Insurance Pricing and Capitalisation Given Market Incompleteness and Frictional Costs

Mark Johnston
PricewaterhouseCoopers,
GPO Box 2650, Sydney NSW 1171, Australia

14 July 2003
Revised 11 August 2004

Abstract

In this paper we re-visit the principles of insurance pricing, using a modern economic valuation framework. We do not seek to explain all asset purchasing decisions; rather we start from the observable fact that individuals hold concentrations of wealth in particular assets, such as their family home, and engage in activities, such as driving cars, that produce concentrations of liability. We then seek to understand what economic valuation models imply about the prices of insurance policies that are designed to mitigate the risks associated with such assets and liabilities.

Our primary conclusion is that for typical insured risks, there will be a range of feasible insurance premiums — where a premium is defined as *feasible* if it makes entering into the insurance contract value-creating for both policyholders and shareholders. Where in the feasible range premiums will be, or should be, set, will be determined by the level of competition and regulation in the market for insurance policies (the "consumer insurance market"). The existence of this *feasible range* requires the existence of a positive *insurance surplus*, the latter being defined here as the difference between the sum of the values placed by consumers upon the policies in a portfolio and the market value of this portfolio. The existence of this surplus relies on two assumptions — firstly, that the risks being insured cannot be offset by securities traded on capital markets (implying that capital markets are incomplete); secondly, that consumers are averse to these risks (in fact to the non-traded components of them).

The existence of this surplus implies that insurance enterprises could be set up in a manner that improves the lot of both policyholders and those who provide capital to the insurance enterprise. In the latter part of this paper we investigate the corporate form of the insurance enterprise, showing how to determine the feasible combinations of shareholders' funds and premium, in the presence of taxes, expenses, and limited liability. We show that expenses and taxes, which can be viewed as "frictional costs" of this form of risk financing, induce upper and lower bounds on the feasible levels of capitalisation for an insurance company.

Since securities markets are incomplete with respect to the risks being insured, one cannot determine insurance prices by looking at security prices alone. One can view the consumer insurance market as "completing" the market for these risks — it is thus a primary source of pricing information.

We finish with an illustration of how pricing information might be inferred from the consumer insurance market.

This paper was developed with application to pricing of general insurance in mind, but the methods and results may be applicable to life insurance also, and may be relevant to the current debate about the notion of fair value of insurance liabilities in international accounting standards.

1 Introduction

1.1 Background

When buying an insurance policy, a consumer exchanges a known sum of money (the premium) for an uncertain payoff. The insurance payoff is designed to offset other risky payoffs the consumer has — for example, car insurance pays off just when the consumer has to pay a large car repair bill after a crash. Because of this offsetting effect, and because the events insured are typically of a magnitude that would cause financial hardship if the consumer was not insured, it has been broadly accepted since the early days of insurance that consumers will be willing to pay a premium that exceeds the *risk-neutral value* — the expected insurance payoff discounted at the risk free rate[1]. Consequently, consumers have been willing to purchase policies from insurers who set their prices using traditional actuarial methods, which typically include a risk premium in some form.

[1]Moller [Mol02] attributes this logic to Daniel Bernoulli's 1738 consideration of the risk preferences of individuals in gambling settings

More recently, actuaries have applied economic valuation models, such as the *Capital Asset Pricing Model (CAPM)*, to insurance companies. The 1981 paper of Myers and Cohn (eventually published in 1987 [MC87]) contains three contributions:

1. They suggest that definitions of premium fairness are best framed in terms of present values (of the components of the insurer's balance sheet), rather than in terms of rates of return.

2. They propose a definition of a "fair" policy pricing method as one which sets premiums at a level that makes accepting the business an NPV-zero proposition for investors, meaning that the market value of the shareholders' equity claim is equal to the amount of shareholders' funds that is required to support the policies. They take the latter amount to be a fixed proportion of the premium, with the proportion determined exogenously — for example, by a regulator.

3. In calculating the market value of the shareholders' equity claim, and of the other components of the insurer's balance sheet, they employ the CAPM.

They do point out that their second and third contributions are distinct — that the use of CAPM is not necessary, as their "Discounted Cash Flow" method can be applied regardless of which method is used to determine the present values of the balance sheet components.

In the CAPM, the expected or required rate of return on an asset is calculated from the "beta" of the asset. An example of the application of the Myers-Cohn approach can be found in an issues paper developed for the NSW Motor Accidents Authority [Act00] on "Capital and Profit in CTP Insurance". In that paper it is suggested that an assumption of zero beta for the technical insurance liabilites might be reasonable. In the absence of taxes this would lead to a conclusion that CTP insurance should be priced at the expected level of payoffs, plus expenses, discounted at the risk-free rate (so with no loading for risk). Even when taxes are considered the application of CAPM and the Myers-Cohn fair premium principle described in the paper leads to premiums that are considered by some practitioners to be low relative to accepted practice [Lee01].

1.2 Outline of this paper

In this paper we re-visit the principles of insurance pricing, using modern economic valuation methods. We shall start (here) by agreeing with the Myers-

Cohn suggestion that discussions about fair premiums are best[2] framed in terms of present values, rather than rates of return.

We then consider the second contribution of Myers and Cohn's paper — their proposed definition of a fair pricing principle: set the premium at the level that makes entering into the insurance contract an NPV-zero proposition for shareholders. In Section 3 we propose a related definition: the *feasible range of insurance premiums* is the range of premiums that make entering into the insurance contract an NPV-positive proposition for both the shareholders and the customers of the insurance company. Where such a range exists, the Myers-Cohn definition of the "fair" price represents the lower boundary of the feasible range. Where in the feasible range premiums will be, or should be, set, will be determined by the level of competition and regulation in the market for insurance policies (the "consumer insurance market"). We argue that in "typical" un-regulated markets premiums might be set somewhere above the lower boundary of the feasible range, so that in some regulated markets it may also be appropriate to set premiums above the Myers-Cohn "fair" price.

The existence of a feasible range of insurance premiums requires the existence of a positive *insurance surplus*, the latter being defined here as the difference between the sum of the values placed by consumers upon a portfolio of insurance policies and the market value of this portfolio. The existence of a positive surplus in typical insurance contexts would be taken as obvious by practising actuaries — that's where the expenses and taxes are paid from, for a start. However, we feel it is worthwhile to exhibit within a rigorous economic valuation framework the existence of such a surplus, as much of the current debate around fair pricing is cast in terms of economic valuation models, but leaves one with the impression that the surplus may have been forgotten. We will see that the key attribute of the securities market necessary for the existence of an insurance surplus is incompleteness — certain states of the world of interest to consumers, such as states where their house burns down, cannot be replicated by trading securities, so there may be an opportunity to create value for both shareholders and consumers through the creation of an insurance enterprise.

In Section 4 we examine the corporate form of the insurance enterprise, investigating how the insurance surplus and the feasible range of premiums are affected by limited liability, expenses and taxation. We will see that in the presence of all three of these characteristics, the range of feasible levels of asset backing is bounded, with positive lower bound, and finite upper bound. This means that insurance companies should be neither too weakly nor too

[2] "more clearly and generally" [MC87]

strongly capitalised.

We provide, in Section 2, an overview of the economic valuation framework mentioned above. This framework, known as the *stochastic discount factor* approach, contains CAPM as a special case, and has the additional merit that it can be applied in incomplete markets, and is thus capable of being applied to the valuation of insurance policies from the consumer's point of view.

We finish by discussing the practical implications of the principles developed, and illustrate how information about the feasible range of premiums in regulated insurance markets might be inferred from prices observed in comparable un-regulated markets.

2 Background — Valuation of Risky Assets

Modern equilibrium asset pricing theory is cast in terms of *stochastic discount factors*. In a single-period setting, these are random variables that allow assets to be priced according to formulae like:

$$p = E(m\,x), \tag{1}$$

where x is the payoff of the asset (a random variable), p denotes the price of the asset, and m is the stochastic discount factor (for example see [Coc01, LW01]).

The stochastic discount factor is the marginal rate of substitution between consumption at the end of the period and consumption at the start of the period. *A priori*, different agents may have different marginal rates of substitution, but if assets are traded in a securities market then agents will adjust the quantities they hold, and prices will adjust accordingly, until at equilibrium the marginal rates of substitution will agree. Agents may disagree on the values of assets which are not traded, however, as such assets are not subject to this equilibrium-forming process. This means the agents may have different discount factors, but that the projection of each onto the space of traded assets must be the same, and must equal the "market" discount factor [Coc01].

A set of payoffs and prices is arbitrage-free if and only if there exists a strictly positive stochastic discount factor.

Supposing for now the existence of a risk-free asset, which yields a guaranteed gross return[3] of R_f for an investment of 1, Equation (1) immediately

[3]Whenever we discuss returns in this paper, we shall mean the *gross return*. This is defined as the asset payoff divided by the asset price: $R = x/p(x)$.

yields: $E(m) = 1/R_f$. Applying the definition of covariance, Equation (1) can be re-stated as:

$$p = \frac{E(x)}{R_f} + \text{cov}(m, x). \tag{2}$$

This shows that the price of a risky payoff is calculated as the expected payoff, discounted at the risk-free rate, plus a risk-adjustment term that is of the nature of a co-variance. The discount factor co-varies negatively with consumption, so that assets that tend to pay off in high-consumption states attract a negative risk-adjustment, while those that pay off in low-consumption states will be valued above the risk-neutral value (meaning the first term on the right-hand side of (2)). This conclusion is reminiscent of a conclusion of the CAPM, but is shown to hold in a more general setting. The CAPM is a special case of the stochastic discount factor approach (for example see [Coc01] and [Joh04]).

3 The Insurance Surplus

3.1 Existence of the Insurance Surplus

We will now apply this economic valuation approach to argue for the existence of a surplus that is created by insurance — the difference between the sum of the values placed by consumers upon a portfolio of insurance policies and the market value of this portfolio. For simplicity we will assume a one-period setting, where all the policies in the portfolio pay off at some single time in the future. However, we note that the asset valuation framework described above applies in a multi-period setting also.

Suppose there are K customers who have assets they would like to insure. Suppose the insurance policy of each customer is designed to pay off in states of the world where the assets it is written against have declined in value, and that in such states the customer's consumption would otherwise be low — for example because the asset constitutes a large part of the customer's wealth.

Let X_k be a set of assets (payoffs) accessible to customer k, which includes his insurance policy. Let m_k denote this customer's stochastic discount factor, which prices assets in X_k.

We will assume the existence of a market of traded assets, X, that are accessible to all K customers: $X \subset X_k, \forall k$. This set would include the stocks of insurance companies, and other companies, for example. We also assume that portfolio formation holds in X (so that X is a vector subspace of each X_k), and thus conclude the existence of a stochastic discount factor $m \in X$ that prices assets in X. We also conclude that this *market discount factor*,

m, is the projection of m_k onto the space of traded assets ($m = \text{proj}_X m_k$), which means that each customer will value traded assets at market value. We can thus write the customers' discount factors as $m_k = m + m_k^\perp$, for each k, where $m_k^\perp$ is a payoff orthogonal to X — that is, each customer's discount factor can be decomposed as the sum of the market discount factor, m, and a payoff orthogonal to the space of traded assets, $m_k^\perp$.

Let w_k be the payoff of the insurance policy to customer k. The payoff w_k is negatively correlated with the customer's consumption, and hence positively correlated with his stochastic discount factor, m_k: $\text{cov}(m_k, w_k) > 0$. Thus the value placed on the policy by the customer, v_k, exceeds the risk-neutral value (meaning the value that would be placed upon it by a risk-neutral agent):

$$v_k = E(m_k\, w_k) = \frac{E(w_k)}{R_f} + \text{cov}(m_k, w_k) > \frac{E(w_k)}{R_f}$$

We can't talk about the "market value" of w_k, as this payoff is available only to the owner of the specific assets insured, and can't be replicated through traded assets (assets in X). That is, we can calculate the quantity $E(m\, w_k)$, as m and w_k are both in X_k, but it is not a price (unless w_k has no "idiosyncratic" component, i.e. $w_k^\perp = 0$), as m is a discount factor only on X, not the whole of X_k. In other words, m is a discount factor that represents the market pricing operator, but this operator is only defined for traded assets, and the insurance policy w_k is not traded, so it is not meaningful to talk about its market price.

Suppose the individual policies are aggregated to form an insurance liability. The payout (outflow of cash) by the holder of this aggregate portfolio is the random variable $L = \sum_{k=1}^K w_k$. Let us suppose that this aggregate liability is to be traded in the market. If the payoff L is not in X we will need to augment X with the span of L, and re-define m accordingly. Then the market value of L is:

$$v = E(m\, L) = E\left(m \sum_{k=1}^K w_k\right).$$

This is the price at which investors would be willing to trade the liability L.

We would like to investigate the difference between the value placed upon the policies by customers, and the market value of the aggregate liability. Accordingly, we define the *insurance surplus*, Γ, as:

$$\Gamma = \left(\sum_{k=1}^K v_k\right) - v = \left(\sum_{k=1}^K E(m_k\, w_k)\right) - E\left(m \sum_{k=1}^K w_k\right)$$

For simplicity of notation below we will omit the limits on the sums over k – they should all be taken to be from 1 to K.

Our basic proposition is that for typical insurance risks, the insurance surplus, Γ, will be positive. We can exhibit a number of reasonable sets of assumptions under which this is the case. For example, consider the situation mentioned in the introduction — a "zero-beta" portfolio:

Proposition 1 *If the aggregate insurance liability is uncorrelated with the market discount factor (i.e.* $\mathrm{cov}(m, \sum_k w_k) = 0$*), and customers are averse to the risks they have insured (i.e.* $\mathrm{cov}(m_k, w_k) > 0$ *for all* k*), and a risk-free asset is traded, then* $\Gamma > 0$*.*

Proof: Since $\mathrm{cov}(m, \sum_k w_k) = 0$ we have:

$$\begin{aligned} v &= E\left(m \sum_k w_k\right) \\ &= E(m)\, E\left(\sum_k w_k\right) + \mathrm{cov}(m, \sum_k w_k) \\ &= E(m)\, E\left(\sum_k w_k\right) \\ &= \sum_k E(m)\, E(w_k), \end{aligned}$$

so

$$\begin{aligned} \Gamma &= \sum_k E(m_k\, w_k) - E\left(m \sum_k w_k\right) \\ &= \sum_k \left(E(m_k\, w_k) - E(m)\, E(w_k)\right) \\ &= \sum_k \left(E(m_k)\, E(w_k) + \mathrm{cov}(m_k, w_k) - E(m)\, E(w_k)\right). \end{aligned}$$

Now since there is a risk free asset traded, i.e. there exists an asset in X that has a certain payoff, say R_f for an investment of 1, pricing this with each of m and m_k leads to the conclusion that $E(m) = E(m_k) = 1/R_f$. We thus have:

$$\Gamma = \sum_k \mathrm{cov}(m_k, w_k) > 0.$$

□

In fact we can generalise the above quite readily. Assuming again that a risk-free asset is traded, so that $E(m) = E(m_k)$, one sees that

$$\begin{aligned}
\Gamma &= \sum_k E(m_k\, w_k) - E\left(m \sum_k w_k\right) \\
&= \sum_k \left(E(m_k)\, E(w_k) + \operatorname{cov}(m_k, w_k)\right) - E(m)\, E\left(\sum_k w_k\right) - \operatorname{cov}\left(m, \sum_k w_k\right) \\
&= \sum_k \operatorname{cov}(m_k - m, w_k) \\
&= \sum_k \operatorname{cov}(m_k^{\perp}, w_k) \\
&= \sum_k \operatorname{cov}(m_k, w_k^{\perp})
\end{aligned}$$

If all assets in each X_k are traded, then each $w_k^{\perp}$ is a zero vector, and the surplus is zero. Thus we see that incompleteness of the securities market is necessary in order for a positive surplus to exist. The insurance surplus will be positive if each policyholder is averse to the non-traded part of their insured risk (i.e. $\operatorname{cov}(m_k, w_k^{\perp}) > 0$ for each k).

3.2 Allocation of the Insurance Surplus

It would be interesting to consider other particular circumstances under which $\Gamma > 0$, but for now let us stop and consider the consequences of the existence of a positive insurance surplus.

A simple insurance company could be formed by charging customers some premium, and agreeing to pay out the insurance liability L when it arises. Let's assume for now we set up this company with unlimited liability (i.e. it guarantees to pay the claims) — we will address the effects of limited liability later. We will also ignore expenses and taxes for now.

If we charge customers premiums that correspond in aggregate to the market value of the liability, i.e. v, the surplus Γ will accrue entirely to policyholders. Taking out the insurance policy will be an NPV-positive decision for them, and writing the policy will be an NPV-zero decision for shareholders. The latter is the rule that Myers and Cohn propose be adopted as the characterisation of a "fair" premium.

The other extreme is to transfer the surplus entirely to shareholders, by charging customers $v + \Gamma$. Taking out the insurance policy would then be an NPV-zero decision for customers, as they would be charged an amount that

corresponds with the value they place upon the insurance policy. If customers were charged any more than this they would not purchase the policies, as they would value them at less than cost. Writing the policy would be an NPV-positive decision for shareholders. Shareholder value of Γ would be created through writing the policies.

If we define a *feasible* aggregate net premium level as being one for which the decision to enter into the insurance contract will be NPV-positive for both customers and shareholders, then the range of feasible aggregate net premium levels is from v to $v + \Gamma$. In circumstances in which there exists a positive insurance surplus, this range will contain more than one value. The "fair" premium defined by Myers and Cohn is the least value in the feasible range. Both consumers and shareholders will be willing to enter into insurance contracts at any price within the feasible range. It is not the case, as is claimed in the seminal paper of Myers and Cohn [MC87], that setting prices above the lower end of this range implies a "wealth transfer" from consumers to shareholders. On the contrary, consumers entering into insurance contracts at prices in the interior of the feasible range have their expected utility improved. They are better off ("wealthier") having entered into the insurance contract than they were beforehand. So are shareholders. The implicit assumption of a zero-sum game is not valid in incomplete markets. That's why insurance can be a beneficial enterprise. Of course, consumers would always like to pay even less for the same service, but whether such an opportunity will be available is a matter for competition.

In a more realistic insurance company structure, there will be expenses incurred in running the company, and there will be taxes that need to be paid to governments. In such a setting the feasible range will shrink, as the insurance surplus of the raw liability needs to be large enough to cover these flows of cash to other stakeholders (employees, suppliers, government). We observe consumers buying insurance and shareholders investing in insurance companies, and this can be taken as evidence that an insurance surplus exists at the raw liability level.

3.2.1 Imperfections, Regulation, Competition, "Fairness"

All prices within the feasible range should be acceptable to capital market participants (shareholders), and to participants in the market for insurance policies (consumers). Deciding where premiums will be, or should be, set, will require consideration of the imperfections of each of these markets. Competition in the capital markets is considered strong, but there are acknowledged imperfections that could be relevant to insurance pricing. These may cause shareholders in insurance companies to demand compensation different from

v, which is the stand-alone value of the liability according to some valuation model. The stochastic discount factor approach could be used to analyse the costs of these imperfections — so-called *frictional costs*.

Altering the valuation model, or valuing the frictional costs, will result in the boundaries of the feasible range moving (e.g. employing CAPM might give one value of v, while employing a deflator might give another; and modelling the possibility of financial distress might lower the value that consumers are prepared to pay). As the assumptions underlying a valuation will always be debateable, the boundaries of the feasible range will always be "fuzzy". Market participants should of course seek to improve their valuation methods. However, once a particular valuation approach has been agreed as reasonable, and the feasible range calculated according to this approach, there remains the question of where in the feasible range premiums will be, or should be, set. This will be determined by the level of competition and regulation in the market for insurance policies (the "consumer insurance market"). In an un-regulated monopoly market, premiums will be set at the upper boundary of the feasible range. At equilibrium in a perfectly competitive market, premiums will be set at the lower boundary of the feasible range. In a "typical" or "realistic" market, premiums might be set somewhere in between. Note that in order to actually reach the upper boundary of the feasible region, the insurer would potentially need to tailor premiums specifically to each customer — as to achieve the maximum while retaining all customers they would need to charge each customer the maximum he or she will bear.

One context in which the Myers-Cohn arguments are being debated is regulated insurance markets. There, the debate centres around the theoretical framework to be used by the regulator in deciding whether a premium proposed by an insurance company is fair. Arguments specific to insurance need to be combined with considerations from regulatory economics in order to settle the issue — in particular, the aims of the regulator need to be considered. If the aim of the regulator is to ensure that customers pay the least possible price, then the Myers-Cohn definition of "fair" premium might justfiably be applied. However, the regulator's aims may well be less extreme. For example the Motor Accidents Authority, which regulates Compulsory Third Party (CTP) car insurance in the state of New South Wales in Australia, aims to have a CTP scheme that is affordable, fair and accessible[4]. Industry practitioners assess the fairness and accessibility aims to imply that the regulator strives to ensure that a healthy number of insurers offer CTP services, that these insurers remain financially viable and willing to com-

[4]see the MAA website: `http://www.maa.nsw.gov.au`

mit capital to the CTP business, and that prices are not too volatile. One might reasonably argue that premiums should be similar to those that would be charged in an efficiently-functioning un-regulated market — a "typical" insurance market — so that premiums set somewhere away from the lower boundary of the feasible range might be justified.

Since risk pooling and diversification are fundamental to insurance, there are significant economies of scale, and so the typical market may differ from a perfectly competitive market, characterised by the presence of many firms.

3.3 What's so special about insurance ?

The CAPM, with all its acknowledged imperfections, is a very popular model, applied daily in many corporations across the world to value risky assets and make decisions. So why can't insurance customers apply CAPM to value their insurance contracts ? Let's think about a consumer who owns a home, and assume that this home constitutes a large part of the consumer's wealth. We claim that the consumer will be very averse to fluctuations in the value of this house caused by it potentially burning down, for example, and so will be willing to buy insurance for prices greater than the expected payoff discounted at the risk free rate (the risk-neutral value). If the consumer tried to value the policy using CAPM, however, the risk-neutral value is precisely what the answer would be, presuming the home burning down is un-correlated with the wealth portfolio ("the market"). We could argue about whether the assumptions of CAPM apply in this case, but it is perhaps easier to argue that the conclusions of CAPM don't apply. In particular, CAPM concludes that all investors should hold some combination of the market portfolio and the risk-free asset. They shouldn't have a large part of their wealth tied up in one particular house, for example — instead they should securitise the house, and sell off units in it, and only retain a minuscule proportion of it under their own ownership, and invest the resulting proceeds in a well-diversfied portfolio of assets, including many other houses. What we observe, however, is different from this, and we see that consumers, for reasons beyond the scope of CAPM, indeed have large parts of their wealth tied up in particular assets, and that there are certain risks associated with these assets, such as fire, which no traded security will offset. Consequently, they will be averse to these risks, and will be willing to pay a risk premium for an insurance policy that offsets them. Thus the incompleteness of the securities market is a precondition for insurance to be worthwhile. The shareholders of the insurance company are not exposed to the concentration of risk that the policyholder is, so they require a lesser risk premium to take on that risk. Furthermore the diversification achieved by aggregating many insured risks reduces the

capital the shareholder must commit to support the insurance company.

We are not saying that it is unreasonable to apply the CAPM to value the aggregate insurance liability from the shareholders' point of view, we are just saying that valuing the insurance policies from the policyholders' point of view is beyond the scope of any complete-markets model, such as the classical CAPM.

4 Corporate Structure and the Insurance Surplus

In this section we shall investigate how the feasible range of premiums is affected by the typical structure of a corporate insurance firm. In particular, we shall consider the effects of limited liability, expenses, and taxes.

4.1 Limited liability

We shall consider an insurance company with a single line of business, in a one-period setting. The company sells insurance policies to customers at time $t = 0$, receiving premium of P. The payouts under the policies are uncertain at time zero, but become known at a later point in time, $t = T$, with the aggregate payout to customers being L_T. We will denote the market value of this payout at time zero as L_0.

In this section we will consider the effect of limited liability, in the absence of expenses and taxes. These other factors will be addressed in subsequent sections.

We shall assume that shareholders contribute funding of Q at time zero, so that the total assets of the company at time zero amount to:

$$A_0 = P + Q. \tag{3}$$

We will assume that all these funds are applied to the purchase of financial assets at market value (so we are ignoring for now any other assets the firm may need, such as office equipment etc. — the effect of these will be similar to that of expenses, which will be discussed below). The balance sheet of the company, at market values, thus looks like that shown in Table 1.

The balance sheet balances, i.e.:

$$A_0 = L_0 + E_0, \tag{4}$$

which, using (3), may be re-written as:

$$P = L_0 + (E_0 - Q). \tag{5}$$

Assets			Liabilities
Investments	A_0	L_0	Technical Insurance Liability
		E_0	Shareholders' Equity

Table 1: Opening balance sheet of a simple insurance company (at market values)

Now $E_0 - Q$ is the difference between the market value of the equity claim and the amount of funds contributed by shareholders — so it is the NPV from the shareholders' point of view. Thus the premium is the market value of the insurance liability, plus the NPV that accrues to shareholders.

The Myers-Cohn "fairness" criterion is that the NPV from the shareholders' point of view should be zero, and from (5) we see that this means that premiums should be set at the level: $P = L_0$, i.e. aggregate premium should be set equal to the market value of the insurance liability. As we have mentioned above, the premiums observed in typical insurance markets may be NPV-positive for shareholders, meaning that they are set at something above L_0.

Now let us consider what happens at time T. By this point, the value of the assets will have risen (hopefully) to an amount A_T. If the funds were invested in risky assets, the amount A_T will be a random variable, and may of course be less than A_0. We will assume here that the funds have in fact been invested in risk-free assets, so that:

$$A_T = R_f\, A_0 = R_f\, (P + Q),$$

where R_f is the risk-free return between time zero and time T. This simplifying assumption will allow us to deal with a single random variable — the aggregate amount of claims — but many of our conclusions will carry over to the case where funds are invested in risky assets.

At time T the aggregate amount of claims becomes known, taking the value C, say (a random variable). Policyholders have first claim on the available funds (A_T), with the remainder accruing to shareholders. The balance sheet of the company at time T is thus like that shown in Table 2. We can consider this a balance sheet at market values at time T, or a statement of the cash flows at time T, with the "Assets" representing cash flows into the company ("payoffs"), and the "Liabilities" representing cash flows from the company to stakeholders ("payouts"). From the points of view of the various stakeholders (in this case the policyholders and the shareholders), these payouts represent the payoffs of the claims they hold over the company's assets.

Assets		Liabilities	
Investments	A_T	$L_T = \min(C, A_T)$	Technical Insurance Liability
		$E_T = \max(A_T - C, 0)$	Shareholders' Equity

Table 2: Closing balance sheet of a simple insurance company (at market values)

Let us consider qualitatively the impact of limited liability. In this case, the cash flow to equity at time T is: $\max(0, A_T - C)$, i.e. the down-side is cut off at zero, and so the time-zero market value of the equity claim, E, will be correspondingly increased, compared with the unlimited liability case. So, assuming the level of asset backing remains constant, the introduction of limited liability will increase the value of the equity claim, so that the Myers-Cohn fair premium will be reduced. This makes sense, as in the event that assets are insufficient to meet liabilities at time T, the customers will get back less than they were promised, and so the price they are charged up-front should be reduced.

4.1.1 Structure of the feasible region

To draw out the impact of limited liability we will distinguish between the amounts claimed and the amounts paid under each policy. Let c_k be the amount claimed by customer k, so that the aggregate claim is $C = \sum_k c_k$. As in Section 3, w_k will denote the amount paid to customer k, so that $L_T = \sum_k w_k$. We know that the aggregate payout will be limited by the available funds: $L_T = \min(C, A_T)$, but in order to value the individual policies we would need to specify a scheme for the allocation of any shortfall. We will assume the shortfall between amount claimed and amount paid is allocated to the individual policyholders in proportion to the amounts claimed, so that

$$w_k = c_k \frac{L_T}{C}.$$

We'll write C_0^S for the value the shareholders place on the portfolio of claims, and C_0^P for the sum of the values the policyholders place on their claims:

$$C_0^S = E(m\,C), \qquad C_0^P = \sum_k E(m_k\,c_k).$$

Note that as $L_T \leq C$, $E(m\,L_T) \leq E(m\,C)$ (no-arbitrage), so the value shareholders place on the aggregate liability will be bounded above by the value they place on the claims portfolio (C_0^S). Similarly for the policyholders.

Let us suppose that we are dealing with a set of risks that are insurable, at least under the assumption that the claims are guaranteed to be paid. According to our discussion above this means that the insurance surplus is positive under the assumption that claims are guaranteed to be paid:

$$\Gamma_\infty = C_0^P - C_0^S > 0.$$

We will also assume that the aggregate claim, C, is a random variable taking values between 0 and $C^{\max}$, with probability density function ϕ (the upper bound $C^{\max}$ may be ∞). For simplicity we will assume that ϕ is positive between these bounds, and we will continue to assume that a risk-free asset is traded.

Shareholders pay Q at time zero to receive E_T at time T. The expression for the net present value to shareholders is thus:

$$\mathrm{NPV}^S = -Q + E(m\, E_T).$$

Similarly, policyholders, in aggregate, pay P at time zero to receive L_T at time T. The aggregate net present value accruing to policyholders is:

$$\mathrm{NPV}^P = -P + \sum_k E(m_k\, w_k).$$

We are particularly interested in the contours $\mathrm{NPV}^P = 0$ and $\mathrm{NPV}^S = 0$ in the (P, Q)-plane, as these will define the boundaries of the feasible region.

4.1.2 The shareholders' zero-value curve

Now $E_T = A_T - L_T$ and $L_T = \min(C, A_T) = C - D_T$, where we write D_T for the shortfall or deficit between what is claimed and what is paid:

$$D_T = \max(C - A_T, 0).$$

Hence $E_T = A_T - C + D_T$, so

$$E(m\, E_T) = E(m\, A_T) - E(m\, C) + E(m\, D_T) = A_0 - C_0^S + E(m\, D_T),$$

which says that equity value is the value of the assets, less the value of the claims, plus the value of the deficit, the latter arising from the option of shareholders to walk away in the event of insolvency. Now $A_0 = P + Q$ (from (3)), and D_T depends on P and Q only through the asset backing $A_T = R_f\,(P + Q)$, so

$$\mathrm{NPV}^S = P - C_0^S + E(m\, D_T(A_T(P, Q))), \tag{6}$$

The contour $\mathrm{NPV}^S = 0$ has equation:

$$P = C_0^S - E(m\, D_T(A_T(P,Q))). \tag{7}$$

We'll show that this is a contour that passes monotonically as a function of A_T from the origin of the (P,Q)-plane towards the vertical line $P = C_0^S$. First note that if $A_T = 0$, $D_T = C$, so $E(m\, D_T) = C_0^S$, so this contour passes through $A_T = 0, P = 0$, i.e. $P = 0, Q = 0$.

To see that the contour is monotonic, suppose $a > b$ and $0 < a, b < C^{\max}$. Then $D_T(b) - D_T(a)$ is a non-negative random variable which is positive with positive probability: $D_T(b) \geq D_T(a)$ for all C and $\mathrm{Prob}(D_T(b) > D_T(a)) > 0$, so the absence of arbitrage requires it have positive value, i.e. $E(m\, D_T(b)) > E(m\, D_T(a))$. This shows that the value of the deficit is a strictly decreasing function of asset backing. As the asset backing A_T becomes large, the value of the deficit approaches zero, or becomes zero at $C^{\max}$ if $C^{\max}$ is finite.

To infer the slope of the contour, we differentiate the right-hand side of Equation (6) with respect to P and Q, and thus calculate the gradient vector of NPV^S, which gives us the direction of steepest ascent for the shareholders' NPV. Now

$$\frac{\partial}{\partial P}\mathrm{NPV}^S = 1 + \frac{d}{dA_T}\, E(m\, D_T(A_T))\frac{\partial A_T}{\partial P},$$

and

$$\frac{\partial A_T}{\partial P} = R_f.$$

Writing

$$E(m\, D_T(A_T)) = \int_{A_T}^{C^{\max}} m(c)\,(c - A_T)\,\phi(c)\,dc,$$

we see that:

$$\frac{d}{dA_T}\, E(m\, D_T(A_T)) = -\int_{A_T}^{C^{\max}} m(c)\,\phi(c)\,dc = -P(C \geq A_T)\, E(m|C \geq A_T).$$

This expression increases monotonically from $-1/R_f$ at $A_T = 0$ to zero at $A_T = C^{\max}$ or as A_T becomes large if $C^{\max}$ is infinite. Hence

$$\frac{\partial}{\partial P}\mathrm{NPV}^S = 1 - R_f\, P(C \geq A_T)\, E(m|C \geq A_T)$$

is zero at $A_T = 0$ and increases monotonically to one as A_T becomes large. Proceeding in a similar fashion for the derivative with repect to Q we find that

$$\nabla \mathrm{NPV}^S = \begin{cases} \begin{pmatrix} 0 \\ -1 \end{pmatrix} & \text{if } A_T = 0 \\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{if } A_T \geq C^{\max} \end{cases}$$

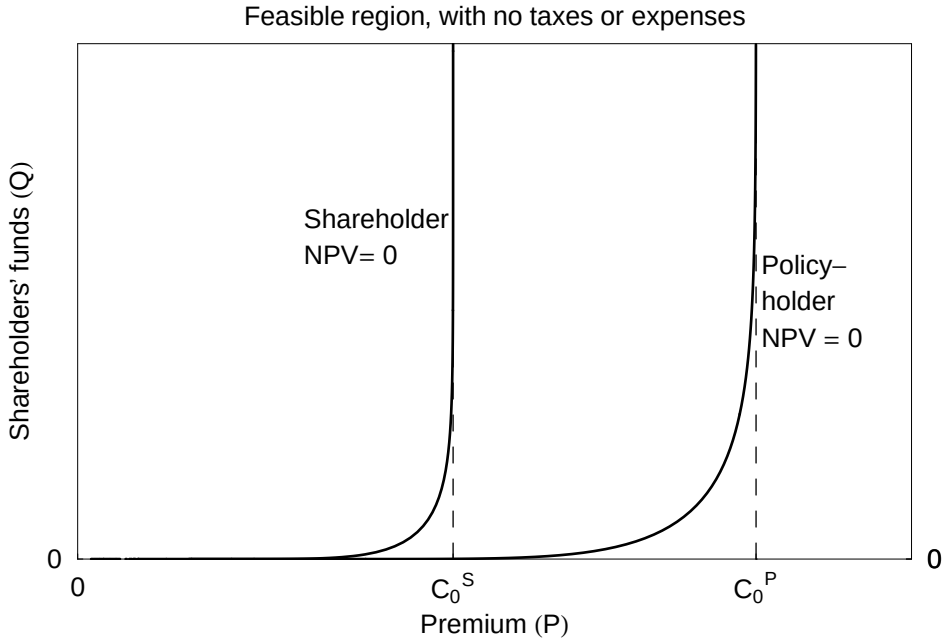

Figure 1: Zero-value contours for shareholders and policyholders, with limited liability, but no taxes or expenses. The dashed lines are the asymptotes of these curves as $Q \to \infty$. The feasible region is the region containing premiums greater than those on the shareholders' zero-value curve and less than those on the policyholders' zero-value curve.

with the gradient moving monotonically beween these extremes. In particular we see that the shareholders' zero-value curve is horizontal at the origin. The curve is thus as shown in Figure 1.

4.1.3 The policyholders' zero-value curve

We will write d_k for the shortfall or deficit of policyholder k, $d_k = c_k - w_k$. Then

$$d_k = c_k \left(1 - \frac{L_T}{C}\right) = \begin{cases} 0 & \text{if } C \le A_T \\ c_k \left(1 - \frac{A_T}{C}\right) & \text{if } C > A_T \end{cases}$$

so d_k is non-negative, and depends on P and Q only through A_T. Since $w_k = c_k - d_k$ we have:

$$\text{NPV}^P = -P + \sum_k E(m_k\, c_k) - \sum_k E(m_k\, d_k).$$

The first sum here is the sum of the values placed by policyholders on their claims, which we have denoted by C_0^P. Hence the equation of the contour

$\text{NPV}^P = 0$ is:

$$P = C_0^P - \sum_k E(m_k\, d_k(A_T(P, Q))). \tag{8}$$

We can view this expression as describing the $\text{NPV}^P = 0$ contour in the form $P = f(A_T)$, and since each d_k is non-negative, the contour lies not to the right of the vertical line $P = C_0^P$ (policyholders will not pay more for the insurance than they think their claims are worth).

As above, we can show that value of the deficit is lower at higher levels of asset backing — taking $d_k(A_T)$ to mean the random variable d_k for the particular level of asset backing A_T, one can see that if $a > b$ and $0 < a, b < C^{\max}$ then $d_k(b) - d_k(a) \geq 0$, for any (positive) values of c_k and C, and that $\text{Prob}(d_k(b) - d_k(a) > 0) > 0$, and assuming m_k is strictly positive, so that policyholder k prices in an arbitrage-free manner, this means that the value of $d_k(b) - d_k(a)$ is positive also, i.e. $E(m_k\, d_k(a)) < E(m_k\, d_k(b))$, for each k, and hence $\sum_k E(m_k\, d_k(a)) < \sum_k E(m_k\, d_k(b))$. This means that the policyholders' aggregate value of the deficit is a strictly decreasing function of A_T while A_T is less than the maximum possible claim. As the asset backing A_T becomes large, the value of the deficit approaches zero, or becomes zero at $C^{\max}$ if $C^{\max}$ is finite.

If the asset backing A_T is zero, no claims are paid, $d_k = c_k$ for each k, and so $\sum_k E(m_k\, d_k) = \sum_k E(m_k\, c_k) = C_0^P$, and so the right-hand side of Equation (8) is zero, which shows that the $\text{NPV}^P = 0$ contour passes through the point $P = 0$, $A_T = 0$, which in terms of P and Q is $P = 0$, $Q = 0$.

The feasible region, being the set of combinations of aggregate premium, P, and shareholders' funds, Q, for which entering into the insurance arrangement will be NPV-positive for both shareholders and policyholders, is thus as shown in Figure 1.

It might be helpful at this point to consider why the feasible region has the shape shown in Figure 1. From the policyholder perspective, there is a maximum premium they are prepared to pay. This is an increasing function of asset backing, because increasing asset backing reduces the likelihood of the insurer defaulting on the payment of claims, but it is bounded above by the value the policyholders place on the claims. From the shareholder perspective, there is a minimum premium for a given level of initial asset backing which provides sufficient return for the risks to which the shareholders' capital is exposed. As the asset backing decreases, this minimum level decreases, because the shareholders' option to walk away in the event of insolvency becomes "in the money". The feasible region of premiums is bounded below by requirements for shareholder returns and above by the maximum that policyholders are willing to pay.

4.2 The impact of expenses and taxes

A corporate insurance firm will have various cash outflows other than the payment of claims. For example, we might typically consider customer acquisition expenses, which are paid at the time the policy is written, and are known, and claims handling expenses, which are paid at the time the claim is paid, and are unknown (random). Also, taxes will need to be paid on any profit made by shareholders, either underwriting profit, or profit arising through capital gains on the financial assets held. Thus modified cash flow statements at time zero and time T are as shown in Tables 3 and 4. We have assumed here that expenses are the highest-priority claim on the assets at time T, being paid before policyholders' claims.

Cash Out			Cash In
Investments	A_0	P	Aggregate Premium
Customer Acquisition Expenses	X_0	Q	Shareholders' Funds

Table 3: Opening cash flow statement of a simple insurance company, with taxes and expenses

Cash In			Cash Out
Investments	A_T	X_T	Claims Handling Expenses
		L_T	Claims Paid
		G_T	Tax Paid
		E_T	Equity Cash Flow

Table 4: Closing cash flow statement of a simple insurance company, with taxes and expenses

We will now consider the impact on the feasible region of the inclusion of taxes and expenses.

4.2.1 The impact of expenses

In our simplified one-period model, we could model these expenses as a known cash outflow of X_0 at time zero, and a stochastic cash outflow of X_T at time T. If the claims handling expenses are some particular function of the aggregate amount of claims, C, we could handle this in our model without introducing any further fundamental uncertainties by specifiying $X_T = f(C)$. For simplicity, however, we will assume that the claims handling expenses are

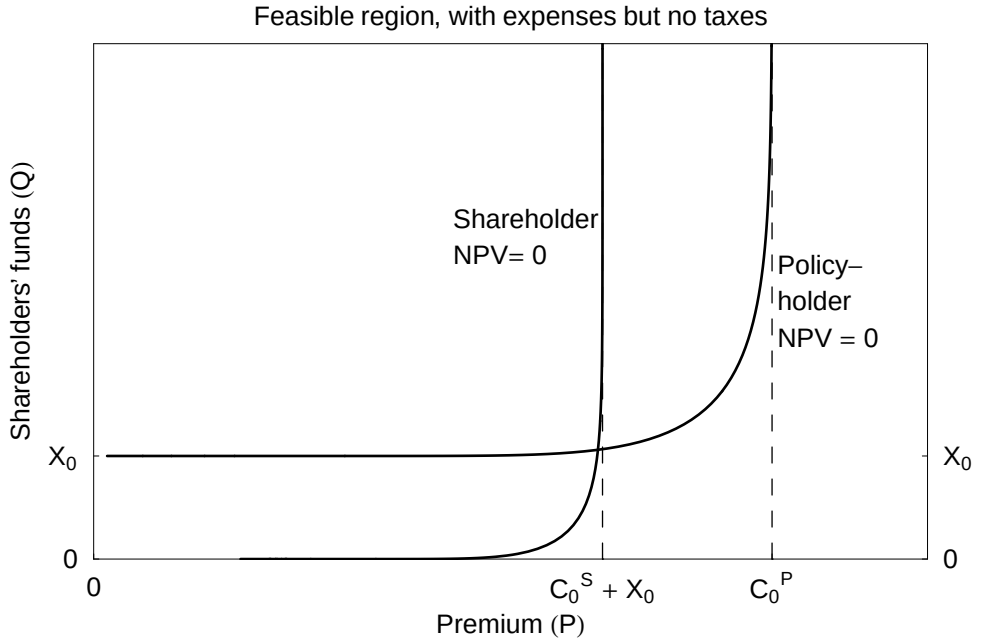

Figure 2: The feasible region, with limited liability and expenses, but no taxes.

known also, and then we can incorporate their value into X_0 by discounting at the risk-free rate.

The only impact on our model, then, is on the expression for the initial asset backing — rather than all of the premium and shareholders' funds being invested in financial assets, some of these funds must first be used to pay the expenses:

$$A_0 = P + Q - X_0.$$

Consequently, $A_T = R_f\,(P + Q - X_0)$. Referring to expression (8), we see that the equation of the contour $\mathrm{NPV}^P = 0$ is the same as before, but with Q replaced by $Q - X_0$. The contours of NPV^P will thus be translated upwards by X_0 in the (P, Q)-plane. Similarly we find that the contours of NPV^S will be translated to the right by X_0. The feasible region is thus as shown in Figure 2. Points to note include:

- The lower bound for the amount of funding required is now positive. Policyholders will not deal with the insurer unless the insurer has sufficient capital to pay the expenses.

- Inclusion of expenses causes the feasible region to diminish in size, and if expenses are sufficiently large, the feasible region will be wiped out.

The critical value is $X_0 = \Gamma_\infty$, i.e. expenses equal to the insurance surplus in the unlimited liability case.

4.2.2 The impact of corporate income tax

We will assume tax is paid at time T, at the rate τ on profit and capital gains. We'll take the underwriting profit to be the premium less the claims paid, less the expenses: $P - L_T - X_0$. The capital gain is $A_T - A_0$. The total "income" is therefore:

$$P - L_T - X_0 + A_T - A_0 = A_T - L_T - Q.$$

We will assume tax is only paid if this quantity is positive, so the tax paid is:

$$G_T = \max\left(0, \tau\left(A_T - L_T - Q\right)\right).$$

We may write this as a function of the amount claimed, as follows:

$$G_T(c) = \begin{cases} 0 & \text{if } c \geq A_T - Q, \\ \tau\left(A_T - Q - c\right) & \text{if } 0 \leq c < A_T - Q. \end{cases}$$

The NPV from the policyholders' point of view does not change, but the NPV from the shareholders' point of view is reduced by the value of the tax paid. The value of the tax paid is:

$$\int_0^{A_T(P,Q)-Q} m(c)\,\tau\left(A_T(P,Q) - Q - c\right)\phi(c)\,dc.$$

By calculating the gradient of this quantity and subtracting it from the gradient of NPV^S calculated previously, we see that:

$$\nabla \mathrm{NPV}^S = \begin{cases} \begin{pmatrix} 0 \\ -1 \end{pmatrix} & \text{if } A_T = 0 \\ \begin{pmatrix} 1-\tau \\ -\tau\left(1 - \frac{1}{R_f}\right) \end{pmatrix} & \text{if } A_T \geq C^{\max} \end{cases}.$$

Accordingly, the contour $\mathrm{NPV}^S = 0$ is no longer asymptotically vertical, rather it slopes up and to the right (as long as $R_f > 1$). The feasible region is thus as shown in Figure 3. Points to note include:

- The set of feasible funding levels is now bounded above. If the firm is too strongly capitalised, the amount of tax paid will be sufficient to wipe out the insurance surplus
- Inclusion of taxes causes the feasible region to diminish in size, as shareholders require additional premium to compensate them for having to pay tax, while policyholders gain no corresponding benefit.

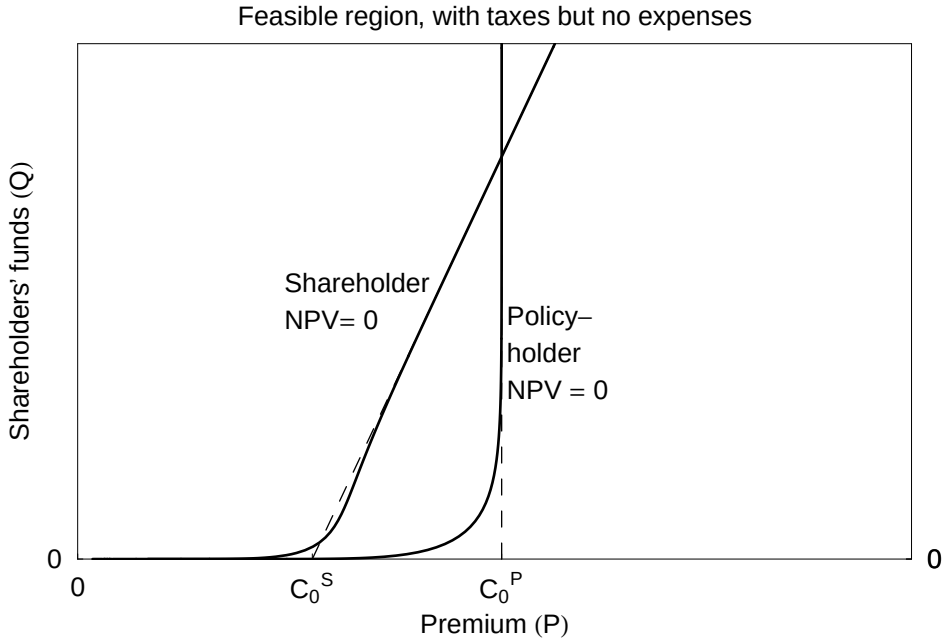

Figure 3: The feasible region, with limited liability and taxes, but no expenses.

4.2.3 The feasible region under limited liability, expenses and taxes

A real insurance firm with have limited liability, expenses, and taxes, and if we combine the results of the previous sections we see that the feasible region will look like that shown in Figure 4. Points to note include:

- The set of feasible funding levels is now bounded both above and below. The lower bound arises from the fact that policyholders require a positive amount of asset backing, after expenses have been paid, in order to derive any value from the insurance contract. The upper bound arises from the fact that taxes wipe out the insurance surplus if the company is too strongly capitalised.

- These upper and lower bounds on feasible funding levels can of course be expressed as upper and lower bounds on the probability of ruin.

We reiterate that the existence of an insurance surplus at the raw liability level, which arises from the incompleteness of securities markets, is a necessary condition for the existence of a feasible combination of premium and shareholders' funds — for otherwise there would be no capacity to pay the

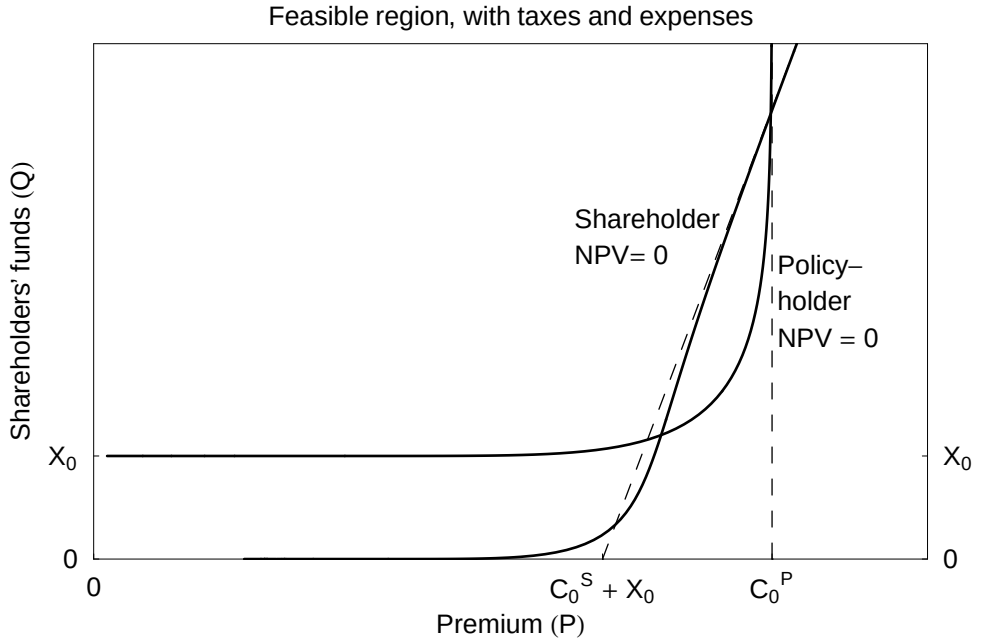

Figure 4: The feasible region, with limited liability, expenses and taxes.

taxes and expenses that arise in the corporate form of the insurance enterprise.

The model we have presented does not contain all the features of a real insurance company — in particular, funds are often invested in risky assets, rather than risk-free assets, so this is an effect that should be considered, but it would involve introducing a new fundamental uncertainty, being the return on these risky assets, so we shall defer consideration of this matter. We expect that the broad conclusions we have drawn regarding the existence and shape of the feasible region will persist to the case of risky assets.

4.3 Mutuals

We have discussed above the corporate form of the insurance enterprise, in which the owners and customers of the firm are two distinct groups, though they may have some intersection. It will be instructive to consider another form of the insurance enterprise — a "mutual" — in which the owners and customers are the same group. We will consider a simple unlimited-liability mutual, and we'll leave tax at zero for simplicity. Using the same notation as above, at time zero policyholders pay an aggregate amount P of premium, with policyholder k paying p_k. Expenses of X_0 are paid, and the balance $A_0 = P - X_0$ is invested in risk-free assets, accumulating to $A_T = R_f\, A_0$ by

time T. At time T claims totalling C are paid, with customer k receiving c_k. If the claims are less than the amount of assets available, the surplus $A_T - C$ is distributed to policyholders in some proportion. Let us assume it is distributed in proportion to premiums paid by each customer. Likewise, if the claims exceed the assets available, the policyholders must contribute the deficit $C - A_T$. Again we will assume that the deficit is distributed in proportion to premiums paid by each customer. Accordingly, the cash flow to customer k at time T is:

$$c_k + \frac{p_k}{P} \min(A_T - C, 0) - \frac{p_k}{P} \min(C - A_T, 0),$$

being full payment of the customer's claim, plus his share of any surplus, less his share of any deficit. This simplifies to:

$$c_k + \frac{p_k}{P} (A_T - C).$$

The net present value for policyholder k is then:

$$\begin{aligned} \mathrm{NPV}_k^P &= -p_k + E\left(m_k \left(c_k + \frac{p_k}{P}(A_T - C)\right)\right) \\ &= -p_k + E(m_k\, c_k) + \frac{p_k}{P}\frac{A_T}{R_f} - \frac{p_k}{P} E(m_k\, C). \end{aligned}$$

Now $\frac{A_T}{R_f} = A_0 = P - X_0$ so we have:

$$\mathrm{NPV}_k^P = E(m_k\, c_k) - \frac{p_k}{P} X_0 - \frac{p_k}{P}\ E(m_k\, C).$$

This says that the net present value to customer k is the value he attributes to his claim, less his share of the expenses, less the value to him of his share of the aggregate claim. Now if the aggregate claim is traded, each customer will value it the same way, so $E(m_k\, C) = E(m\, C)$ for all k. Summing over customers we thus obtain the aggregate NPV to policyholders as:

$$\mathrm{NPV}^P = \sum_k \mathrm{NPV}_k^P = -X_0 + \sum_k E(m_k\, c_k) - E(m\, C).$$

This is the value of the insurance surplus in the unlimited-liability, zero-expense case (Γ_∞), less expenses. Comparing this with Figure 2, we see that this is the width of the feasible region at very high capitalisation levels. It is the maximum NPV which could be achieved by policyholders buying insurance from a corporate insurance enterprise. This NPV would be achieved if premiums were set at the left hand edge of the feasible region.

This shows that the only way policyholders could be happier with a corporate insurer over an unlimited-liability mutual is if it has a lower level of expenses. Or, if they prefer it for factors unrelated to the valuation of risk, for example if it provides better service in the event of an accident, or more convenient payment options, *etc.*, in other words if the insurance it offers is a more attractive consumer product.

The argument can be generalised to the limited-liability case by "replicating" a limited-liability mutual by assuming that the policyholders contribute all of the capital to the corporate insurer, and thus capture all of the surplus as either policyholders or shareholders. Of course, not every potential policyholder has the available capital to participate in an unlimited-liability mutual, or to purchase shares in a corporate insurer. This will limit the extent to which mutuals can meet the needs of all potential insurance customers, and implies a need for at least some part of the market to be serviced by corporate insurers. Another interesting observation that can be made from Figure 2 is that policyholders could achieve greater NPV from a strongly capitalised corporate insurer than they would at a sufficiently weakly capitalised mutual, even if they pay higher premiums in the former case. Even though policyholders do not capture the entire surplus when the corporate form is used, they may capture a larger absolute amount of value as the insurance surplus is an increasing function of asset backing.

These considerations support our conclusion that insurance pricing will be determined by competition in the market for insurance policies. This competition should drive cost savings, which would allow premiums to reduce over time. If mutuals could readily be formed at low cost, this possibility would keep a lid on insurance premiums, and limit the extent to which shareholders could capture a large share of the insurance surplus by keeping premiums high, all other things being equal. Instead they must compete on costs and service in order to create shareholder value, just as participants in other industries must. The current trend seems to be away from mutuals and towards a prevalence of corporate insurers, and many successful mutuals have de-mutualised in recent times. This apparent success may indicate that the corporate form is a more effective vehicle for driving cost savings and service improvements, and providing less restricted access to capital.

5 Applying these principles using specific valuation methods

In this section we consider how the principles set out in this paper might be applied. We first consider their application using a risk-adjusted discount rate approach, and then consider the use of the stochastic discount factor approach.

5.1 Risk-adjusted discount rate approach

5.1.1 Applying the CAPM from the shareholders' point of view

We will consider the unlimited liability case, with no expenses or taxes. To value the insurance liability using CAPM, one needs to determine a beta for this liability, β_L, then calculate the risk-adjusted discount rate for the liability, $E(R_L)$, using the CAPM formula, then estimate the expected future cash flow of the liability, $E(L_T)$, and finally value the liability via:

$$L_0 = \frac{E(L_T)}{E(R_L)}. \tag{9}$$

There are two ways that β_L is typically determined (application of each of these methods is illustrated in the paper of Myers and Cohn [MC87]). One can estimate it directly; this requires an estimate of the covariance of the liability return with the "market" return. We could obtain this either through examination of return data, or via a thought experiment (though it's hard to see how the latter would give anything other than just the sign of the beta, or a conclusion that it is about zero).

Alternatively, if one feels that the betas of the other components of the balance sheet are more easily estimated, one can exploit the fact that beta, as a function of returns, is linear, and apply this to Equation (4) to yield:

$$A_0\,\beta_A = L_0\,\beta_L + E_0\,\beta_E. \tag{10}$$

This is just a weighted average cost of capital formula, and suffers from the circularity of such formulae when applied in a corporate valuation context — while A_0 is known, and E_0 could be determined using the Myers-Cohn fairness criterion, $E_0 = Q$, the remaining needed weight is L_0, the quantity we are trying to calculate. This can be overcome via iteration, however, so that Equation (10) can be used to calculate β_L from β_A and β_E. Another way of expressing the Myers-Cohn fairness criterion is to say that the market-to-book ratio for equity should be one: $E_0/Q = 1$. If one wanted to calculate

the liability beta consistent with a greater market to book ratio, one could simply take E_0 in (10) to be some other multiple of Q.

The β_L produced from (10) will only ever be a beta for the aggregate insurance liabilities of an insurance company, as there is only one equity beta to start from, whereas the direct method has the advantage that it could yield different betas for each line of business.

5.1.2 Reverse-engineering pricing parameters from premiums charged

If an insurance company is contemplating entering or creating a new market, it will need to consider whether the returns it could obtain in that market are as large as the returns it will require. The latter alone do not determine a price that can be charged — one also needs to take account of what the market will bear. One way to do this for a particular liability class would be to "reverse-engineer" pricing parameters from known prices for comparable insurance portfolios. These comparable portfolios could be the same line of business at different companies, or in different geographical markets, or they could be different but similar lines of business.

For example, if we know the expected payout, $E(L_T)$, and the aggregate premium charged to customers for a particular portfolio of insurance policies, P, we can infer the implied risk-adjusted discount factor, $E(R_L)$, from (9) and (5). Note that in doing this we cannot distinguish between the two components on the right-hand side of Equation (5) — we are in effect solving the equation:

$$P = L_0 + (E_0 - Q) = \frac{E(L_T)}{E(R'_L)}, \tag{11}$$

for a risk-adjusted discount rate $E(R'_L)$. This discount rate will include compensation to shareholders for taking on the expected payouts of the insurance liability, and the systematic risk associated with these, as well as any excess return permitted to them by the competitive structure of the particular consumer market they are operating in. That is, we are solving for a "realistic" liability value L'_0, which consists of the "theoretical" or "perfect-markets" liability value L_0, plus a term, $E_0 - Q$, that in the theoretical or perfect-markets world would represent value creation for shareholders, but in reality may consist in part of other factors such as compensation for systematic modelling risk, strategic risk, and frictional costs. It could however contain a degree of value creation that is permitted by the competitive conditions in the consumer insurance market being examined.

5.1.3 Competitive and regulatory considerations

Let us consider the implications for regulated markets of setting insurance premiums based purely on a securities market view of the insurer's balance sheet. Generalising Equation (5) to the case where expenses exist (for simplicity we'll leave taxes at zero), we obtain:

$$P = L_0 + X_0 + (E_0 - Q). \tag{12}$$

As pointed out above, the Myers-Cohn fair premium is obtained by setting $E_0 = Q$ in this equation, while an insurer's wish to maintain a higher market-to-book ratio would correspond to setting E_0 to be some greater multiple of Q. Our arguments will apply to either case, so let's assume that some particular market-to-book ratio is judged "fair".

Suppose a particular insurer innovates and reduces expenses through efficiency (so their X_0 comes down). Then the "fair" premium for that insurer would come down. If regulation was based on a requirement to charge the "fair" premium, the insurer would be required to reduce their premiums as soon as the innovation was made. In a typical un-regulated market, a firm which innovates and reduces costs would in fact be able to maintain prices at or near former levels for a period of time, perhaps some years, and thus create additional shareholder value, until other firms caught up and prices were driven down by competition. One might consider it reasonable to allow such an outcome in a regulated market as well, as the consumer would benefit in the medium term from the cost-reduction incentives thus created.

For another example, suppose the managers of regulated portfolios decide to increase their salaries. This would increase X_0 and hence increase the "fair" premium, and if the regulatory regime merely requires the "fair" premium to be charged, the decision would be self-funding from the shareholders' point of view, but would obviously be detrimental to consumers. Such an increase in expenses could even push premiums out of the feasible region, above the levels that consumers consider value-creating, but in a compulsory class they would still be required to pay.

These examples show that the regulator must consider more than just a definition of fair premium based on market values of the insurer's balance sheet components — they should consider whether the premiums are value-creating for consumers, they should have some means of assessing whether expenses are reasonable, and they should impose a regulatory regime which encourages insurers to innovate and reduce costs. Characterising the pricing from the policyholder's point of view, in comparable un-regulated consumer insurance markets, as we have described above, might be one way to do this.

5.1.4 Choice of calculation method

The method described in the Section 5.1.2 amounts to agreeing a "fair rate of return" for the particular portfolio being priced. The trouble with expressing the method in terms of rates of return, i.e. risk-adjusted discount rates, or betas, is that these should not be expected to be stable across similar products. For example, the beta of an option over a stock is in general different from the beta of the stock, and options with different strike prices will have different betas, and the betas will vary according to state of the world (e.g. they will be different at the various nodes of a binomial tree used to price an option). That is, they vary not just with the underlying risk, but with the form of the claim held over that risk. This is why formulations such as risk-neutral probabilities, or deflators, or stochastic discount factors are preferred in financial option pricing — the same stochastic discount factor will price all derivatives over an asset. If a form for the stochastic discount factor over the underlying risk is assumed, with certain free parameters, these parameters will not depend on the particular conditions of the policy, such as different levels of excess, or different caps. It also means that parameters inferred from re-insurance quotes are comparable to the parameters of the underlying portfolio. This stability means stochastic discount factors could be a better basis for characterisation of "fair returns" than betas or discount rates are. Furthermore, as the CAPM typically involves an assumption that returns are normally distributed, it cannot correctly capture the effect of variations in the probability of ruin, or situations where the underlying aggregate claim distribution is not normal, whereas the stochastic discount factor approach can. However, the stochastic discount factor approach suffers from the problem that there is little theoretical guidance as to what form of stochastic discount factor should be used in any particular situation. It also requires more information about the claim payments, needing the full distribution, rather than just the expected payoff. It remains to be seen which particular calculation method represents the best compromise between accuracy and applicability. As a first step, we now illustrate how the stochastic discount factor method could be applied.

5.2 An illustrative application of the stochastic discount factor approach

We will consider an example to illustrate how the reverse-engineering method described in the previous section could be applied, and to make the concepts discussed in the section on the feasible region a little more tangible. To map a realistic insurance situation onto our one-period model, we will assume

the risk-free rate is known, and imagine we have simulated future claims, and discounted them back to the present day at the risk-free rate, to obtain a distribution of a random variable that we will call the risk-free present value of claims — or in the typical terminology, the distribution of inflated and discounted claims. As we only have a single-period model, we'll write this random variable as C/R_f, where R_f is the risk-free return over the period, and the length of the period is some sort of average duration. We'll take the length of the period to be 3.7 years, and the risk free return to be $R_f = (1.06)^{3.7} = 1.2406$.

As the pricing operator is linear, by assumption, our conclusions will be scale-invariant, so we will assume the distribution of C/R_f has mean of one, or equivalently that all quantities are expressed as multiples of the expected claim (as a risk-free present value).

We shall assume that C has a log-normal distribution, with coefficient of variation of 24%. Claims administration expenses are often assumed to be a percentage of claims, and so their risk-free present value will be that percentage of C/R_f, as present values are linear in cash flows. We will assume they are known, and express them as a percentage of expected claims. We shall take this ratio to be 5% here. We shall assume customer acquisition expenses are also expressed as a percentage of expected claims, and assume this value is 12%. Accordingly, we have total expenses as $X_0 = 0.17\,E(C/R_f) = 0.17$. We shall take the effective tax rate to be $\tau = 25\%$.

In asessing the value of their insurance contracts, policyholders will apply their individual discount factors m_k to their individual insurance claims c_k. In this first application we wish to simplify the analysis so we will assume that the preferences of policyholders can be captured by an "aggregate discount factor" m^P, which is applied to payoff of the portolio of claims, C. There will be multiple combinations of values of the c_k which lead to any particular value of C, and in general these combinations may not be valued in aggregate by policyholders in the same way. In that case $m^P|C = c$ can perhaps be thought of as some sort of average discount factor over those states of $\{c_k\}$ which add up to c. The situation we are dealing with thus looks like that shown in Figure 5.

In order to proceed we need to specify the form and parameters of the stochastic discount factors, as functions of the value of C. For illustrative purposes, we shall assume that both the market and policyholder aggregate discount factors have the power-law form:

$$m_\lambda(c) = \frac{1}{R_f}\, e^{\frac{-\lambda^2}{2}}\, e^{\frac{\lambda\,\mu}{\sigma}}\, c^{\frac{-\lambda}{\sigma}},$$

where μ and σ are the mean and standard deviation of the normal distribution

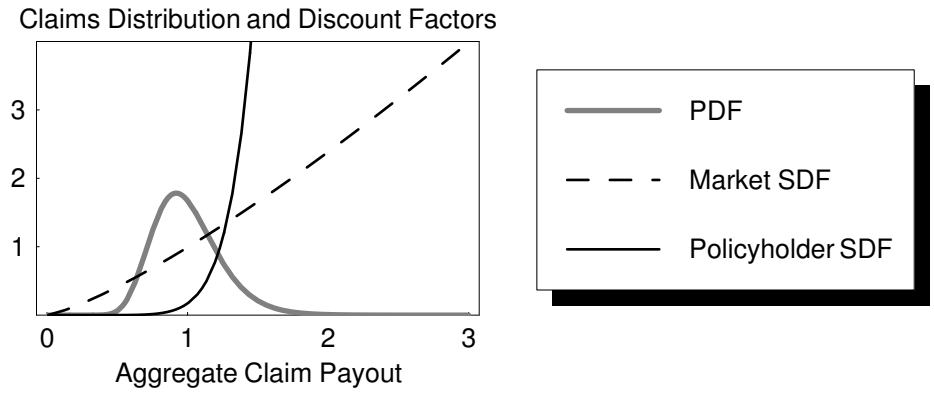

Figure 5: Aggregate claim payout distribution, along with the policyholders' aggregate discount factor (m^P) and the shareholders' discount factor (m).

underlying C (i.e. $C = e^{(\mu+\sigma Z)}$, where Z is a standard normal variable)[5]. Here λ is a parameter (which Wang [Wan02] calls the *market price of risk*). We will assume that the market discount factor m is given by this form with parameter λ^S, and the policyholder aggregate discount factor m^P has this form with parameter λ^P. The existence of the insurance surplus in the unlimited liability case is then equivalent to the condition $\lambda^P < \lambda^S$. The condition that policyholders are averse to the risks insured amounts to $\lambda^P < 0$. We will assume that λ^S is known, and equal to -0.3. In summary, our assumed parameter values are:

$$\mu = 0.1876, \qquad \sigma = 0.2366, \qquad X_0 = 0.17, \qquad \tau = 0.25, \qquad \lambda^S = -0.3$$

Given this information, we can draw the contour $\text{NPV}^S = 0$. If we knew the value of λ^P, we could draw the contour $\text{NPV}^P = 0$ and thus construct the feasible region. We don't know the value of λ^P, but we will show how information about λ^P can be inferred from observed prices. First we observe that as λ^P decreases (becomes a larger negative number), the contour $\text{NPV}^P = 0$ moves to the right on our (P, Q)-plane. Suppose there are multiple insurers offering prices in the insurance market, and that the products offered are identical, except that the different insurers may have different levels of financial strength (as measured by the probability of ruin). Suppose the following combinations of premium and probability of ruin are observed: $(1.47, 0.1\%)$, $(1.43, 0.5\%)$, and $(1.41, 2\%)$. We can solve for the

[5]This particular form of the discount factor can be derived by assuming that C satisfies the assumptions of Wang's specialisation of Buhlmann's equilibrium model [Wan03, Joh03, Joh04]

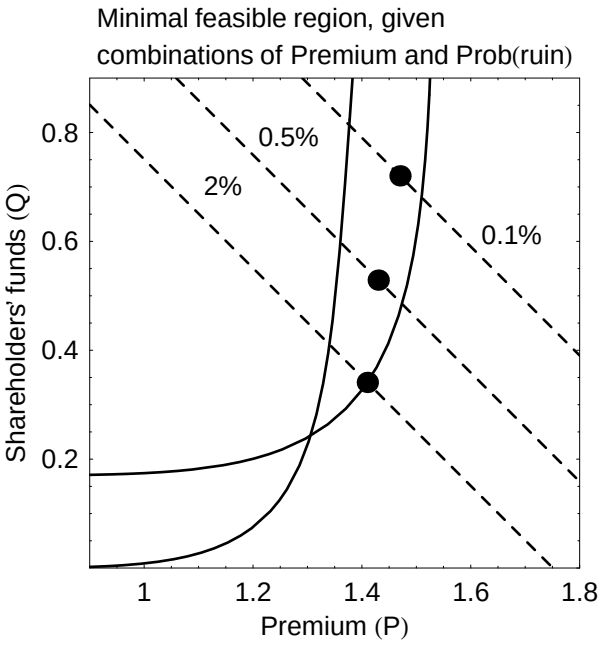

Figure 6: The NPV-zero contours at the maximum feasible value of λ^P consistent with observed combinations of premium and probability of ruin. The dashed lines are the contours of total funding corresponding to the specified probabilities of ruin, and the dots are the points on these lines corresponding to the specified premiums.

greatest value of λ^P that makes these three points feasible. The solution in this case turns out to be $\lambda^P = -1.81$. This situation is shown in Figure 6. We see that the bound on λ^P is in this case determined by the combination $(1.41, 2\%)$. As the other two combinations lie on higher value contours (to the left of the contour on which the $(1.41, 2\%)$ point lies), we deduce that policyholders prefer them, despite the premiums being higher.

Care would need to be taken to capture the correct scaling of λ^P with portfolio size, and with the duration of the liability. The true value of λ^P might be lower again, but this approach would yield values for λ^P that correspond to prices charged in real insurance markets, taking account of competitive pressures. In drawing a single $\text{NPV}^S = 0$ curve we have assumed that all three insurers have the same levels of expenses. In reality their expense levels may differ, so they might have their own specific $\text{NPV}^S = 0$ curves which are translations to the left or right of that shown.

6 Conclusions

We have argued that an insurance surplus can arise due to the incompleteness of securities markets, and that the existence of this surplus implies the existence of a range of insurance premiums that are value-creating for both shareholders and policyholders. We have shown how this range of feasible premiums is affected by the typical corporate form of the insurance enterprise, and how it varies with the level of asset backing. We have shown that for given, positive, levels of expenses and taxation, the range of feasible levels of asset backing has a positive lower bound and a finite upper bound.

In developing our conclusions we have employed the stochastic discount factor approach, because of its generality, and its applicability to incomplete markets.

Potential further developments include treatment of more of the features of a typical corporate insurance firm, such as investment in risky assets, and multiple lines of business. As real insurance companies exist in a multi-period setting, it would be useful to extend the approach to account for this, and to see whether any qualitative changes arise. Further investigation into the quantitative application of the approach would also be instructive.

Another potential avenue of exploration would be to work directly from the utility functions of individuals who are potentially both policyholders and shareholders. Such an approach is described in [Tay95], where Taylor sets out an equilibrium model of insurance pricing and capitalisation. The model accounts for insurer default risk, but takes taxes and expenses as zero. Under these assumptions equilibrium insurance prices, insurer capital levels and stock prices can jointly be determined. It would be interesting to extend that type of approach to the situation described herein, in which the frictional costs of insurance are significant, to see if any light could be shed on appropriate allocation of the insurance surplus, i.e. where within the feasible region prices and capitalisation should emerge.

Since securities markets are incomplete with respect to the risks being insured, one cannot infer fair — or even feasible — pricing of insurance by looking at securities markets alone. One can view the market for insurance policies (the consumer insurance market) as "completing" the market for these risks — it is thus a primary source of pricing information. Since the frictional costs of this form of risk financing (such as expenses and taxes) are relatively high, they should ideally be incorporated into any analysis. The presence of high "frictional costs" leads us away from a view of insurance as a pure financial product, and towards a view of insurance as a consumer product, with the expenses being treated as production costs, and with products being differentiated by factors such as the quality of service. Insurers have

the opportunity to create value by competing on costs and service, just as participants in other industries must.

Acknowledgements

The author would like to thank Tim Jenkins for many helpful discussions during the development of this paper, Tony Coleman for suggesting the area of research and providing helpful references, and Insurance Australia Group for sponsoring the research.

References

[Act00] Taylor Fry Consulting Actuaries. Capital and profit in CTP insurance. NSW Motor Accidents Authority Issues Paper, May 2000.

[Coc01] John H. Cochrane. *Asset Pricing*. Princeton University Press, Princeton, New Jersey, 2001.

[Joh03] Mark E. Johnston. New directions in risk modelling. IAA Workshop on Risk Modelling in Banking and Finance, April 2003.

[Joh04] Mark E. Johnston. The stochastic discount factor for the exponential-utility capital asset pricing model. Working Paper, Department of Actuarial Studies, University of New South Wales, 2004.

[Lee01] Audrey Lee. Re: Capital and profit. Letter to David Bowen, General Manager of the Motor Accidents Authority of NSW, September 2001.

[LW01] Stephen F. LeRoy and Jan Werner. *Principles of Financial Economics*. Cambridge University Press, Cambridge, United Kingdom, 2001.

[MC87] Stewart C. Myers and Richard A. Cohn. A discounted cash flow approach to property-liability insurance rate regulation. In J. D. Cummins and S. E. Harrington, editors, *Fair Rates of Return in Property-Liability Insurance*, pages 55–78. Kluwer Nijhoff Publishing, Dordrecht, 1987.

[Mol02] T. Moller. On valuation and risk management at the interface of insurance and finance. *British Actuarial Journal*, 8(IV):787–827, 2002.

[Tay95] Greg Taylor. An equilibrium model of insurance pricing and capitalization. *The Journal of Risk and Insurance*, 62(3):409–446, 1995.

[Wan02] Shaun S. Wang. A universal framework for pricing financial and insurance risks. *ASTIN Bulletin*, 32(2):213–234, 2002.

[Wan03] Shaun S. Wang. Equilibrium pricing transforms: New results using Bühlmann's 1980 economic model. *ASTIN Bulletin*, 33(1):57–73, 2003.