(d_1(x), d_2(x), \dots, d_n(x))$$

and the i^{th} column of $\mathbf{A}(x)$ is the right-eigenvector associated with $d_i(x)$ (Cox and Miller 1965, Chapter 4.5). Equations (6) and (7) lead to the following expression for $\mathbf{P}(t, x)$:

$$\mathbf{P}(t, x) = \mathbf{A}(x)\text{diag}(e^{td_1(x)}, \dots, e^{td_n(x)})\mathbf{A}(x)^{-1}. \quad (8)$$

In this paper equation (8) is used to compute $\mathbf{P}(t, x)$.

Once $\mathbf{P}(t, x)$ is known for $x = 0, 1, \dots$, and $0 \leq t \leq 1$, we must develop an expression to compute $p_{ij}(t, x)$ for $x = 0, 1, \dots$, and $t > 1$. Suppose $t = k + s$ where $k = 1, 2, \dots$ and $0 \leq s < 1$. It follows that

$$\mathbf{P}(k + s, x) = \left(\prod_{r=1}^k \mathbf{P}(1, x + r - 1) \right) \mathbf{P}(s, x + k). \quad (9)$$

Next, as premiums and benefits are paid m times per year, we need expressions for transition probabilities at m^{thly} intervals. Consider the form of $p_{ij}(1/m, x + h/m)$ where $h = 0, 1, \dots, m - 1$. Under the piecewise constant assumption of equation (2) $p_{ij}(1/m, x + h/m)$ is independent of h for $h = 0, 1, \dots, m - 1$. Let us define $y_{ij}^{(m)}(x)$ as

$$y_{ij}^{(m)}(x) = p_{ij}\left(\frac{1}{m}, x + \frac{h}{m}\right). \quad (10)$$

In other words, $y_{ij}^{(m)}(x)$ is the probability that a person currently age $x + h/m$ and in state i will be in state j at age $x + (h + 1)/m$ where $h = 0, 1, \dots, m - 1$. We now define the $n \times n$ matrix

$$\mathbf{\Gamma}_x^{(m)} = \{y_{ij}^{(m)}(x)\}_{i,j=1}^n. \quad (11)$$

It follows that, for $t = k + h/m$, $k = 0, 1, \dots$, and $h = 0, 1, \dots, m - 1$,

$$P\left(k + \frac{h}{m}, x\right) = \left(\prod_{r=1}^k (\Gamma_{x+r-1}^{(m)})^m\right) (\Gamma_{x+k}^{(m)})^h \quad (12)$$

and $p_{ij}(t, x)$ can be determined. There is no real advantage to using equation (12) over equation (9) except when m is large. If m is large, say $m = 52$ (i.e., weekly payments), we can approximate $y_{ij}^{(m)}(x)$ as follows:

$$y_{ij}^{(m)}(x) = \begin{cases} \frac{1}{m} \mu_{ij}(x) & \text{if } i \neq j; \\ 1 + \frac{1}{m} \mu_{ii}(x) & \text{if } i = j. \end{cases} \quad (13)$$

2.2 Determination of the Net Premium

Premiums are assumed to be payable weekly in advance. A premium is only payable if the policyholder is either in state 1 (super healthy) or state 2 (ultimate healthy) at the start of the week in the policy year under consideration if premiums are waived during periods of sickness.

The annual net premium P is determined by equating the actuarial (expected) present value of future net premiums and the actuarial (expected) present value of future benefits at policy inception. To determine the net premium we need an expression for an m^{thly} annuity due payable for z years whenever x is in state j , which is:

$${}_{ij}\ddot{a}_{x:\overline{z}|}^{(m)} = \frac{1}{m} \sum_{r=0}^{zm-1} v^{r/m} p_{ij}\left(\frac{r}{m}, x\right) \quad (14)$$

and an expression for an m^{thly} annuity immediate payable for z years whenever x is in state j , which is:

$${}_{ij}a_{x:\overline{z}|}^{(m)} = \frac{1}{m} \sum_{r=1}^{zm} v^{r/m} p_{ij}\left(\frac{r}{m}, x\right). \quad (15)$$

It follows that the actuarial present value (APV) of the future premium is

$$\text{APV of Future Premiums} = P \left({}_{11}\ddot{a}_{x:\overline{z}|}^{(m)} + {}_{12}\ddot{a}_{x:\overline{z}|}^{(m)} \right).$$

The PHI benefit is assumed to be paid weekly during periods of sickness at the rate of $\$B$ per year. The PHI benefit is only payable if the policyholder is in either state 3 (short-term sick) or state 4 (long-term sick) at the end of the week in the policy year under consideration. Hence, the actuarial present value of the PHI benefits is

$$\text{APV of Future Benefits} = B \left({}_{13}a_{x:\overline{z}|}^{(m)} + {}_{14}a_{x:\overline{z}|}^{(m)} \right).$$

Therefore, we can find P from

$$P = \frac{B \left({}_{13}a_{x:\overline{z}|}^{(m)} + {}_{14}a_{x:\overline{z}|}^{(m)} \right)}{\left({}_{11}\ddot{a}_{x:\overline{z}|}^{(m)} + {}_{12}\ddot{a}_{x:\overline{z}|}^{(m)} \right)}. \quad (16)$$

3 PHI Data and Parameter Values

The parameter values used in this model have been influenced by the data contained in CMIR 12. As the data used in CMIR 12 are somewhat outdated, it is not necessary to input into our model precisely the output values emanating from CMIR 12.³ Therefore CMIR 12 is simply used as a guide to choosing parameter values for this paper.

For convenience the ages are grouped into 5-year age bands with the forces of transition assumed to be constant over each 5-year age band. The age bands are 30-34, 35-39, ..., 60-64. Next we describe the way in which each parameter value has been chosen.

$\mu_{23}(x)$ (Unstable Healthy \rightarrow Short-Term Sick): This parameter is based on the sickness inception rate, σ_x , described in Part C of CMIR 12. We use the values of σ_x for a deferred period of 13 weeks because the data sets for the shorter deferred periods (i.e., one week and four weeks) may be less typical of the general insured population. The values for the deferred period of 13 weeks are found in Table C16 of CMIR 12 (p. 74).

The force of sickness, σ_x , in CMIR 12 should be applied to the whole of the healthy population (i.e., states 1 and 2 combined) whereas $\mu_{23}(x)$ is a force that operates only on lives in state 2 (i.e.,

³CMIR 12 is based on data collected between 1975 and 1978. Subsequent work by Clark and Dullaway (1995), Haberman and Walsh (1998), and Renshaw and Haberman (2000) have suggested that PHI experience has changed since 1978.

the healthy state). It could be argued, therefore, that the values of σ_x taken from CMIR 12 should be adjusted. Because CMIR 12 is being used merely as a guide, no adjustments have been made, i.e., $\mu_{23}(x) = \sigma_x$.

$\mu_{16}(x)$ (**Super Healthy \rightarrow Dead**): Under CMIR 12 the morality rate for healthy lives is assumed to be that of male permanent assurances 1979-82, duration 0. The rates are shown in Table E17 (p. 132) under the column headed $m(x)$. In our model, we have divided healthy lives into *super healthy* and *ultimate healthy* states. Because lives in the latter state will experience higher mortality rates than those in the former, we have decided to assume: $\mu_{16}(x) = 0.80m(x)$, i.e., 80 percent of the mortality rates for male permanent assurances of 1979-82, duration 0.

$\mu_{26}(x)$ (**Ultimate Healthy \rightarrow Dead**): We assume $\mu_{26}(x) = 1.20m(x)$, i.e., 120 percent of the mortality rates for male permanent assurances of 1979-82, duration 0.⁴

$\mu_{32}(x)$ (**Short-Term Sick \rightarrow Ultimate Healthy**): Recovery rates are described in Section 3, Part B of CMIR 12. On page 34 of CMIR 12 various values of $\rho_{y+z,z}$, the transition intensity from sick to healthy at current age $y+z$ and current duration of sickness z , are displayed. These recovery rates vary markedly by duration of sickness (measured in weeks). In view of the relatively simple approach adopted in our model, we will use a constant parameter value, i.e., $\mu_{32}(x) = 2.5$ at all ages.

$\mu_{36}(x)$ (**Short-Term Sick \rightarrow Dead**): These mortality intensities are described in Section 6, Part B of CMIR 12. On page 39 of CMIR 12 the values of $\nu_{y+z,z}$ at various ages are displayed where $\nu_{y+z,z}$ is the transition intensity from sick to dead at current age $y+z$ and current duration of sickness z measured in weeks. For our calculations, we will use the values at 15 weeks duration of sickness, which is when the transition intensities reach their peak, i.e., $\mu_{36}(x) = \nu_{x,15}$. Interpolated values have been used where necessary.

$\mu_{34}(x)$ (**Short-Term Sick \rightarrow Long-Term Sick**): CMIR 12 does not provide explicit parameter values for $\mu_{34}(x)$. Having considered the or-

⁴The overall effect of the mortality assumptions for $\mu_{16}(x)$ and $\mu_{26}(x)$ can be considered to be broadly consistent with CMIR 12. As suggested by Cordeiro (1995), net premium values are likely to be less sensitive to the parameter values chosen for the forces of mortality.

der of magnitude of all the other forces in the model, we assume $\mu_{34}(x) = 0.1$ at all ages.

$\mu_{46}(x)$ (**Long-Term Sick → Dead**): We can again consider the mortality intensities $v_{y+z,z}$ that were described under $\mu_{36}(x)$ above. It seems appropriate to use these intensities at a suitably long sickness duration. We will use the values at duration five years (260 weeks) that are shown on page 39 of CMIR 12, i.e., $\mu_{46}(x) = v_{x,260}$.

$\mu_{56}(x)$ (**Lapse → Dead**): Because only super healthy policyholders lapse their policies, we will assume that $\mu_{56}(x) = \mu_{16}(x)$.

$\mu_{12}(x)$ (**Super Healthy → Ultimate Healthy**): CMIR 12 is not able to provide explicit parameter values for $\mu_{12}(x)$. It seems reasonable, however, to ensure that our estimates of $\mu_{12}(x)$ should be such that the aggregate mortality rates implied within our model approximately reflect the U.K. Male Permanent Assurances 1979–82 (duration 0) mortality table. The values for $\mu_{12}(x)$ that meet this constraint are, for simplicity, chosen by inspection.

$\mu_{15}(x)$ (**Super Healthy → Lapse**): Finally, having set the other parameters, $\mu_{15}(x)$ is varied in order to investigate its effect on the net premium rate.

Table 1 displays the parameter values. Table 2 shows the number of lives in each state at various sample ages given 100 super healthy lives entering state 1 at age 30, using the data in Table 1 and assuming $\mu_{15}(x) = 0.05$ for all x .⁵ For example, Table 2 shows that, by age 65, 12.0 percent of the lives would have died, 50.6 percent would have lapsed, and none of the lives would still be in the super healthy state.

The next step is to calibrate the model, i.e., to check if the model can produce the expected proportions of lives that are healthy, sick, or dead at various ages similar to those shown in CMIR 12 (Table E14, page 126). Table 3 displays these comparisons. The proportions are similar, particularly up to age 55. In Section 4.1 we will make another reasonableness check by comparing the net premium implied by our model with that implied by CMIR 12.

⁵The assumption $\mu_{15}(x) = 0.05$ is consistent with the assumption of Sanders and Silby (1986) who use a lapse rate of 5 percent per annum for policy duration greater than two years.

Table 1
Summary of Parameters

Age x	$\mu_{16}(x)$	$\mu_{26}(x)$	$\mu_{46}(x)$	$\mu_{36}(x)$	$\mu_{23}(x)$	$\mu_{12}(x)$
30-34	0.0003	0.0005	0.0172	0.1108	0.1982	0.0270
35-39	0.0004	0.0006	0.0190	0.1180	0.1766	0.0150
40-44	0.0006	0.0010	0.0215	0.1251	0.1560	0.0480
45-49	0.0011	0.0017	0.0239	0.1379	0.1408	0.1100
50-54	0.0019	0.0028	0.0271	0.1507	0.1337	1.1000
55-59	0.0031	0.0046	0.0303	0.1694	0.1375	1.5000
60-65	0.0049	0.0073	0.0343	0.1880	0.1576	2.0000

Notes: We have assumed (i) constant forces of transition over successive 5-year age bands (i.e., age 30-34, 35-39, ..., 60-64); and (ii) $\mu_{56}(x) = \mu_{16}(x)$, $\mu_{32}(x) = 2.5$, and $\mu_{34}(x) = 0.1$ for all x .

Table 2
Percent of Lives in Each State at Sample Ages

Age	State					
	1	2	3	4	5	6
30	100	0	0	0	0	0
31	92.6	2.5	0.1	0	4.8	0
32	85.7	4.7	0.3	0	9.3	0
.
.
.
.
50	13.4	30.3	1.5	1.5	50.0	3.3
.
.
.
65	0	32.4	1.9	3.1	50.6	12.0

Notes: Using the data from Table 1 and $\mu_{15}(x) = 0.05$.

Table 3
Comparing Percentages of Healthy, Sick and Dead Lives
Under CMIR 12 (Table E14) with Our Model

Age	CMIR 12 (Table E14)			Our Model		
	Healthy	Sick	Dead	Healthy	Sick	Dead
35	98.4	1.1	0.5	98.8	0.9	0.3
40	97.3	1.4	1.3	97.6	1.4	1.0
45	95.8	1.9	2.3	96.0	2.1	1.9
50	93.2	2.8	4.0	93.7	3.0	3.3
55	88.9	4.4	6.7	90.4	4.1	5.5
60	81.6	7.4	11.0	87.1	4.5	8.4

Notes: Our model uses the data from Table 1 and $\mu_{15}(x) = 0.05$.

4 The Main Results

The PHI policy under consideration here is a 35-year term policy issued to a life age 30. The sickness benefit is paid weekly during periods of sickness at the rate of £1,000 per annum. Premiums are paid weekly and are waived during periods of sickness. Benefits are paid on a weekly basis. There is no deferred period, and the benefits and premiums cease at the age of 65. The valuation rate of interest is set to 6 percent per year. The forces of transition used are given in Table 1.

4.1 Sensitivity of Net Premiums to Various Parameters

Sensitivity of P to $\mu_{15}(x)$: Figure 2 shows how the net premium varies as the lapse rate $\mu_{15}(x)$ takes values between 0 and 1. The net premium is relatively insensitive to the lapse rate. For example, the net premium decreases from £33.79 per annum to £26.36 per annum as the lapse rate increases from 0 to 0.2. This relative insensitivity is due to the fact that only super healthy lives lapse their policies, and their reserves are relatively small. Lapse rates of more than 0.4 would be unrealistic. For example, it can be shown that if $\mu_{15}(x) = 0.4$, over 83 percent of the insured population age 30 at the outset would have lapsed their policy during the first five years of the policy.

It is surprising that the net premium decreases rather than increases as the lapse rate increases, which is counter-intuitive. Standard actuarial logic suggests that the net premium should increase, because when the lapse rate is small, there are large numbers of lives in the system who are in the super healthy state and therefore continue to pay premiums without receiving any PHI benefit payments. This tends to suppress the net premium averaged over all the policyholders in the system. As the lapse rate increases, more of the super healthy lives leave the system by lapsing, which will tend to increase the average premium payable in respect of the remaining, relatively unhealthy, insured population.

So why does the net premium decrease as the lapse rate increases? Figure 3 shows how the numerator and the denominator of the right side of equation (16) vary as the lapse rate increases. We show scaled versions of the numerator and the denominator in order to fit them on the same graph. Both numerator and the denominator decrease, as would be expected, because the effect of lapses is to remove lives from state 1 before they have an opportunity to enter states 2, 3, or 4. The rate of decrease is the result of the complicated interaction between the different forces within the model. It can be seen that the numerator decreases at a faster rate than the denominator, and, therefore, the overall effect is that the net premium decreases.

Finally, before discussing other sensitivity issues, it is worth comparing the net premiums calculated using the model described in this paper with those derived from the data in CMIR 12. The data contained in Table F1 on page 228 of CMIR 12 suggest that the net premium for a policy similar to that described earlier in this section, but with premium and benefit payments made continuously and with a deferred period of one week, should be £24.24 per annum. The net premium figures shown in Figure 2 are of the same magnitude and hence provide some comfort that our model (including the parameter values chosen) is consistent with the model described in CMIR 12.

Sensitivity of P to $\mu_{12}(x)$: Figure 4 shows how the net premium changes when the parameter values for $\mu_{12}(x)$ given in Table 1 are increased or decreased 10 percent. If $\mu_{12}(x)$ is increased 10 percent, the net premium increases between 4.9 percent (when the lapse rate, $\mu_{15}(x) = 0$) and 8.4 percent (when $\mu_{15}(x) = 1.0$). If $\mu_{12}(x)$ is reduced 10 percent, the net premium decreases between 5.1 percent (when $\mu_{15}(x) = 0$) and 8.7 percent (when $\mu_{15}(x) = 1.0$).

Figure 2
Sensitivity of Net Premium to Lapse Rate, μ_{15}

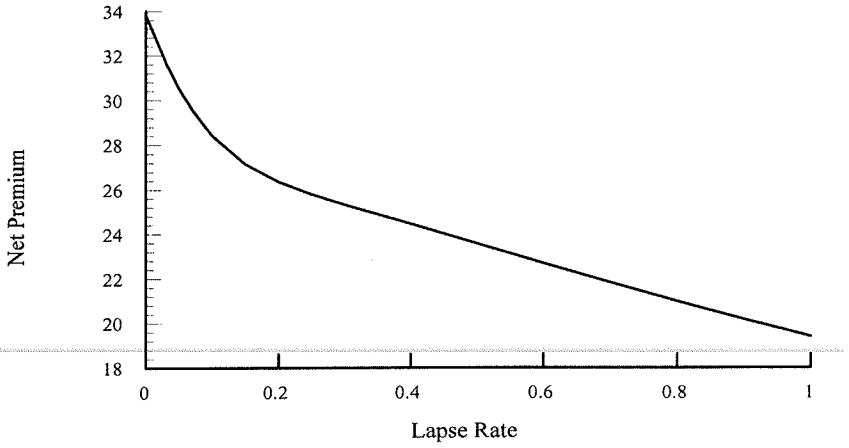
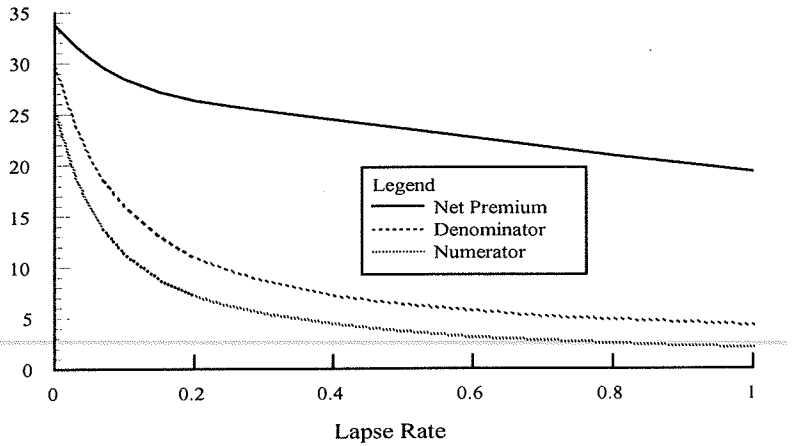


Figure 3
Variations in the Numerator and Denominator of the Net Premium as the Lapse Rate Increases



The net premium is expected to move in the same direction as $\mu_{12}(x)$. An increase in $\mu_{12}(x)$ causes more lives to move from the super healthy to the ultimate healthy state where they are exposed to the risk of sickness inception, which, in turn, will lead to an increase in the premium required.

Sensitivity of P to $\mu_{23}(x)$: Figure 5 shows how net premiums change when the parameter values for $\mu_{23}(x)$, the sickness inception rate, are altered 10 percent. The net premium increases approximately 8.6 percent when the $\mu_{23}(x)$ values are increased 10 percent and decreases approximately 8.9 percent when the $\mu_{23}(x)$ values are decreased 10 percent. These results (in terms of relative sensitivities) are largely unaffected by the level of lapse rate assumed. As expected, an increase in the sickness inception rate causes an increase in the net premium required.

Cordeiro (1995) extends the work described in CMIR 12 by considering the effect on net premiums in changes in the sickness inception rates for various deferred periods and entry ages. Cordeiro finds that, for the CMIR 12 model and data, if the sickness inception rate is doubled, the net premium is approximately doubled. The results of this paper are therefore consistent with those of Cordeiro (1995).

Sensitivity of P to $\mu_{32}(x)$: Figure 6 shows how net premiums change when the parameter value for $\mu_{32}(x)$, the recovery rate, is increased or decreased 10 percent (i.e., changed from 2.5 at all ages to 2.75 or 2.25, respectively).

The net premium increases approximately 8.3 percent when the recovery rate is reduced 10 percent and decreases approximately 7.2 percent when it is increased 10 percent. Again, the level of lapse rate has little effect on these relative sensitivities. It is to be expected that an increase in the recovery rate should lead to a reduction in the amount of PHI premium required.

Cordeiro (1995) has investigated the effect that changes in the recovery rates have on net premiums based on the CMIR 12 model and data. Cordeiro discovers that a 10 percent increase in the recovery intensity leads to a 27.6 percent reduction in the net premium for entry age 30 and deferred period one week. Therefore, the net premium is less sensitive to a change in the recovery intensity under the model described in this paper than under the model used by Cordeiro (1995).

Figure 4
Net Premium Sensitivity to a $\pm 10\%$ Change in $\mu_{12}(x)$

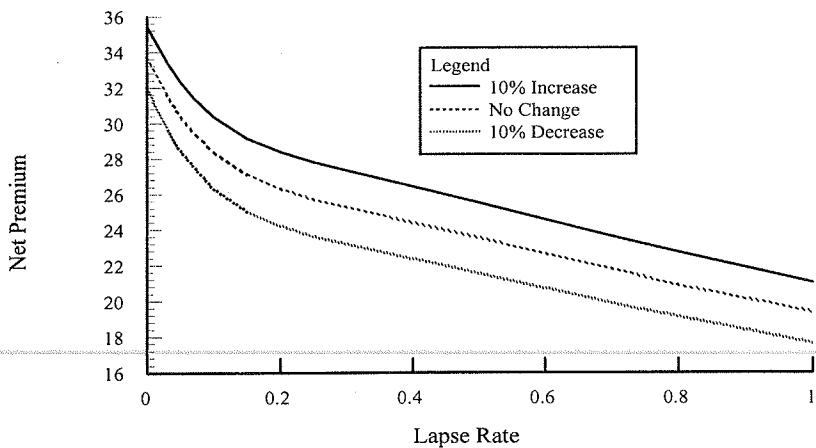
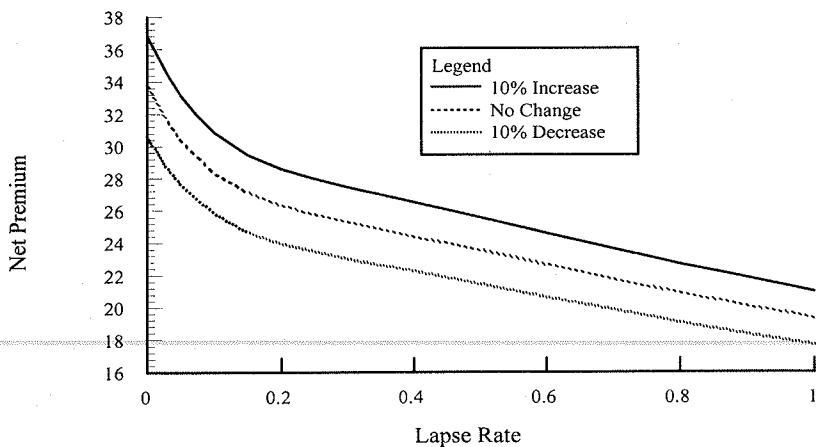


Figure 5
Net Premium Sensitivity to
A $\pm 10\%$ Change in the Sickness Inception Rate $\mu_{23}(x)$



Sensitivity of P to $\mu_{34}(x)$: Figure 7 shows the changes in net premiums when the parameter values for $\mu_{34}(x)$ are increased or decreased 10 percent.

It can be seen that the net premium is relatively insensitive to changes in $\mu_{34}(x)$ because a 10 percent increase/decrease in the latter causes only a 4.0 percent increase/decrease in the net premium. As expected, an increase in the long-term sickness inception rate leads to an increase in the net premium required.

4.2 The Relationship Between $\mu_{12}(x)$ and $\mu_{32}(x)$

In Section 3, we explain how the parameter values for μ_{12} are chosen so that the aggregate mortality rates within the model broadly reflect the male permanent assurances 1979–82, duration 0. We now analyse how sensitive the values of $\mu_{12}(x)$ are to a change in the other parameters, in particular to a 50 percent increase in the recovery rate, $\mu_{32}(x)$. In other words, we retain all the parameter values summarized in Table 1 except for $\mu_{32}(x)$, which we increase from 2.5 at all ages to 3.75, and $\mu_{12}(x)$, which we need to recalibrate in order to ensure that the aggregate mortality rates still reflect the mortality table mentioned above. The results are summarized in Table 4.

Table 4
Comparison of $\mu_{12}(x)$ Values
When $\mu_{32}(x)$ Increases

Age	$\mu_{12}(x)$ Values when	
	$\mu_{32}(x) = 2.5$	$\mu_{32}(x) = 3.75$
30–34	0.027	0.045
35–39	0.015	0.025
40–44	0.048	0.074
45–49	0.110	0.180
50–54	1.100	1.500
55–59	1.500	1.900
60–64	2.000	2.400

A 50 percent increase in $\mu_{32}(x)$ requires an increase in $\mu_{12}(x)$ of approximately the same order of magnitude up to age 50 in order to leave the aggregate mortality rates within the model unaltered.

Figure 8
Impact on Net Premium of Increasing $\mu_{32}(x)$
(From $\mu_{32}(x) = 2.50$ to $\mu_{32}(x) = 3.75$)

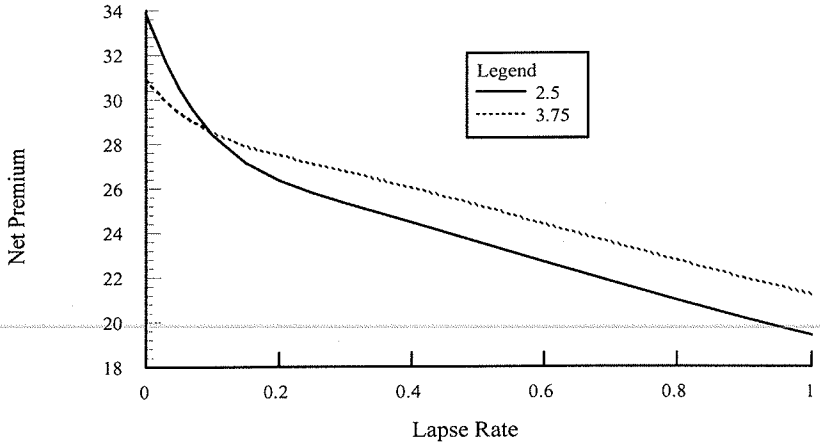
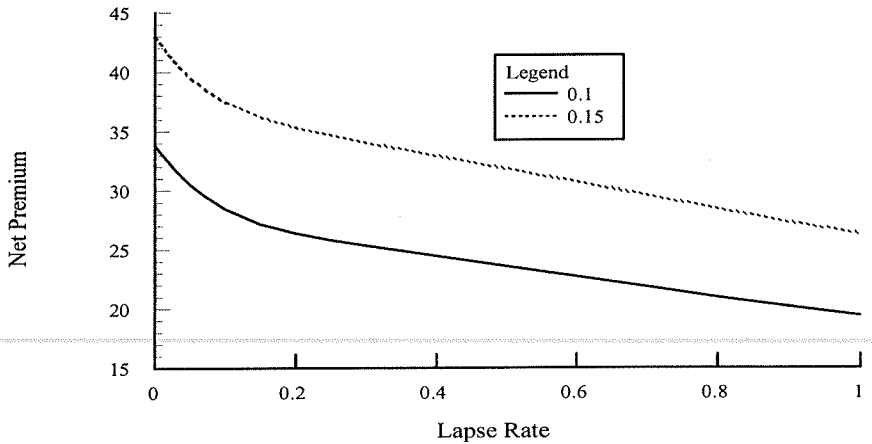


Figure 9
Impact on Net Premium of Increasing $\mu_{34}(x)$
(From $\mu_{34}(x) = 0.10$ to $\mu_{34}(x) = 0.15$)



This result involving changes to $\mu_{34}(x)$ and $\mu_{12}(x)$ contrasts with the result in Section 4.2 where increasing $\mu_{32}(x)$ and recalibrating $\mu_{12}(x)$ has a neutral effect on the net premium. This feature further illustrates how complicated the interaction between the transition intensities is within the model.

5 Closing Comments

An objective of this paper is to develop a simple, practical U.K. style PHI model that can be used by actuaries who do not have access to complex models such as CMIR 12 or the detailed data required to use such models or who are interested in rough estimates for net premiums for PHI models.

One of the main difficulties that needs to be overcome in maintaining the simplicity of the model, however, is that the forces of transition between different states may depend not only on the age of the policyholder, but also on the time spent in the current state. For example, the longer a policyholder remains in the sick state, the less likely he or she is to recover. That is, there is duration-dependence. This factor usually leads to a semi-Markov model being used. However, convenient expressions for the transition probabilities are then hard to obtain.

The problem of duration-dependence is handled, in part, by increasing the number of states to differentiate between short-term and long-term stays in a particular status. This enables the model to be Markov rather than semi-Markov and therefore leads to tractable solutions. The model also includes lapses.

Using a particular policy, we test the sensitivity of the net premium to changes in the most significant model parameter values ($\mu_{12}(x)$, $\mu_{15}(x)$, $\mu_{23}(x)$, $\mu_{32}(x)$, and $\mu_{34}(x)$). Not surprisingly, the net premium is relatively insensitive to changes in the lapse rate ($\mu_{15}(x)$) because only the most healthy lives are assumed to lapse their policies and they have small reserves. We also find that when any of the forces of transition, $\mu_{23}(x)$, $\mu_{32}(x)$, or $\mu_{34}(x)$, are increased, the resultant change in the level of net premium depends little on the level of the lapse rate. As a result, actuaries may initially ignore lapse rates when considering rough estimates for net premiums for PHI models.

By contrast, however, when the force of transition from the super healthy to the ultimate healthy state ($\mu_{12}(x)$) is increased, the extent to which the net premium increases depends on the level of the lapse rate. This shows that actuaries should probably spend more of their energies trying to obtain accurate estimates of $\mu_{12}(x)$.

Let $p_{ir} = \mathbb{E} [X_{ij}^r]$ for $r = 1, 2, \dots$. The first three cumulants of $S(t)$ are

$$\kappa_1 = t \sum_{i=1}^m \lambda_i p_{i1}, \quad \kappa_2 = t \sum_{i=1}^m \lambda_i p_{i2}, \quad \text{and} \quad \kappa_3 = t \sum_{i=1}^m \lambda_i p_{i3}.$$

Using the SDS principle, the accumulated risk premium received in $(0, t)$ (ignoring interest) is $\Pi^{\text{SDS}}[S(t)] = \Pi^{\text{SDS}}(t)$, where

$$\Pi^{\text{SDS}}(t) = \kappa_1 + \alpha_1 \kappa_2^{\frac{1}{2}} + \alpha_2 \kappa_3^{\frac{1}{3}}. \tag{6}$$

It must be pointed out that although $\Pi^{\text{SDS}}(t)$ is the accumulated risk premium received in $(0, t)$, it does not specify the amount of premium received in an intermediate period $(0, s)$ for $0 < s < t$. Let $\Pi^{\text{SDS}}(s|t)$ denote the accumulated risk premium received in $(0, s)$ for $0 < s < t$. All that is known is $\Pi^{\text{SDS}}(0|t) = 0$ and $\Pi^{\text{SDS}}(t|t) = \Pi^{\text{SDS}}(t)$. How must $\Pi^{\text{SDS}}(s|t)$ be defined for fixed t ? There are several possibilities, for example,

$$\Pi^{\text{SDS}}(s|t) = \int_0^s d\Pi^{\text{SDS}}(y) \quad 0 < s < t$$

or

$$\Pi^{\text{SDS}}(s|t) = c_t s \quad 0 < s < t.$$

where c_t is a constant for fixed t . As premiums are usually collected at a constant rate, we propose the second approach with

$$c_t = \frac{\Pi^{\text{SDS}}(t)}{t}. \tag{7}$$

Let $\theta(t)$ denote the relative security loading in $\Pi^{\text{SDS}}(t)$ so that

$$\Pi^{\text{SDS}}(t) = (1 + \theta(t))\kappa_1$$

and

$$\begin{aligned} \theta(t) &= \frac{\alpha_1 \kappa_2^{\frac{1}{2}} + \alpha_2 \kappa_3^{\frac{1}{3}}}{\kappa_1} \\ &= \frac{\alpha_1 (\sum_{i=1}^m \lambda_i p_{i2})^{\frac{1}{2}}}{t^{1/2} \sum_{i=1}^m \lambda_i p_{i1}} + \frac{\alpha_2 (\sum_{i=1}^m \lambda_i p_{i3})^{\frac{1}{3}}}{t^{2/3} \sum_{i=1}^m \lambda_i p_{i1}}. \end{aligned}$$

Notice that for fixed α_1 and α_2 , $\theta(t) \rightarrow 0$ as $t \rightarrow \infty$. This property of $\theta(t)$, i.e., converging to zero for long-term contracts, also exists for other premium calculation principles such as the standard deviation principle and makes these premium calculation principles unsuitable for long-term contracts.

Consider a time horizon of t years. Let $U(\tau)$ denote the surplus at time τ ($0 < \tau < t$), then

$$U(\tau) = u + c_t\tau - S(\tau)$$

with $U(0) = u \geq 0$ being the initial surplus. The ruin probability within t years given an initial surplus of u , $\psi(u, t)$, is defined as

$$\psi(u, t) = \mathbb{P}[T(u) \leq t] \tag{8}$$

where $T(u) = \min\{\tau : \tau > 0 \text{ and } U(\tau) < 0\}$. It is evident that the function ψ depends on the size of u , c_t , and the time horizon t .

For a compound Poisson process with a fixed relative security loading on the risk premium, two well-known results are that the probability of ruin depends only on the size of the relative security loading, and that it increases as the size of the loading decreases. These results are used to determine c_t .

Specifically, to determine the premium rate c_t , we set ψ of equation (8) at an acceptable level and then solve the resulting equation for c_t . If ϵ is our acceptable probability of ruin (typically, $\epsilon < 0.05$), we must solve the equation

$$\psi(u, t) = \epsilon.$$

As $\psi(u, t)$ is a complicated function of the premium rate, c_t is determined directly through simulations. Note that for fixed u and t , $\psi(u, t)$ decreases as the relative security loading increases, i.e., as c_t increases. This inverse relationship enables us to search for solutions using the bisection method.

4 The Determination of Parameters α_1 and α_2

The c_t obtained using simulations is actually the premium rate needed to cover m classes of risks at the acceptable level of the probability of

ruin. Hence, the value of $c_t t$ is an aggregate of m classes of premiums collected over t years. The question here is how do we allocate $c_t t$ among these m classes? Though there are several approaches that can be used, we opt for the one that allows us to set the m premiums via the SDS premium calculation principle, i.e., we choose the parameters so that the α_1 s are the same for each class and the α_2 s are the same for each class (α_1 and α_2 may be different). This means that the premium for each class satisfies the SDS premium calculation principle.

Let c_{it} denote the premium allocated to the i^{th} class. Set

$$\begin{aligned} c_t t &= \sum_{i=1}^m c_{it} \\ &= \sum_{i=1}^m \left(\lambda_i p_{i1} t + \alpha_1 (\lambda_i p_{i2} t)^{\frac{1}{2}} + \alpha_2 (\lambda_i p_{i3} t)^{\frac{1}{3}} \right). \end{aligned} \quad (9)$$

Because we only have one equation but two unknown parameters, we need to impose a relation between α_1 and α_2 . We assume that

$$\alpha_1 = \gamma \alpha_2 \quad (10)$$

where $\gamma > 0$ is a known constant. In practice, γ can be chosen in accordance with the insurers' preferences and claim experiences.

Combining equations (9) and (10), we get

$$c_t t = \sum_{i=1}^m \left(\lambda_i p_{i1} t + \gamma \alpha_2 (\lambda_i p_{i2} t)^{\frac{1}{2}} + \alpha_2 (\lambda_i p_{i3} t)^{\frac{1}{3}} \right). \quad (11)$$

For a given γ , we can easily solve equation (11) for α_2 . Then, α_1 can be obtained using equation (10).

4.1 Simulation Assumptions

The following assumptions are used:

- There are two classes, i.e., $m = 2$.
- The time horizons used are $t = 10, 50, 100$.
- The t -year ruin probability is set to be 0.05.
- The initial reserves used are $u = 10, 20, 30$.

- The premium is paid continuously at a constant rate of c_t per year.
- For $i = 1, 2$, $N_i(t)$ is a Poisson process with $\lambda_i = 10$. Hence, the claim number process $N(t)$ is a Poisson process with $\lambda = 20$. This implies that the inter-occurrence time random variables (i.e., the times between successive claims) are exponential with mean $1/\lambda$; see Bowers et al., (1997, Chapter 13.3).
- Two pairs of claim size distributions are used. They are specified in two cases:

Case 1: (Exponential-Lognormal Pair) The claim size X_{1j} has an exponential distribution with density $f_1(x) = e^{-x}$, and X_{2j} has a lognormal distribution, i.e., $\ln X_{2j} \sim N(\mu, \sigma^2)$, where $\mu = -\ln(2)/2$ and $\sigma^2 = \ln 2$. In this case, $p_{11} = 1$, $p_{12} = 2$, $p_{13} = 6$, and $p_{21} = 1$, $p_{22} = 2$, $p_{23} = 8$; and

Case 2: (Gamma-Pareto Pair:) The claim size X_{1j} has a gamma distribution with density

$$f_1(x) = \frac{\eta^\eta x^{\eta-1} e^{-\eta x}}{\Gamma(\eta)}$$

where $\eta = 4$. The claim size X_{2j} has a Pareto distribution with density

$$f_2(x) = \frac{\beta + 1}{\beta} \left(\frac{\beta}{\beta + x} \right)^{\beta+2}$$

where $\beta = 3$. In this case, $p_{11} = 1$, $p_{12} = 1.25$, $p_{13} = 1.875$, and $p_{21} = 1$, $p_{22} = 3$, $p_{23} = 27$.

The simulation is performed as follows. Let T_k denote the occurrence time of the k^{th} claim (Z_k) and define $V_k = T_k - T_{k-1}$ with $T_0 = 0$. The V_k s are called the *inter-occurrence time random variables*. Define W_k as

$$W_k = u + \sum_{r=1}^k (c_t V_r - Z_r).$$

Ruin occurs if W_k is ever negative for any $k = 1, 2, \dots, N(t)$ where $N(t)$ is the total number of claims generated by the two classes in $(0, t)$.

Step 1: As $T_n = V_1 + \dots + V_n$ for $n = 1, 2, \dots$, generate the sequence of inter-occurrence time random variables V_k s until the condition

$$T_n \leq t < T_{n+1}$$

occurs, then stop; see Ross (1990) for more on generating pseudo-random variables;

Step 2: Assign $N(t) = n$ and $W_0 = u$;

Step 3: For $k = 1$ to $N(t)$, do the following:

1. Generate a uniform (0,1) random number U . If $U < \lambda_1/\lambda$, then generate Z_k from the claim distribution of class 1 (i.e., the distribution of X_{1j}), else generate Z_k from the claim distribution of class 2 (i.e., the distribution of X_{2j});
2. Compute $W_k = W_{k-1} + c_t V_k - Z_k$;
3. If $W_k < 0$, then ruin occurs. Return to Step 1 to start another simulation;
4. If $W_k \geq 0$, then go back to Step 3.1 above to continue the loop;

Step 4: If $W_k \geq 0$ for $k = 1$ to $N(t)$, then ruin does not occur. Return to Step 1 for another simulation.

For each of the two cases and for each u and t , we perform 10,000 simulations. We choose the value of c_t that yields 500 ruins out of the 10,000 simulations (as the ruin probability is set to be 0.05). Then, based on equation (11), we use

$$c_t t = \left(10t + \gamma \alpha_2 (20t)^{\frac{1}{2}} + \alpha_2 (60t)^{\frac{1}{3}}\right) + \left(10t + \gamma \alpha_2 (20t)^{\frac{1}{2}} + \alpha_2 (80t)^{\frac{1}{3}}\right)$$

for Case 1, and

$$c_t t = \left(10t + \gamma \alpha_2 (12.5t)^{\frac{1}{2}} + \alpha_2 (18.75t)^{\frac{1}{3}}\right) + \left(10t + \gamma \alpha_2 (30t)^{\frac{1}{2}} + \alpha_2 (270t)^{\frac{1}{3}}\right)$$

for Case 2, with γ varying from 0 to 5 in steps of 0.1, to calculate α_2 . Once α_2 is obtained, we compute α_1 using equation (10).

4.2 Numerical Results

The results are summarized in Figures 1 to 4 and Tables 1 and 2. Figures 1 and 2 show that α_1 decreases as u increases, while Figures 3 and 4 show that α_1 increases as t increases. Similar observations also hold for α_2 because of equation (10). Notice that in the first row of Table 1, the c_t values for $t = 10, 50, 100$ are the same. This suggests that in both cases, the c_t value with $u = 10$ and $t = 10$ is close to the largest premium for a probability of ultimate ruin of 0.05. The second observation is that for fixed t , the larger the value of u , the smaller the value of c_t . This is consistent with Figures 1 and 2.

Table 1
Values of c_t for Various Values of u and t

u	Case 1			Case 2		
	Exponential-Lognormal			Gamma-Pareto		
	$t = 10$	$t = 50$	$t = 100$	$t = 10$	$t = 50$	$t = 100$
10	27.40	27.40	27.40	29.94	29.94	29.94
20	23.18	23.30	23.30	23.68	24.12	24.12
30	21.39	22.16	22.18	21.79	22.39	22.39

Table 2
Values of c_{1t} and c_{2t} with $u = 10$ and $t = 50$

γ	Case 1			Case 2		
	Exponential-Lognormal			Gamma-Pareto		
	α_1	c_{1t}	c_{2t}	α_1	c_{1t}	c_{2t}
0.1	1.0104	13.5535	13.8468	1.2435	13.0559	16.8853
0.5	2.9879	13.6134	13.7869	3.7965	13.3845	16.5567
1.0	3.9556	13.6427	13.7576	5.1071	13.5532	16.3880
1.5	4.4343	13.6572	13.7431	5.7711	13.6387	16.3025
2.0	4.7200	13.6659	13.7344	6.1724	13.6903	16.2509
2.5	4.9097	13.6716	13.7287	6.4412	13.7249	16.2163
3.0	5.0449	13.6757	13.7246	6.6337	13.7497	16.1915
3.5	5.1461	13.6788	13.7215	6.7785	13.7683	16.1729
4.0	5.2247	13.6812	13.7191	6.8912	13.7828	16.1584
4.5	5.2876	13.6831	13.7172	6.9816	13.7945	16.1467
5.0	5.3389	13.6847	13.7156	7.0556	13.8040	16.1372

Figure 1
 α_1 Vs. γ for Exponential-Lognormal with $t = 50$

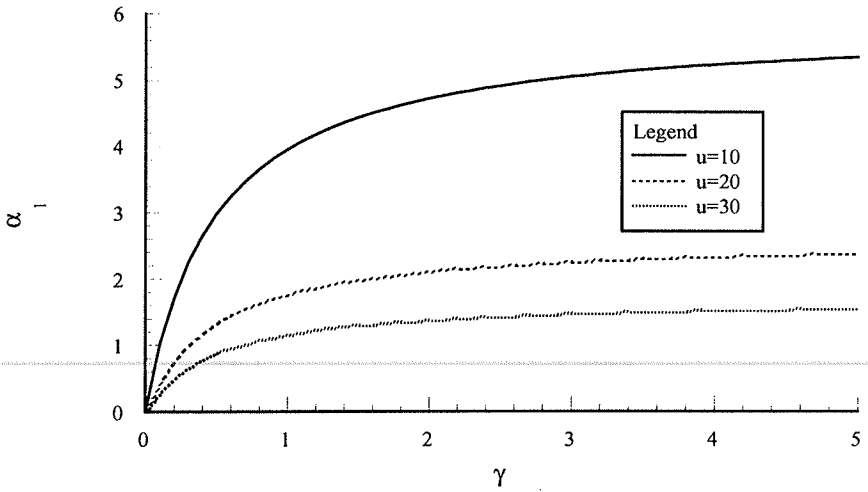


Figure 2
 α_1 Vs. γ for Gamma-Pareto with $t = 50$

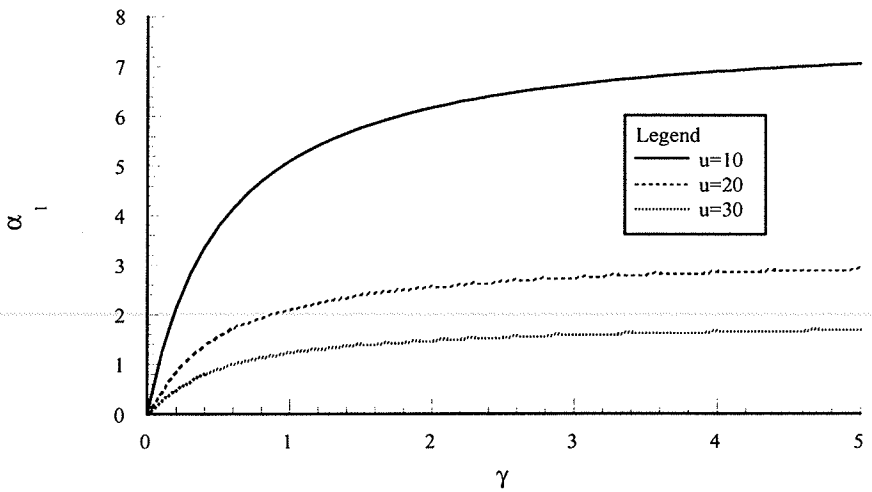


Figure 3
 α_1 Vs. γ for Exponential-Lognormal with $u = 10$

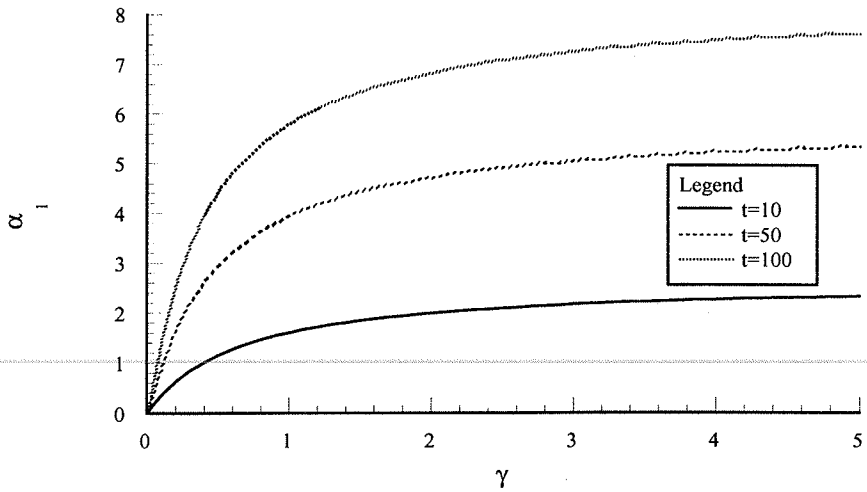


Figure 4
 α_1 Vs. γ for Gamma-Pareto with $u = 10$

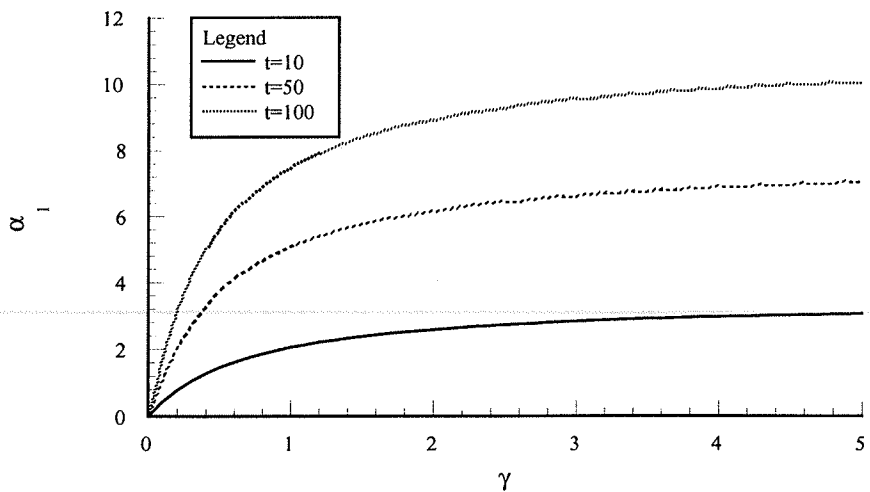


Table 2 displays c_{1t} and c_{2t} for $u = 10$ and $t = 50$. In both cases c_{2t} exceeds c_{1t} . In the exponential-lognormal case, the third cumulant of the lognormal is slightly larger than that of the exponential so c_{1t} and c_{2t} differ only by a small margin. Moreover, c_{2t} exceeds c_{1t} because the lognormal is riskier (i.e., has a heavier right tail) than the exponential. In the gamma-Pareto case, the differences are much larger because the Pareto has a larger second cumulant and a much larger third cumulant, i.e., the Pareto is much riskier than the gamma. In both cases, $c_{2t} - c_{1t}$ decreases as y increases because α_1 (α_2) becomes larger (smaller) when y increases, so a heavier (lighter) weight is put on the standard deviation (skewness) term.

5 Closing Remarks

There are three important points that must be addressed:

1. Ruin probabilities are difficult to obtain because they do not usually have closed-form solutions, so the method of simulations is a natural way to deal with the problem. One advantage of simulation is flexibility. It can be used in practical situations with real insurance data as well as more complex models that include factors such as correlated risks and investment performance.
2. From the practical point of view, the value of t should not be set too large because it leads to lower risk loading factors. If the insurance market is such that one can split the time horizon into smaller time periods, then the insurer may receive higher risk loadings over each period. For example, a 10-year horizon may be split into five 2-year horizons.
3. The question of allocating premiums among the m classes has no unique solution. For example, we can allocate the premiums according to their proportion of the total risk loadings. Specifically, using equation (2), we define

$$c_{it} = \lambda_i p_{i1} t + \frac{(\alpha_1 + \alpha_2 \varphi_i^{\frac{1}{3}}) \sigma_i}{\sum_{i=1}^m (\alpha_1 + \alpha_2 \varphi_i^{\frac{1}{3}}) \sigma_i} \left(\alpha_1 \kappa_2^{\frac{1}{2}} + \alpha_2 \kappa_3^{\frac{1}{3}} \right)$$

where $\sigma_i^2 = \text{Var}[X]$ and $\varphi_i = \mathbb{E}[(X_{ij} - \mathbb{E}[X_{ij}])^3] / \sigma_i^3$ is the coefficient of skewness of X_{ij} . As before, we set $c_t t = \sum_{i=1}^m c_{it}$.

References

- Ash, R.B. and Doléans-Dade, C.A. *Probability and Measure Theory, Second Edition*. San Diego, Calif.: Harcourt Academic Press, 2000.
- Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. and Nesbitt, C.J. *Actuarial Mathematics, Second Edition*. Schaumburg, Ill.: Society of Actuaries, 1997.
- Buhlmann, H. "An Economic Premium Principle." *ASTIN Bulletin* 11 (1980): 52-60.
- Gerber, H.U. *An Introduction to Mathematical Risk Theory*. Philadelphia, Pa.: Huebner Foundation, 1979. (Distributed by Irwin, Inc., Homewood, Ill.)
- Goovaerts, M.J., de Vylder, F. and Haezendonck, J. *Insurance Premiums: Theory and Applications*. Amsterdam, Holland: North-Holland, 1984.
- Kamps, U. "On a Class of Premium Principle Including the Esscher Principle." *Scandinavian Actuarial Journal* (1998): 75-80.
- Ramsay, C.M. "Loading Gross Premiums for Risk Without Using Utility Theory." *Transactions of the Society of Actuaries* 45 (1993): 305-336.
- Ross, S.M. *A Course in Simulation*. New York, N.Y.: Macmillan, 1990.
- Schmidt, K.D. "Positive Homogeneity and Multiplicativity of Premium Principles on Positive Risks." *Insurance: Mathematics and Economics* 8 (1989): 315-319.
- Van Heerwaarden, A.E., Kaas, R. and Goovaerts, M.J. "Properties of the Esscher Premium Calculation Principle." *Insurance: Mathematics and Economics* 8 (1989): 261-267.
- Wang, S. "Insurance Pricing and Increased Limits Ratemaking by Proportional Hazards Transform." *Insurance: Mathematics and Economics* 17 (1995): 43-54.

