© *The Journal of Risk and Insurance*, 2006, Vol. 73, No. 1, 123-151

Can Bonus-Malus Alleviate Insurance Fraud?

Ignacio Moreno
Francisco J. Vázquez
Richard Watt

Abstract

Traditionally, insurance companies attempt to reduce (or even eliminate) fraud via audit strategies under which claims may be investigated at some cost to the insurer, with a penalty imposed upon insureds who are found to report claims fraudulently. However, it is also clear that, in a multiperiod setting, bonus-malus contracts (increases in subsequent premiums whenever a claim is presented) also provide an incentive against fraud. In this article, we consider a model in which, conditional upon the client renewing his contract, the *only* mechanism used to combat fraud is bonus-malus. In this way, our model provides the opposite pole to the pure audit model. We show that in our simplified setting there exists a bonus-malus contract that will eliminate all fraud in all periods, while guaranteeing nonnegative expected profits to the insurer and participation by the insured. We also consider the dynamics of the solution, the effect of an increase in risk aversion on the solution, and the welfare implications.

Introduction

Whenever one party to a contract unilaterally holds information concerning the true value of a relevant parameter, then we say that an adverse selection problem exists. This contrasts with moral hazard problems, in which the holder of the private information not only unilaterally observes a relevant quantity, but he can also alter its value. Hence, moral hazard concerns unilateral information concerning a *variable* that is controlled by the informed party, while adverse selection concerns unilateral information concerning a *parameter*. It is also important to distinguish between the point of time at which the information asymmetry occurs in relation to the date at which the contractual relationship is formalized. We say that the asymmetry is "*ex ante*," when it has already occurred when the contract is signed, and "*ex post*" when the contract is signed under conditions of symmetric information, but when an asymmetry will occur later on. In this article, we consider a situation in which, under an assumption

Ignacio Moreno, Francisco J. Vázquez, and Richard Watt are from Dept. Análisis Económico, Fac. C.C.E.E., Universidad Autónoma de Madrid, Ctra. Colmenar km. 15, Madrid 28049, Spain.
The model used in the article is based on Moreno (2003) and Moreno, Vázquez, and Watt (2002).

of symmetric information on the probability of each state of nature, an insurance contract between the insurer and the insured is signed in the knowledge that during the course of the contract a loss that is only observed by the insured may occur. Thus, we are considering a case of *ex post* adverse selection, in which insurance fraud in the sense of misrepresenting the true loss amount may occur.[1]

Under this interpretation of insurance fraud, it is interesting to note that the solutions that are traditionally posed to combat the adverse selection are very different from the solutions that are posed in other economic models of adverse selection. Traditionally, a principal is assumed to resolve an adverse selection problem by ensuring that costs (usually risk-bearing costs) are placed upon the "good" agent type, in order that the "bad" agent type finds no advantage in passing himself off as being of the good type. Such a solution is typically known as "self-selection," since when used, it implies that the agent (the informed party) will voluntarily act in such a way that he discloses the information that he possesses. On the other hand, the theory of insurance fraud has always taken the stance that the way in which an insurer should combat fraud is by using an audit scheme under which claim reports may be investigated to discover if they are fraudulent or not (see Picard (2000) for a survey of the economic theory of insurance fraud). In short, the audit mechanism will detect[2] fraud with a given probability, and in this case, some type of (financial) penalty can be levied upon the insured. In this way, an insured who misreports a claim enters into a lottery which can be made to have a lower expected utility than the option of honest reporting. Thus fraud can be eliminated, although at some cost to the insurer (the cost of maintaining the audit mechanism) if the misreporting lottery is made worse than the option of reporting truthfully. In such a model, it is straightforward to see that in fact the insurer may not be interested in eliminating all fraud, since there exists a strictly positive cost of reducing fraud (the audit cost)—indeed, the optimal audit scheme will equate the marginal cost and the marginal income[3] of fraud.

A solution to any adverse selection problem can always be posed using some type of learning mechanism under which the principal may gather further information, at some cost to himself, concerning the value of the relevant parameter. However, this type of solution is not invoked in traditional adverse selection models. On the other hand, it is also quite puzzling that a traditional solution (self-selection) has not ever been considered for insurance fraud problems. In this article, we introduce a very simple model of insurance fraud as an *ex post* adverse selection problem, and we show that a self-selection solution exists under which no fraud is ever committed.

The model can be thought of as the opposite pole to the pure audit model, in which self-selection is not used as a solution mechanism. In the model, conditional upon the

[1] We do not consider any other type of fraud, for example, deliberate accidents.

[2] By "detect" we mean that sufficient evidence suggesting that fraud has been committed has been gathered, for an impartial third party (e.g., a law court) to pass a verdict declaring that indeed fraud has been committed.

[3] The marginal income from fraud will be measured in terms of fraudulent claims that will not be paid, plus any financial penalty that can be collected from insureds who are found to report fraudulently.

client remaining in the contractual relationship, no audits at all are carried out, and so the only mechanism used to combat fraud is the self-selection mechanism. Naturally, the most logical scenario is that neither polar model is appropriate, and instead a mixture of both mechanisms will be optimal, with each mechanism serving to temper the severity required of the other. However, before considering the optimal strategy of the insurer as a mixture of all possibilities, we must firstly establish that a solution under self-selection is indeed feasible.

Clearly, in the pure audit model the costs implied by the existence of the option of committing fraud, compared to a situation in which fraud is not possible, are borne by the insurer. However, in a multiperiod setting, a second option exists based on premiums that involve bonus-malus (see Lemaire, 1985, 1995, for a full actuarial treatment of this type of contract in automobile insurance). If an insured's future premiums are increased whenever a claim is presented, there exists a direct incentive against presenting fraudulent claims. Assuming that no auditing whatsoever is used, then bonus-malus implies that a fraudulent claim report will provide current positive surplus (over and above the option of honest reporting) to the insured at the cost of future negative surplus (assuming that he continues with the insurance contract). In principle, it may be possible to design a bonus-malus contract that provides sufficient future costs for all fraud to be eliminated at no cost to the insurer.[4] Showing that indeed this is possible, and analyzing the dynamics of the solution, pointing out certain comparative statics results, and considering the welfare implications of the solution, is the modest objective of this article.

It is important to clarify that in this article, we are not attempting to model any currently existing real world contractual environment. As far as we are aware, bonus-malus contracts are not used anywhere as a mechanism to address insurance fraud. The objective of the article is to put forward the idea that a certain degree of premium uncertainty in the bonus-malus sense can be of aid for reducing insurance fraud, and thus that perhaps it is worthwhile to consider including bonus-malus aspects into the antifraud strategy of an insurer. We use a very simple model to present this idea, since hopefully the more transparent is the presentation of the mechanism under which bonus-malus combats *ex post* adverse selection, the clearer will be the underlying message of the article.

The article continues as follows: the following section outlines the basic model together with all of the relevant assumptions, and the self-selection solutions for both a perfectly competitive and a monopoly scenario are presented. In the next section we discuss the dynamics of the solution, and in the subsequent section, we consider the effect of an increase in risk aversion upon the solution. The penultimate section contains the welfare implications of the self-selection mechanism as compared to a pure audit model, while the final section concludes and offers some general directions for future research. The arguments presented in the article are supported by a strong graphical orientation. However, the most important results are set aside as theorems, lemmas, etc., but all mathematical proofs are relegated to the appendix.

[4] Note that, since an experience rating scheme implies that each individual's lottery remains dependent upon his own outcome, it must violate the mutuality principle (see Borch, 1962), and hence cannot be first best efficient. It may, however, be second best efficient.

THE MODEL, ASSUMPTIONS, AND STATIC SOLUTION

In order to highlight the incentive effects that bonus-malus type premiums can have for insurance fraud, we present a highly simplified theoretical model. The objective is not to explicitly model any particular real-world situation, but rather to attempt to make as clear as possible our underlying message—that by discriminating premiums according to claims history, an incentive against fraud is provided that can imply the reduction of reliance upon auditing claims, with the obvious cost savings for the insurer.

Concretely, the assumptions that we make are the following.

1. **The insurable risk.** We assume a representative individual, with utility function for money of $u(x)$ which is assumed to be strictly increasing and concave. We assume that the individual has, at the beginning of any given period, an endowed wealth of w, and an endowed binomial lottery defined by a loss of L with probability p. Naturally, we assume $L \leq w$. The lotteries in each period are independent, and the outcomes are private information to the individual.[5] The endowed expected utility of the individual in each period is given by

$$Eu(w, L) = pu(w - L) + (1 - p)u(w) \equiv \bar{u}.$$

2. **Insurance coverage.** In any given period, the individual may insure his loss lottery with an insurance company. We shall be interested in the case when the insurer acts in a perfectly competitive market, as well as the case of a monopoly insurer. For simplicity, we assume that all insurance contracts have full coverage, that is, in the case of a loss, the indemnity is equal to L. Of course, this is equivalent to assuming that, at all times, the premium for an indemnity of x in the case of an accident is given by $px + k$ where k is a constant, in which case, it is well known (and very easy to prove) that for all risk-averse individual, partial coverage is never optimal.[6] It is also true that for this type of premium, full coverage is both efficient and profit maximizing so long as the insurer is risk neutral. The full coverage assumption is not really restrictive in the model, since it only contemplates the existence of a single good (money). That is, although we shall describe the model in terms of premiums, the entire model can be cast equivalently in terms of the insured's wealth, in which case differences between coverage levels and premiums paid are all accumulated in a single variable—state contingent wealth.
3. **Discounting.** The intertemporal nature of the model demands that we explicitly deal with how the insured thinks about the future when he decides upon an optimal action in the present, that is, how the individual converts future utility into present value terms. Indeed, how the insured individual thinks about the

[5] More complex lottery environments can be introduced at a considerable cost of mathematical complexity, but without altering the underlying results of the model in any significant way. It should also be noted that the most obvious application of the model is to first party insurance.

[6] Risk averse consumers will demand full coverage if k is not greater than the risk premium corresponding to the endowed situation, and no coverage otherwise (for an elementary treatment of this result see Eeckhoudt and Gollier, 1992). In the model of this article, the option of no-coverage is eliminated using a participation constraint.

future is the key ingredient to getting a bonus-malus system to control for fraud. In order to keep the model simple, and to a certain extent realistic, we make the following assumptions with respect to discounting. We assume that, in any given period i, the utility of period $i+1$ carries a discount factor of 1, while all periods beyond $i+1$ have a discount factor of 0. In words, the assumption is that, when considering any present choices, the agent considers present utility and follow-on period utility to be equally valuable, and does not weigh any further utilities at all. Aside from the obvious mathematical simplifications that this assumption provides, it can also be thought of as a rudimentary manner in which bounded rationality can be brought into the model. Naturally, since we are considering bonus-malus contracts with a binomial risk in each period, starting off at period i, in period $i+1$ there are two possible options to take into account (the situation following a claim, and that following a claim-free period), therefore, in period $i+2$ there are 4 options, and so on.[7] Surely, taking into account all options is far beyond the ability of any reasonable individual. On the other hand, taking present period and next period utility at equal value is also a reasonable approximation in an infinite period world.[8] It is important to note that this simple assumption on discounting *does not* imply that the individual thinks that his life time or the number of insurable periods in the future, is limited to a single future period. Indeed, as we shall now point out, that the individual explicitly recognizes that his insurable future is far longer than a single period is also crucial to our model.

4. **Anticipation of honesty in follow-on period.** Not-withstanding our rather extreme assumption on discounting, we assume that the insured fully understands that the insurance relationship goes beyond the two periods that are contemplated at any given moment of choice. Thus, our insured individual fully understands in period i that his problem will indeed be replicated as a new choice problem in period $i+1$, even though with (possibly) different parameter values. Therefore, period $i+1$ is understood to be both the second of a pair of periods (where the current period, i, is the first of this pair of periods) and the first of a subsequent pair of periods (where period $i+2$ is the second of that pair). Now, it is natural to assume that the insurance relationship must, at some currently unknown future date, become void (possibly the insured asset disappears, is sold, or the insured individual shifts away from that country, or the insurance company closes down, etc.), and so there will be a final period at some time—we just do not know exactly when (although we do assume that it is "sufficiently" distant in time for it to be validly neglected in present value terms). Thus the insured understands that at some point in the distant future, say period T, his contract will not be renewed (for reasons that are entirely exogenous to the working of the contract itself), and

[7] In general, when a decision in period i is considered, there are 2^n possible situations in period $i+n$ for $n = 1, 2, \ldots, \infty$.

[8] Again, the entire model can be recast with a more traditional assumption on discounting, for example, exponential discounting. However, the mathematical complexities then make most of the analysis intractible, and only serves to mask the underlying message that the model is trying to send. In fact, the discounting assumption in the model can be thought of as a very rudimentary approximation to exponential discounting, so the results that our discounting assumption yields can also be generated in a model with full exponential discounting that sufficiently approximates the current model.

so in period T he will be subject to an audit such that he is necessarily honest in this period.[9] So the bonus-malus scheme must keep him honest in period $T-1$, assuming honesty in T. But then we are valid in assuming honesty in $T-1$ when choices are made in period $T-2$, and so on. Using this type of backward induction, we are justified in assuming that when contemplating what to do in period i (commit fraud or not), our insured individual anticipates that in period $i+1$ he will be honest.

5. **Bonus-malus premiums.** Since the loss lotteries are discrete (either a loss of L occurs, otherwise the loss is 0), and since only the individual observes the outcome of the loss lottery (that is, whether a loss occurred or not), an opportunity for fraud exists whenever the outcome is that no loss occurs, in which case the individual may fraudulently report a loss to the insurer. We assume that, whenever a client commits to renew his contract, the insurer never audits any claims, but rather simply pays out on any claim that is reported. However, in the period following a claim, the insurer will discriminate the premium demanded of the insured according to whether or not a claim was reported. As an endogenous result in our model it turns out that reporting a claim will result in a higher premium in the follow-on period than not reporting a claim.[10] Hence, reporting a fraudulent claim results in a present benefit to the insured at the cost of a higher future premium. The insurer's objective is to design the premiums in period $i+1$ such that fraud will not be committed in period i, for all i. Thus, as opposed to the pure audit model where the contract balances the utility of the uncertain lottery that the client receives by presenting a fraudulent claim against the certain situation of not acting fraudulently, here we balance the certainty scenario of fraud (high current income but high future premium) against the no-fraud option yielding lower current income and a lower future premium. We assume that in the case of perfectly competitive environments, while the true outcomes of the loss lotteries are private information for the insured, his claims, history is common information over all insurers, so that there is no benefit to be gained by leaving one insurer for another.[11]
6. **Commitment.** Our assumption on discounting is not costless. Recall that the incentive not to commit fraud is the future costs to the individual due to the higher premium track that subsequently occurs. On one hand, by keeping the next period discount rate at 1, we are effectively imposing the maximum penalty that our restricted model allows, but on the other hand, we are eliminating from

[9] Given a sufficiently distant perception of T the present value of this audit for the insurer is negligable, and indeed it is not calculable with any degree of certainty. Thus we ignore this cost in any current period.

[10] Exactly how each of the two possible follow-on premiums compare with the current period premium will depend on the parameters of the model.

[11] Again, this is very common practice in Spain, and in Europe in general, as well as in other countries. Indeed, in New Zealand a strong marketing factor for private insurance is that a client changing companies retains any bonus premium that has accummulated from earlier contracts with other companies. Thus information sharing on client's claim histories is done there also. Claims information sharing makes it hard (maybe impossible) for clients to "game the system," by placing a claim with one insurer, and then leaving to start anew with a new insurer later, and is clearly beneficial to all insurers.

consideration the future steam of utilities that the individual can obtain by either continuing with the contract or canceling it. Whenever the expected premiums in future periods are fair, then continuing the contract in each of these future periods offers expected utility that is greater than the alternative (the reservation utility). Thus, the short-time horizon that we assume in the model comes at the cost of having to make some sort of assumption regarding commitment to continue by the insured. Extending the time horizon that the insured takes into account would have the effect of allowing us to relax any commitment assumptions on behalf of the client (assuming that the discount rate is sufficiently high) but would also complicate the calculations required enormously.[12] Concretely, we take the following assumption, which seems to be the most reasonable. At the beginning of any given period *i*, the client pays the premium corresponding to the current period and then decides whether or not to sign the renewal contract for the follow-on period. This renewal contract will stipulate two possible premiums, one that corresponds to the contingency that no claim is presented in the current period (period *i*), and another that corresponds to the contingency that a claim is presented in the current period, and signing it commits the client to paying the premium that corresponds at the beginning of the follow-on period. Effectively then, our assumption is that there is a one period commitment on behalf of the client,[13] or that in order to break the contract, the client must inform the insured of this decision one period in advance, as is often the case in real-world contracts.[14]

7. **Non-renewal clauses.** A necessary condition for a sequence of bonus-malus contracts to self-police insurance fraud is that the insured always commits to continue at least one more period. Indeed, if the decision is ever not to continue, then there can be no direct punishment for fraud in the current period. The only options are to increase the time horizon of the insured (so that canceling the contract has the implied current discounted opportunity cost of having no insurance from then on), or to introduce an exogenous policing mechanism that provides sufficient disincentive for a client who does decide to cancel to never commit fraud in his

[12] In any case, there are examples of insurance contracts where the insured does indeed face a legal committment to continue. For example, in order to circulate with a car in most countries, a minimum of insurance coverage (that is typically subject to experience rating) must be purchased to cover third party damages.

[13] An alternative to assuming committment is to assume that at the outset a bond is posted which is held by the insurer against any unpaid premium (that is, to cover for any precommitted contract that is reneged upon).

[14] Certainly this is the case in Spain, and most other European countries. Insurance premiums are normally paid by contractually allowing the insurer to place a direct debit from the insured's bank account. If a stipulated period notification of cancellation (normally one month) is not given, the insurer has the contractual right to charge the premium for the renewal of the contract (normally a yearly contract) directly from the insured's bank account. Thus claiming and exiting from the contract becomes more difficult for the insured. Naturally, this lock-in is relative, since the insured can simply ensure that his bank account balance is at 0 when the time to renew comes along for the premium not to be paid, but clearly doing this implies other costs. For example, even if the balance in the account that was supplied for premium payments is reduced to 0, if the insured has other accounts open at the same bank with positive balances, then the premium payment can be differed to those other accounts. So to avoid the premium, the insured would have to change banks entirely.

final insured period. We opt for the simplest possible mechanism, which is that a nonrenewing client is always fully audited in his last period (assuming a claim is presented).[15] We simply assume that the audit that is used is sufficient for the client not to claim fraudulently in his last period. However, as will be noted below, the credible commitment by the insurer to this audit will be sufficient for no clients to ever want to cancel their contracts as an endogenous best response, and so outside of exogenous occurrences no such audits will ever need to be carried out.

Naturally, individuals that really suffer an accident will suffer the consequences of the bonus-malus contract as if they had acted fraudulently. Given this, the bonus-malus contract must be designed under several reasonable conditions. Firstly, it must be incentive compatible (which we take as meaning that it must provide sufficient incentive that fraud is eliminated in all periods).[16] Secondly, it must ensure participation by the insured in the sense that he would rather enter, and continue with, the contractual relationship than to not insure (thereby retaining his endowed expected utility in all periods). Thirdly, in order that the concept of insurance itself is present, the contract must be such that if the individual legitimately suffers an accident he will still rather present a claim than not (we refer to the decision to not report an accident that really happened as "type-2 fraud"). Finally, of course, the scheme must offer nonnegative expected profits to the insurer, in order that she also is willing to participate.

The problem facing the insurer can be set up according to a constrained optimization framework. In particular, each of the constraints will be discussed in greater detail in order that the feasible set for all problems can be found. In all cases, we assume that the current period premium is q_0, and in the following period the premium will be $q_1(q_0)$ if a claim is reported in the current period, and $q_2(q_0)$ otherwise. In this way, whether or not a claim is reported in the current period, the situation in the next period is formally identical, but where the new q_0 will be either q_1 or q_2. Given this, we shall concentrate on solving for any particular general period, and then clearly the analysis can always be replicated in the next period, and so on.

The Insured's Participation Constraint

We take the insured's participation constraint to require that the individual will agree to commit himself to continue the contractual relationship by signing the contract offered for coverage in the follow-on period. If the insured agrees to continue with the contract, he will commit to a premium of q_1 if he presents a claim in the current period, and q_2 otherwise. Given that we are assuming that follow-on period utility is not discounted, and assuming that no fraud is committed in the current period when accepting to renew (this will be guaranteed by the incentive compatibility constraint)

[15] Other options are also possible. For example, at the outset of the contract the client could be made to deposit a sum of money of L with the insurer, and from this sum the insurer deducts the premium that would have ensued had the client renewed his contract, as a type of contract cancellation fee, and returns the excess to the nonrenewing client.

[16] Really, it is not clear that the insurer really would like to eliminate all fraud, for many reasons. For example, Crocker and Morgan (1998) show that allowing some fraud can be of aid in alleiviating a traditional adverse selection problem.

accepting the renewal awards expected utility of $u(w - q_0) + [pu(w - q_1) + (1 - p) \times u(w - q_2)]$. On the other hand, if, for any reason, the insured decides not to continue with the contract, his current period claim will be audited with sufficient frequency (and/or penalty for fraud) so that no fraud will be committed in this case. Hence, the decision not to renew yields an expected utility of $u(w - q_0) + \bar{u}$. Therefore, when making the decision to commit to the follow-on or not, the insured will always agree to the follow-on if

$$u(w - q_0) + [pu(w - q_1) + (1 - p)u(w - q_2)] \geq u(w - q_0) + \bar{u}.$$

That is,

$$P(q_1, q_2) \equiv pu(w - q_1) + (1 - p)u(w - q_2) \geq \bar{u}. \tag{1}$$

Note that, since we require that the expected utility of the continuation be never less than the single period reservation utility, the (credible) threat of an audit if the contract is not renewed is sufficient for no such audit to ever be carried out. Therefore, given that Equation (1) will be satisfied, these audits imply no costs to the insurer.

The Insurer's Participation Constraint

The insurer will participate (offer a contract renewal) if doing so does not result in negative expected profits. Consider the follow-on period premiums. Assuming incentive compatibility (see below), with probability p the client will have suffered an accident in the current period, and hence will have presented a claim and so will face the higher premium, q_1. In the same way, with probability $1 - p$ he will face the lower premium, q_2. Hence, given that the expected value of the indemnity (recall, we are assuming full coverage) to the insurer of any client in any period is just $-pL$, in order for the expected profits that the contract offers in the follow-on period to be nonnegative, we require[17]

$$p(q_1 - pL) + (1 - p)(q_2 - pL) \geq 0 \quad \Rightarrow \quad pq_1 + (1 - p)q_2 \geq pL. \tag{2}$$

Finally, in the very first period we have $q_1 = q_2 = q_0 \geq pL$. We define

$$B(q_1, q_2) = pq_1 + (1 - p)q_2$$

so that the expected profit condition[18] is simply $B(q_1, q_2) \geq pL$.

The Type-1 Fraud Incentive Compatibility Constraint

The incentive compatibility constraint ensures that no fraud is committed in the current period. In order to correctly consider the incentive constraint, we must assume

[17] Note that we are assuming that no cross-subsidization over clients occurs.

[18] Again, we are ignoring all costs of any audits. There will be no audits for endogenous reasons (since the insured's participation constraint is satisfied), and audits for exogenous reasons are simply assumed to imply a negligable cost in the present since they are sufficiently in the future, and since they are also discounted by an accumulated probability of occurance.

that no accident has occurred in the current period, which is the only situation in which fraud is possible. Hence, the incentive compatibility constraint requires that the expected utility of not having an accident in the current period and not reporting one is not less than the expected utility of not having an accident and yet reporting one. Recall that, given the way in which the problem repeats itself from one period to the next, we only need to ensure that fraud in the current period is eliminated assuming honesty in the follow-on period. This is because, once the follow-on period arrives, it will itself be the current period of a new problem, and hence the replication of the incentive compatibility constraint will ensure that in that period no fraud will be committed.

Recalling that by assumption we have full coverage, in this case, the expected utility obtained in the current period if the insured does not present a fraudulent claim is simply $u(w - q_0)$. Since no claim is presented, the follow-on period premium will be q_2, yielding a follow-on period expected utility of $u(w - q_2)$, where once again, we are assuming honesty in the follow-on period. On the other hand, if (once no accident has occurred) the individual presents a claim in the current period, he will get current period utility of $u(w - q_0 + L)$ and follow-on period utility of $u(w - q_1)$. Thus, the incentive compatibility constraint can be expressed as

$$u(w - q_0) + u(w - q_2) \geq u(w - q_0 + L) + u(w - q_1).$$

We rewrite this constraint as

$$I_1(q_1, q_2) \equiv [u(w - q_2) - u(w - q_1)] - [u(w - q_0 + L) - u(w - q_0)] \geq 0. \qquad (3)$$

The Type-2 Fraud Incentive Compatibility Constraint

Recall, type-2 fraud is defined as having suffered a current period accident but not reporting it. Once again, type-2 fraud is only an issue once an accident has occurred in the current period. The required constraint is

$$u(w - q_0) + u(w - q_1) \geq u(w - q_0 - L) + u(w - q_2).$$

The left-hand side of the constraint measures the current period utility once an accident has occurred and has been reported, plus the follow-on expected utility at the high premium. The right-hand side measures the current period utility of having an accident and not reporting it plus the follow-on period utility at the low premium.

We write the relevant constraint as

$$I_2(q_1, q_2) \equiv [u(w - q_2) - u(w - q_1)] - [u(w - q_0) - u(w - q_0 - L)] \leq 0. \qquad (4)$$

The Feasible Set

Using Equations (1–4), we can now go on to identify the feasible set for both a perfectly competitive insurer and a monopolist. While this can be done mathematically, here we shall restrict our attention to the graphical analysis in the plane (q_1, q_2). Begin with the two participation constraints, (1) and (2).

Firstly, consider the insurer's participation constraint, (2). By the implicit function theorem, the slope of any particular contour of the function $B(q_1, q_2)$ is given by

$$\left.\frac{dq_2}{dq_1}\right|_B = -\frac{p}{(1-p)},$$

that is, each contour is a decreasing linear function. Furthermore, since the value of B is increased by increases in either q_1 or q_2, the contours that are further from the origin correspond to greater values of expected profits. One particular contour will correspond to the value $B = pL$, that is, the set of points that satisfy the nonnegative expected profit condition is the (strictly convex) set of points on and above the contour $B(q_1, q_2) = pL$. This particular contour intersects the 45° certainty line at the point $q_1 = pL$, and touches the horizontal axis at the point $q_1 = L$, and the vertical axis at the point $q_2 = (\frac{p}{1-p})L$.

Secondly, consider the insured's participation constraint (1). From the implicit function theorem, we get the result that the slope of any particular contour of $P(q_1, q_2)$ is

$$\left.\frac{dq_2}{dq_1}\right|_P = -\frac{pu'(w-q_1)}{(1-p)u'(w-q_2)} < 0.$$

Furthermore, it turns out that the contours of $P(q_1, q_2)$ are negatively sloped and strictly concave.[19] The contours have slope of exactly $-\frac{p}{(1-p)}$ at the point where they cross the line defined by $q_2 = q_1$.

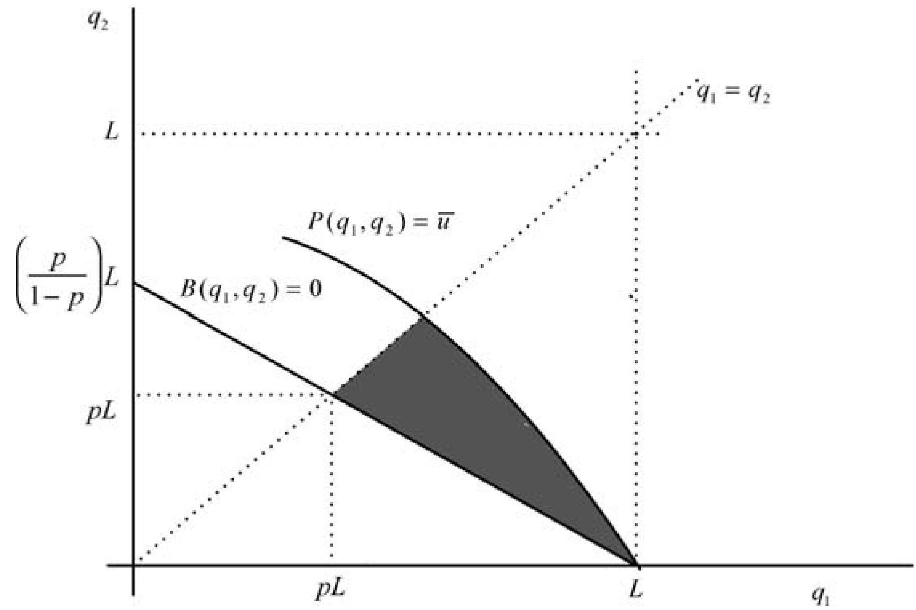
Clearly, since $P(q_1, q_2)$ is decreasing in both of its arguments, the participation constraint (1) is simply the (strictly convex) set of points that lie on and below the contour $P(q_1,q_2) = \bar{u}$. This particular contour touches the horizontal axis at the point $q_1 = L$, that is, exactly the same point at which the contour $B(q_1, q_2) = pL$ touches the horizontal axis, and it touches the line $q_2 = q_1$ at the point $q_1 = q_{ce}$, where $w - q_{ce}$ is the individual's certainty equivalent wealth. Since the individual is strictly risk averse, we know that $q_{ce} > pL$. The set of points that satisfy, simultaneously, the two participation constraints is shown in Figure 1.

Now, consider the incentive compatibility constraints. Consider firstly the type-2 fraud constraint, (4). This constraint saturates along $I_2(q_1, q_2) = 0$. It can be readily checked from the implicit function theorem that this contour of the restriction is a strictly upward sloping function. For any given value of q_0, this contour passes

[19] This is because the second derivative of $\frac{dq_2}{dq_1}|_P$ is

$$\left(-\frac{p}{(1-p)}\right)\left(\frac{-u''(w-q_1)u'(w-q_2) + u'(w-q_1)u''(w-q_2)\left(\left.\frac{dq_2}{dq_1}\right|_P\right)}{u'(w-q_2)^2}\right) < 0.$$

FIGURE 1



through the horizontal (q_1) axis at the point where $q_2 = 0$. The resulting value of q_1 is given by

$$I_2(q_1, 0) = [u(w) - u(w - q_1)] - [u(w - q_0) - u(w - q_0 - L)] = 0. \tag{5}$$

Begin by considering $q_0 = 0$, so that (5) reads $[u(w) - u(w - q_1)] = [u(w) - u(w - L)]$, from which it is clearly true that the relevant value of q_1 is precisely $q_1 = L$. But, applying the implicit function theorem to (5), it can be seen that

$$\left.\frac{dq_1}{dq_0}\right|_{I_2(q_1,0)=0} = \frac{u'(w - q_0 - L) - u'(w - q_0)}{u'(w - q_1)} > 0.$$

That is, the contour moves to the right as q_0 increases.

Now, condition (4) identifies all the points on or above the particular contour satisfying $I_2(q_1, q_2) = 0$. However, since the limit contour has positive slope, it holds that the entire set of points shown in Figure 1 must satisfy the constraint (4) for all $q_0 \geq 0$. Given this, the type-2 constraint can always be ignored.

Finally, consider the incentive compatibility constraint (3). Note directly that, since $u(w - q_0 + L) - u(w - q_0) > 0$, the incentive compatibility constraint can only be satisfied by a point $q_1 > q_2$. Consider the contours (level sets) of the function

$I_1(q_1, q_2)$. Directly from the implicit function theorem, the slope of any given contour for any given value of q_0 is simply

$$\left.\frac{dq_2}{dq_1}\right|_{I_1, q_0} = \frac{u'(w - q_1)}{u'(w - q_2)},$$

which is strictly greater than 1 for all feasible points ($q_1 > q_2$). Therefore, a contour that lies (initially) below the diagonal line $q_1 = q_2$ must approach the line but never reach it, i.e., the contour must lie everywhere below the 45° line. The result is that each particular such contour must be a positively sloped function that asymptotically approaches the line $q_1 = q_2$. It is important to note that the contours may not be concave functions (their concavity depends on the sign of the third derivative of the utility function).

One particular contour will be that corresponding to Equation (3) saturating. Hence, the incentive compatibility condition is simply the set of points on and below this particular contour. To see this, note that starting from a point on the contour, the value of $I_1(q_1, q_2)$ is increased by either an increase in q_1 or a decrease in q_2. This fact also implies that contours corresponding to higher values of q_0 are located below and to the left of contours corresponding to lower values. This is due to the fact that a marginal increase in q_0 will change $I_1(q_1, q_2)$ by $u'(w - q_0 + L) - u'(w - q_0)$ which is strictly negative since the utility function is strictly concave.

Finally, consider the point at which the contour $I_1(q_1, q_2)$ passes through the horizontal (q_1) axis.

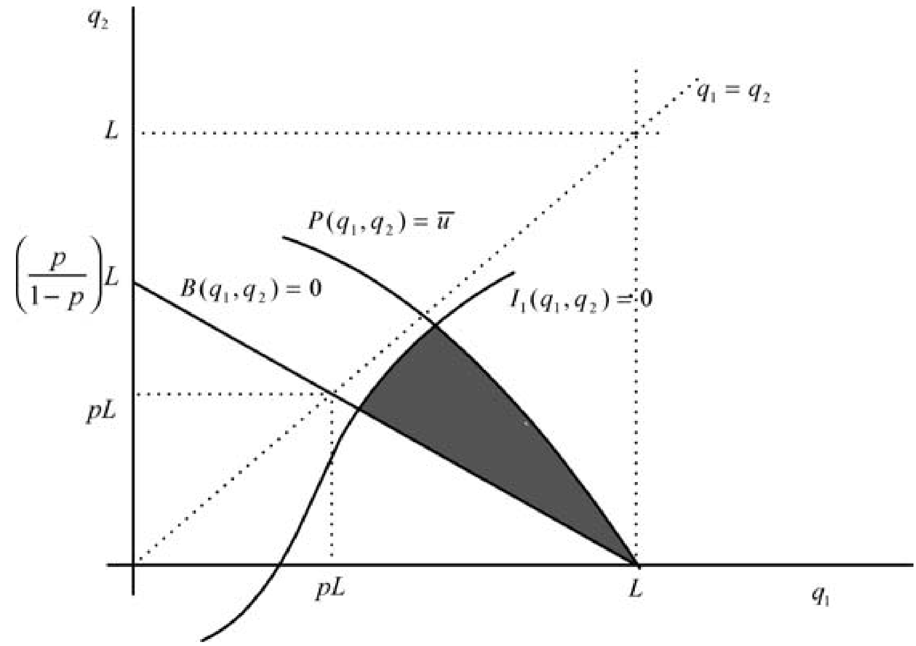
Lemma 1: *Given $0 \leq q_0 \leq L$, the incentive compatibility contour $I_1(q_1, q_2) = 0$ passes through the horizontal axis at a point $\tilde{q}_1$ that satisfies $q_0 \leq \tilde{q}_1 \leq L$.*

Given all of the above analysis, we can now draw the feasible set for the problem. This is done for a value of q_0 that is strictly positive but strictly less than L in Figure 2. Since $I_1(q_1, q_2) = 0$ is not necessarily a strictly concave function, the upper boundary of the feasible set may not be strictly concave, and so there exists the possibility that the set itself is not convex.

The Constrained Optimization Problems

We now go on to consider the two constrained maximization problems, corresponding to a perfectly competitive insurer on the one hand, and a monopolist on the other. We should firstly point out that the (possible) nonconvexity of the feasible set requires that, before attempting any constrained optimization, we firstly check the regularity conditions along the frontier of the set, which guarantees that when we maximize using the Kuhn–Tucker technique, all possible points that are candidate as optimum are considered. It turns out that all regularity conditions are indeed satisfied (see Moreno, 2003). In any case, the graphical solution that we present here clearly indicates both existence and unicity of solution for each of the two cases considered.

FIGURE 2



Consider firstly the case of a perfectly competitive insurer. The insurer's objective is to find the two premiums q_1 and q_2 that maximize the insured's expected utility in the follow-on period, subject to the constraints (1), (2), and (3). The solution is given by

Theorem 1: *For any given value of q_0, the contract $(q_1^c(q_0), q_2^c(q_0))$ that maximizes the expected utility of the insured (the perfect competition solution), while satisfying incentive compatibility, type-2 incentive compatibility, and nonnegative profits, is the (unique) point that saturates both (3) and (2). In this solution, the participation constraint of the insured is satisfied but does not saturate.*

The graphical intuition behind the perfect competition solution is simple. Note that the objective is to move the indifference curve of the client (contours of the function $P(q_1, q_2)$) as close as possible to the origin, without leaving the feasible set. Clearly, this will imply a point on the contour $B(q_1, q_2) = pL$, that is, the insurer's participation constraint will saturate. Recalling that, due to the fact that in the entire feasible set $q_1 > q_2$, all indifference curves will always be steeper than the contour $B(q_1, q_2) = pL$ at the point that the two intersect. Hence, the point that the client most prefers in the feasible set is the intersection of the contours $B(q_1, q_2) = pL$ and $I_1(q_1, q_2) = 0$, and so the simultaneous solution to these two equations, for any given value of q_0, gives us the solution to the perfect competition case.

Corollary 1: *For all values of q_0 such that $0 \leq q_0 \leq L$, the intersection of the two equations $I_1(q_1, q_2) = 0$ and $B(q_1, q_2) = pL$ occurs at coordinates that satisfy $pL > q_2 \geq 0$ and $L \geq q_1 > q_0$.*

Corollary 1 ensures that, independently of the value of the current period contract premium q_0, all renewals will have prices that are never negative nor greater than L, which is a natural aspect to expect from any bonus-malus contract. Furthermore, recall that since the solution occurs below the diagonal line, it always corresponds to a situation in which $q_1 > q_2$. However, it is curious that the exact relationship between q_0 and q_2 cannot be determined in general.

Now assume that the insurer is a monopolist. In this case, the problem that must be solved is to maximize the expected profits in the follow-on period, subject to the constraints (1), (2), and (3). The solution is given by

Theorem 2: *For any given value of q_0, the contract $(q_1^m(q_0), q_2^m(q_0))$ that maximizes the expected profit of the insurer (the monopoly solution), while satisfying incentive compatibility, type-2 incentive compatibility and participation by the insured, is the (unique) point that saturates both (3) and (1). In this solution, the expected profits of the insurer are strictly positive.*

Once again, the graphical intuition is obvious. Clearly, since the constraints are the same as for the perfect competition case, the feasible set is also the same. Indeed, the only change is that now the solution implies moving to higher and higher contours of the function $B(q_1, q_2)$ within the feasible set. Following a very similar reasoning to the perfect competition case, it is easy to see that the solution occurs at the intersection of the two contours $P(q_1, q_2) = 0$ and $I_1(q_1, q_2) = 0$.

Corollary 2: *For all values of q_0 such that $0 \le q_0 \le L$, the intersection of the two equations $I_1(q_1, q_2) = 0$ and $P(q_1, q_2) = 0$ occurs at coordinates that satisfy $q^* > q_2 \ge 0$ and $L \ge q_1 > q_0$.*

It is important to point out explicitly that in neither of the solutions above is the contract designed such that the client repays his claims over time via higher premiums. To see this, consider firstly the perfectly competitive solution. It lies on the line that holds the expected premium equal to the expected indemnity for the current period's contract. Thus, the expected premium is exactly equal to that would ensue in a perfectly competitive model with fully symmetric information on the loss, with the only difference being that it is at a point of positive variance. However, the point is that the expected premium only just covers the corresponding expected indemnity, and so nothing at all is being charged to the client to pay back past indemnities. The monopoly situation is very similar, although it does correspond to an expected premium that exceeds the expected indemnity, just as in a model with fully symmetric information on the loss. The difference here is simply that again the client gets an uncertain premium in the follow-on period (whereas he would get a certain premium in a symmetric information model), but indeed this expected premium is lower than what would be achieved in a model with symmetric information on the loss (where the contract would be at the point at which the autarky indifference curve of the client cuts the 45° line of the graph). Therefore, again, since nothing is charged in the symmetric information model for past claims, so is the case here. This serves to highlight the fact that in the proposed model, the bonus-malus scheme really does serve to eliminate fraud by eliminating the incentive to present fraudulent claims, rather than simply passing the costs of any fraud back onto the client by inflating the premium.

THE DYNAMICS OF THE SOLUTION

The sequence of contracts that is generated by the solution to either problem is interesting in its own right. Here we shall only mention the perfect competition case, although the monopoly case is, of course, very similar. For example, assume that the individual is fortunate and so does not suffer accidents for many periods in a row. In this case, his very first premium will be pL, and his follow-on period premium will be $q_2^c(pL)$, which from Corollary 1 is strictly less than pL. However, since the value of q_0 that will be used to calculate the premiums in the third period is lower than that which was used in the second period, the contour $I_1(q_1, q_2) = 0$ shifts to the left. Given this, we get an intersection between this contour and the line $B(q_1, q_2) = pL$ that is strictly above that of the previous period. Thus, the premiums that are offered for the follow-on period to the second have a higher value of q_2 and a lower value of q_1 than those values corresponding to the previous period. Curiously then, although the individual has (by assumption) not reported a claim, his premium is increased, although in exchange the punishment premium is reduced. If the individual continues the sequence without reporting a claim, his premium sequence is the sequence of q_2 values that the contract generates which oscillates but converges to a fixed value,[20] that we denote by $\bar{q}_2$. This fixed value satisfies $pL < \bar{q}_2 < q_2(pL)$. We shall denote the corresponding malus premium by $q_1^c(\bar{q}_2) = \bar{q}_1$.

On the other hand, an individual who is very unfortunate will find that his premium sequence (the sequence of q_1 values) monotonically increases and converges to the value L. In any case, given that the contract does eliminate all fraud, and given that the probability of an accident is certainly (in realistic cases) low, the probability of such an outcome is very low.

Perhaps the most realistic case is an individual who reports a claim very infrequently. For such individuals, the premium sequence will converge to $\bar{q}_2$ until a claim is reported. Then the premium will jump to the punishment premium $\bar{q}_1$, and will then (assuming again that no claims are reported) proceed to reconverge back to $\bar{q}_2$, although once again oscillating around this fixed value.

It is quite simple to arrive at an approximation for the fixed solution after a long series of no claim reports. Recall that the solution to the perfect competition problem for any given first period premium q_0 is given by the simultaneous solution to the two equations

$$pq_1 + (1 - p)q_2 = pL \quad \text{and} \tag{6}$$

$$u(w - q_0) + u(w - q_2) = u(w - q_2 + L) + u(w - q_1). \tag{7}$$

However, after a long sequence of periods without an accident the system converges to a situation in which $q_0 = \bar{q}_2$. The pair of premiums that will be offered in such a converged solution can be calculated from the two equations

[20] The proof of convergence is given in Moreno (2003).

$$p\bar{q}_1 + (1-p)\bar{q}_2 = pL \quad \text{and}$$
$$u(w-\bar{q}_2) + u(w-\bar{q}_2) = u(w-\bar{q}_2+L) + u(w-\bar{q}_1).$$

Simplifying, this last equation can be written as

$$u(w-\bar{q}_2) = \frac{1}{2}u(w-\bar{q}_2+L) + \frac{1}{2}u(w-\bar{q}_1).$$

Using the fact that $\bar{q}_1 = \frac{pL-(1-p)\bar{q}_2}{p} = L - (\frac{1-p}{p})\bar{q}_2$, we have

$$u(w-\bar{q}_2) = \frac{1}{2}u\left(w-\bar{q}_2+L\right) + \frac{1}{2}u\left(w-L+\left(\frac{1-p}{p}\right)\bar{q}_2\right). \tag{8}$$

Therefore, in the converged solution, the bonus premium must satisfy the requisite that $w-\bar{q}_2$ be the certainty equivalent wealth corresponding to a lottery that pays off $w-\bar{q}_2+L$ and $w-L+(\frac{1-p}{p})\bar{q}_2$ with equal probability. Because by definition, certainty equivalent wealth is equal to expected wealth less the risk premium, and because the expected wealth under this lottery is

$$\frac{1}{2}\left(w-\bar{q}_2+L\right) + \frac{1}{2}\left(w-L+\left(\frac{1-p}{p}\right)\bar{q}_2\right) = w + \frac{\bar{q}_2(1-2p)}{2p},$$

it turns out that the risk premium, π, is given by

$$w-\bar{q}_2 = w + \frac{\bar{q}_2(1-2p)}{2p} - \pi,$$

that is,

$$\pi = \frac{\bar{q}_2}{2p}.$$

However, because the variance of the lottery implied by Equation (8) is just

$$\sigma^2 = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(w-\bar{q}_2+L-\left(w-L+\left(\frac{1-p}{p}\right)\bar{q}_2\right)\right)^2 = \left(\frac{2pL-\bar{q}_2}{2p}\right)^2$$

the Arrow–Pratt approximation for the risk premium is

$$\pi \approx \frac{A\left(w+\dfrac{\bar{q}_2(1-2p)}{2p}\right)}{2}\left(\frac{2pL-\bar{q}_2}{2p}\right)^2,$$

where clearly $A(w) = -\frac{u''(w)}{u'(w)}$ is the Arrow–Pratt measure of absolute risk aversion.

Equating this approximation to the true value of the risk premium, after some simplification we have the result that the converged solution is (approximately) given by the following second-order equation

$$\bar{q}_2 \approx A\left(w + \frac{\bar{q}_2(1-2p)}{2p}\right)\frac{(2pL-\bar{q}_2)^2}{4p}. \tag{9}$$

COMPARATIVE STATICS OF AN INCREASE IN RISK AVERSION

In this short section, we consider how the converged solution to the antifraud bonus-malus system is affected by changes in risk aversion. To that end we have

Theorem 3: *Assuming that $A'(w) \leq 0$ and $p \leq \frac{1}{2}$, then an increase in absolute risk aversion will reduce the value of $\bar{q}_1 - \bar{q}_2$.*

The intuition for this is simple: the bonus-malus system described here eliminates insurance fraud by imposing premium uncertainty upon the insured. The more risk averse is this person, the less premium uncertainty is required to ensure that the relevant incentive compatibility constraint is saturated. The conditions on absolute risk aversion and probability are required since the implied change in the premium will, in general, alter risk aversion itself, and this effect must be controlled for as well.

Indeed, although the calculation is more complex, it can be shown that an increase in risk aversion will imply an upward shift of the curve that indicates the incentive compatibility constraint saturating for any initial premium q_0. Hence, at any particular period of the sequence, the gap between the bonus premium and the malus premium will be reduced (both for a perfectly competitive insurer and for a monopoly insurer).

WELFARE CONSIDERATIONS AND COMPARISON WITH THE PURE AUDIT MODEL

It is, of course, not clear exactly which of the two polar models is preferred by either party to the contract. To see this, note firstly that if one of the underlying characteristics of each polar model is that the insurer gets zero-expected profits (i.e., that the insurance market is perfectly competitive), then she will be indifferent between the two options. However, the position of the insured in this case is unclear. In the pure audit model (assuming that all fraud is eliminated), the premium for coverage must be set sufficiently high in order to recover the costs of maintaining the audit mechanism. On the other hand, the pure bonus-malus contract will offer premium uncertainty at a lower expected value (since there is no audit mechanism to pay for). On the other hand, if we are interested in modeling a monopoly insurer, the insured will always be indifferent between either system since the optimal monopoly strategy will involve saturating the insured's participation constraint, but in this case it is not clear which mechanism offers greater expected profits.

The self-selection solution to the *ex post* adverse selection problem works because the insured retains risk, in the form of premium uncertainty. When comparing with the pure audit model, we should consider the trade-off in welfare terms between premium uncertainty on the one hand, and premium certainty at a higher level on the other. In this section, we shall consider the possibility that the bonus-malus contract that we have outlined can Pareto dominate the pure audit model that is currently the status quo in the real world. Since this is meant as being illustrative only, we shall only present an argument pertaining to the perfect competition model.

If the pure audit model is used, we know that the premium that must be charged is strictly greater than pL, since the insurer suffers the costs relating to the maintenance of the audit mechanism, as well as the costs of any fraud that is not eliminated by the audit system. Let us indicate these costs by assuming that if a pure audit model is used, the total premium charged to the insured is kpL, where $k > 1$, which yields an expected utility of $u(w - kpL)$.

Now, we know that at any particular period, the bonus-malus contract (that eliminates all fraud) will have a certainty wealth level of w^* defined by

$$Eu(w - q) = u(w^*).$$

However, because the risk premium is $\pi = E(w - q) - w^* = w - Eq - w^*$, we have

$$Eu(w - q) = u(w - Eq - \pi)$$

and from the Arrow–Pratt approximation to the risk premium, we know that

$$\pi \approx \frac{\sigma_q^2 A(w - Eq)}{2},$$

where σ_q^2 is the variance of the bonus-malus premiums. Taking the approximation as being valid, and recalling that the measure of relative risk aversion is defined by $R(w) = wA(w)$, and that $Eq = pL$, we have

$$\pi = \frac{\sigma_q^2 R(w - pL)}{2(w - pL)}.$$

Hence, we can conclude that the insured prefers the bonus-malus contract sequence above the pure audit model if

$$kpL \geq pL + \frac{\sigma_q^2 R(w - pL)}{2(w - pL)}.$$

Now, since the bonus-malus contract returns a binomial random variable, we have $\sigma_q^2 = p(1 - p)(q_1 - q_2)^2$, and so the condition for the bonus-malus contract to Pareto dominate the pure audit model can be reformulated as

$$kpL \geq pL + \frac{p(1 - p)(q_1 - q_2)^2 R(w - pL)}{2(w - pL)},$$

that is,

$$k \geq 1 + \frac{(1-p)(q_1 - q_2)^2 R(w - pL)}{2L(w - pL)}. \tag{10}$$

This condition is clearly easier to satisfy, the greater is k. However, the effect of a change in R is far from obvious. On the one hand, the less risk averse is the insured, the lower will be the negative effect of the risk that the bonus-malus contract obliges him to retain. However, as was shown in the previous section, it is also true that the less risk averse is the insured, the greater will be the value of $q_1 - q_2$ that will saturate the incentive compatibility condition. Hence, it is unclear how an increase in the degree of risk aversion will affect the right-hand side of condition (10).

Before going on to an explicit consideration of the condition for the bonus-malus system to Pareto dominate the pure audit system, note that we can find some very simple (albeit poor) sufficient conditions that require far less information. Recall that, independently of the exact period (that is, whether the system has converged or not), it is always true that $q_1 - q_2 < L$, and so

$$\frac{(1-p)(q_1 - q_2)^2 R(w - pL)}{2L(w - pL)} \leq \frac{(1-p)LR(w - pL)}{2(w - pL)} = \left(\frac{L - pL}{w - pL}\right)\frac{R(w - pL)}{2}.$$

Hence, a sufficient condition for (10) to hold is the far simpler requirement

$$k \geq 1 + \left(\frac{L - pL}{w - pL}\right)\frac{R(w - pL)}{2},$$

that is,

$$2(k-1)\left(\frac{w - pL}{L - pL}\right) \geq R(w - pL), \tag{11}$$

where $k - 1$ is the proportional increase in the insurance premium over the expected value of the loss when a pure audit model is used. For example, if we had $p = \frac{1}{4}$, constant relative risk aversion equal to 1 (that is, the insured has logarithmic utility), and $k - 1 = 0.1$, then condition (11) requires that $w \geq 4L$.

Other sufficient conditions can also be found. Note that, since $pq_2 + (1 - p)q_1 = pL$, we have

$$q_1 = \left(\frac{p}{1-p}\right)(L - q_2) \quad \Longrightarrow \quad q_1 - q_2 = \left(\frac{p}{1-p}\right)(L - q_2) - q_2 = \frac{pL - q_2}{1-p}.$$

Therefore, we have

$$\frac{(1-p)(q_1 - q_2)^2 R(w - pL)}{2L(w - pL)} = \frac{(pL - q_2)^2 R(w - pL)}{2(1-p)L(w - pL)}.$$

And since at all points $q_2 \geq 0$,

$$\frac{(pL - q_2)^2 R(w - pL)}{2(1 - p)L(w - pL)} \leq \frac{(pL)^2 R(w - pL)}{2(1 - p)L(w - pL)} = \left(\frac{p^2 L}{(1 - p)(w - pL)}\right) \frac{R(w - pL)}{2}.$$

Hence, a second sufficient condition for (10) to hold is the requirement that

$$k \geq 1 + \left(\frac{p^2 L}{(1 - p)(w - pL)}\right) \frac{R(w - pL)}{2},$$

that is,

$$2(k - 1)\left(\frac{(1 - p)(w - pL)}{p^2 L}\right) \geq R(w - pL). \tag{12}$$

Once again, if we take the example of $p = \frac{1}{4}$, constant relative risk aversion equal to 1, and $k - 1 = 0.1$, then condition (12) requires that $w \geq 2L$. As can be seen, although this sufficient condition is somewhat more complex to calculate, it is also less restrictive in the sense that, given a level of L, it is satisfied by a larger set of wealth data.

Thirdly, independently of the value of p, it is always true that $p(1 - p) \leq 0.25$, and so

$$1 - p \leq \frac{1}{4p}.$$

Given this, we have

$$\frac{(1 - p)(q_1 - q_2)^2 R(w - pL)}{2L(w - pL)} \leq \frac{(q_1 - q_2)^2 R(w - pL)}{8pL(w - pL)}.$$

But since $q_1 - q_2 \leq L$, we have

$$\frac{(q_1 - q_2)^2 R(w - pL)}{8pL(w - pL)} \leq \frac{R(w - pL)L}{8p(w - pL)}.$$

Therefore, we can be sure that Equation (10) is satisfied if

$$k \geq 1 + \frac{R(w - pL)L}{8p(w - pL)},$$

which can be written as

$$\frac{(k - 1)8p(w - pL)}{L} \geq R(w - pL). \tag{13}$$

TABLE 1

	$R = 1$	$R = 2$	$R = 3$	$R = 4$
$p = 0.1$	1.1293	1.1449	1.1364	1.1211
$p = 0.2$	1.1143	1.1257	1.1156	1.1014
$p = 0.3$	1.0995	1.1068	1.0964	1.0833
$p = 0.4$	1.0848	1.0888	1.0779	1.0670
$p = 0.5$	1.0704	1.0718	1.0624	1.0523

With $p = \frac{1}{4}$, constant relative risk aversion equal to 1, and $k - 1 = 0.1$, then condition (13) requires that $w \geq 5.25L$. While this sufficient condition is more restrictive than the other two for the numerical example given, it does perform better than the second one (condition (12)) when $p > \frac{1}{2}$.

To summarize these sufficient conditions, we state

Theorem 4: *Assuming a perfectly competitive insurer, the bonus-malus antifraud mechanism will Pareto dominate the pure audit option if the relative risk aversion of the insured satisfies*

$$R(w - pL) \leq 2(k - 1)\left[\max\left\{\frac{w - pL}{L - pL}, \frac{(1 - p)(w - pL)}{p^2 L}, \frac{4p(w - pL)}{L}\right\}\right],$$

where $k - 1$ is the proportional increase in the insurance premium that is used to cover for fraud-related expenses in the pure audit model.

Finally, in Table 1 we present the limit values of k, as calculated from the true certainty equivalent in the defining equation $Eu(w - \tilde{q}) = u(w - kpL)$, for several different scenarios of relative risk aversion and accident probability, using a constant relative risk-averse utility function, and with parameter values of $w = 3$ and $L = 2$. The bonus-malus system is Pareto dominant over the pure audit system if k is at least as great as the number in the table.

According to Spanish insurance market statistics, between 40 and 50 percent of all clients present at least one claim during each insured period, and so it is most likely that in realistic situations the value of the probability can be set at around 0.4 or even 0.5. Hence, as can be seen in the table, in the majority of such situations we can affirm that the bonus-malus antifraud system will Pareto dominate the alternative pure audit model if, in the pure audit model, the existence of undetected fraud plus the costs involved in maintaining the audit system imply an increase in premiums in the order of between 6 and 8 percent.

CONCLUSIONS

In this article, we have considered the possibility that an insurer may eliminate all fraud (dishonest claim reporting) using only the incentives that can be provided by bonus-malus premiums. We have noted that indeed this appears to be feasible, at least for simplified settings. In particular, for the cases of both a perfectly competitive insurer and a monopolist, the sequence of premiums that is required to eliminate fraud never implies a negative premium nor a premium that exceeds the value of the insured loss.

Clearly, our model provides an opposite pole to the more traditional case in which fraud is controlled by audits, under which dishonest reporting is punished financially. However, we consider that our model is much closer in spirit to traditional adverse selection models in which the principal does not dedicate resources to evaluating agent type, but rather provides an incentive mechanism under which agents self-select. This is exactly what happens in our model, where agent type is defined by having had an accident or not. Furthermore, our model is also similar to traditional adverse selection models in that the cost of asymmetric information is borne by the "good" agent types (in our model, these are the agents who cannot commit fraud since they have legitimately suffered an accident). Concretely, in the traditional adverse selection equilibrium the incentive for bad types to masquerade as good types is eliminated by treating legitimate good types and dishonest bad types in exactly the same way (i.e., with exactly the same contract). This is exactly what the model in this article does; the incentive to commit fraud is eliminated by treating legitimate claimants and fraudulent claimants in exactly the same way, albeit with a treatment that still favors the legitimate claimant over an above the autarky option.[21] In any case, our only objective was to show that all fraud can be eliminated by bonus-malus premiums, with no audit mechanism at all for all clients that renew their contract, and not to advance the conjecture that the model of this article is in any way more optimal or efficient than a pure audit model. Certainly the optimal antifraud strategy would imply some sort of combination of the two extreme options.[22]

We have also discussed briefly the dynamics of the solution, and we have provided a simple way in which the point to which the system converges (under a perfectly competitive insurer) can be calculated (approximately). We have shown that the more risk averse is the insured, the less premium uncertainty is required to eliminate fraud in the solution to the model.

Finally, we have considered the welfare implications of the model, as it compares to the alternative pure audit model. A sufficient condition was found for Pareto dominance of the bonus-malus model, and this condition was simplified to find three other (weaker) conditions that are relatively straightforward to calculate from the parameters of the model. In any case, as a general rule of thumb, assuming that the probability of suffering a loss is in the order of 0.4 as real-life insurance statistics suggest, and assuming a realistic level of relative risk aversion (say below 4), the bonus-malus system will typically Pareto dominate the pure audit model when the existence of undetected fraud plus the costs of the audit system imply an increase in premiums between 6 to 8 percent.

[21] In the pure audit model, the externality costs implied by the informational asymmetry that allows fraud to take place is the sum of fraud that escapes plus the cost of the audit mechanism. These costs are either borne by the insurer, or more likely in real world situations, passed on to the entire collective of insurance clients in the form of inflated premiums.

[22] When both audits and bonus-malus are used together, each will temper the severity required of the other. That is, the existence of the bonus-malus strategy will reduce the probability and/or the fine required of the audit strategy, and the existence of the audit strategy will reduce the variance required of the renewal premiums in the bonus-malus model.

Our analysis may shed some light on why the business of certain types of insurance in some countries (for example, Spain) has traditionally had a lower dependence on audits than in other countries (for example, USA). The basic idea in the paper seems to rationalize why Spanish insurers, who have a traditionally heavy dependence on bonus-malus contracts for automobile insurance, have been able to provide this type of insurance without such a high committal to auditing as in the USA. Perhaps the bonus-malus scheme has inoculated them to a certain extent from the existence of fraud. It would be most interesting to carry out a full empirical consideration of the hypothesis that markets in which bonus-malus contracts are used have a lower dependency on adits than other markets.

The model in this paper is clearly more relevant to certain types of insurance markets (property insurance rather than life or personal injury insurance). The types of fraud claims that occur in a country such as Spain, where personal injury is covered by social health insurance, may be markedly different from the types of fraud that may be more prevalent in other countries (for example the USA, where bodily injury is privately insured). Property damage may be more tangibly documented, making a move up in bonus-malus class easier to justify, easier to price and easier to find the right settlement offer. Similarly, the use of litigation by both the insurer and the insured is much less in Spain than in USA, so that negotiation (and confrontation with suspected fraud in a non-legal arena) becomes easier and more likely, thereby allowing bonus-malus type systems to act more smoothly. However, we feel that the basic message of the model—that some degree of dependence on bonus-malus may allow a lower dependence on audits to control for fraud—can be imported into other types of insurance, and other insurance markets.

Our model can certainly be improved upon in many dimensions. The most obvious would be to reconsider the assumption concerning intertemporal discount factors. Among other details, the assumption made in this article has also required us to assume one period commitment for the insured, that is, that in order to break a contract sequence, the insured must give notice one period in advance. However, if the insured is given a greater time horizon, calculation costs are increased markedly, and it is doubtful that the marginal benefit in terms of a more consistent model would be greater than the marginal costs in terms of mathematical complications.

Other extensions to the model include (but are not restricted to) extending the dimension of the accident set beyond the two accident states that we have assumed, introducing other dimensions (other than loss outcomes) along which insureds may differ (for example, the probability of loss, or perhaps utility type) thereby allowing *ex post* adverse selection to exist along with *ex ante* adverse selection, and studying the optimal antifraud strategy as a combination of the two extremes.

APPENDIX (PROOFS)

Proof of Lemma 1: Since the required value of q_1 occurs where $q_2 = 0$, it must satisfy

$$[u(w) - u(w - \tilde{q}_1)] - [u(w - q_0 + L) - u(w - q_0)] = 0,$$

that is,

$$u(w - \tilde{q}_1) = u(w) - u(w - q_0 + L) + u(w - q_0). \tag{A1}$$

Clearly, so long as $q_0 < L$ we get $u(w) - u(w - q_0 + L) < 0$, so that Equation (A1) requires $u(w - \tilde{q}_1) < u(w - q_0)$, that is, $\tilde{q}_1 > q_0$. On the other hand, if $q_0 = L$, we have $\tilde{q}_1 = q_0$. But since the right-hand side of (A1) is strictly decreasing in q_0, it must also be true that $\tilde{q}_1 < L$ for any $q_0 < L$.

Proof of Theorem 1: The problem can be stated as

$$\max_{q_1, q_2} Eu(q_1, q_2) = pu(w - q_1) + (1 - p)u(w - q_2)$$

subject to

$$\begin{aligned} pu(w - q_1) + (1 - p)u(w - q_2) &\geq \bar{u} \\ pq_1 + (1 - p)q_2 &\geq pL \\ [u(w - q_2) - u(w - q_1)] - [u(w - q_0 + L) - u(w - q_0)] &\geq 0. \end{aligned}$$

Clearly, given that the problem is parameterized by q_0 (as well as L, w, p, and $\bar{u}$), the solution can be expressed as $q^* = (q_1^*(q_0), q_2^*(q_0))$.

We firstly construct the Lagrangian for the problem

$$\begin{aligned} L(q_1, q_2, \lambda_1, \lambda_2, \lambda_3) = {} & Eu(q_1, q_2) + \lambda_1 I_1(q_1, q_2) + \lambda_2[P(q_1, q_2) - \bar{u}] \\ & + \lambda_3[B(q_1, q_2) - pL]. \end{aligned}$$

The two first-order conditions are

$$\begin{aligned} \frac{\partial L}{\partial q_1} &= \frac{\partial Eu}{\partial q_1} + \lambda_1 \frac{\partial I_1}{\partial q_1} + \lambda_2 \frac{\partial P}{\partial q_1} + \lambda_3 \frac{\partial B}{\partial q_1} = 0 \quad \text{and} \\ \frac{\partial L}{\partial q_2} &= \frac{\partial Eu}{\partial q_2} + \lambda_1 \frac{\partial I_1}{\partial q_2} + \lambda_2 \frac{\partial P}{\partial q_2} + \lambda_3 \frac{\partial B}{\partial q_2} = 0. \end{aligned}$$

But since $\frac{\partial P}{\partial q_1} = \frac{\partial Eu}{\partial q_1}$ and $\frac{\partial P}{\partial q_2} = \frac{\partial Eu}{\partial q_2}$, these can be written as

$$\begin{aligned} \frac{\partial L}{\partial q_1} &= (1 + \lambda_2)\tfrac{\partial Eu}{\partial q_1} + \lambda_1 \tfrac{\partial I_1}{\partial q_1} + \lambda_3 \tfrac{\partial B}{\partial q_1} = 0 \quad \text{and} \\ \frac{\partial L}{\partial q_2} &= (1 + \lambda_2)\frac{\partial Eu}{\partial q_2} + \lambda_1 \frac{\partial I_1}{\partial q_2} + \lambda_3 \frac{\partial B}{\partial q_2} = 0. \end{aligned}$$

Recall that we must also fulfill the restrictions that

$$\lambda_i \geq 0 \quad i = 1, 2, 3.$$

Now, since the expected utility function Eu is continuous and the feasible set is nonempty, by the Weierstrass Theorem there must exist a global maximum as the solution to the problem. Furthermore, since the regularity conditions are satisfied, we know that the global maximum must satisfy the Kuhn–Tucker conditions.

We shall locate the solution in a few simple steps.

1. The insurer's participation constraint must saturate, that is $\lambda_3 \neq 0$. To see this, note that if $\lambda_3 = 0$ then the second first-order condition would read

$$\frac{\partial L}{\partial q_2} = (1+\lambda_2)\frac{\partial Eu}{\partial q_2} + \lambda_1 \frac{\partial I_1}{\partial q_2} = 0.$$

However, since

$$\frac{\partial Eu}{\partial q_2} = -(1-p)u'(w-q_2) < 0$$

and

$$\frac{\partial I_1}{\partial q_2} = -u'(w-q_2) < 0,$$

we have $\frac{\partial L}{\partial q_2} < 0$, that is the condition does not hold.

2. The incentive compatibility condition saturates, that is, $\lambda_1 \neq 0$.
To see this, note that if $\lambda_1 = 0$ from the first first-order condition we have

$$\frac{\partial L}{\partial q_1} = (1+\lambda_2)\frac{\partial Eu}{\partial q_1} + \lambda_3 \frac{\partial B}{\partial q_1} = 0.$$

However, using the relevant substitutions, this requires that

$$\frac{\partial L}{\partial q_1} = -(1+\lambda_2)pu'(w-q_1) + \lambda_3 p = 0 \Rightarrow (1+\lambda_2)u'(w-q_1) = \lambda_3.$$

On the other hand, from the second first-order condition we have

$$\frac{\partial L}{\partial q_2} = (1+\lambda_2)\frac{\partial Eu}{\partial q_2} + \lambda_3 \frac{\partial B}{\partial q_2} = 0.$$

Once again, making the relevant substitutions this reads

$$\begin{aligned}\frac{\partial L}{\partial q_1} &= -(1+\lambda_2)(1-p)u'(w-q_2) + \lambda_3(1-p) = 0\\ &\Rightarrow (1+\lambda_2)u'(w-q_2) = \lambda_3.\end{aligned}$$

Therefore, we require that

$$(1+\lambda_2)u'(w-q_1) = \lambda_3 = (1+\lambda_2)u'(w-q_2).$$

However, this implies directly that

$$q_1 = q_2$$

which, as was noted above, cannot ever satisfy the incentive compatibility condition (a necessary condition for the incentive compatibility condition to be satisfied is $q_1 > q_2$).

3. If the incentive compatibility condition saturates, $\lambda_1 \neq 0$, and if the insurer's participation constraint saturates, $\lambda_3 \neq 0$, then the insured's participation constraint must be slack, $\lambda_2 = 0$, since the point of intersection between $I_1(q_1, q_2) = 0$ and $B(q_1, q_2) = pL$ does not saturate $P(q_1, q_2) = \bar{u}$.

Therefore, we can deduce that there is a single point that satisfies the conditions of the solution; the point characterized by the simultaneous solution to:

$$I_1(q_1, q_2) = 0$$
$$B(q_1, q_2) = pL.$$

Let us call this point $q^c = (q_1^c, q_2^c)$. Finally, recall that in the (q_1, q_2) plane, the first of these equations is a strictly increasing function and the second is linear and strictly decreasing. Hence there always exists an intersection, and that intersection is necessarily unique. Furthermore, from Lemma 1, so long as $0 \leq q_0 \leq L$, the fact that $I_1(q_1, q_2) = 0$ is an increasing function implies that it must be true that $q_1^c \in (q_0, L)$, and since we know that $q_1^c > q_2^c$, and that $B(q_1^c, q_2^c) = pL$, it must also hold that $q_2^c \in (0, pL)$.

Proof of Theorem 2: The problem can be stated as

$$\max_{q_1, q_2} \; pq_1 + (1-p)q_2 - pL$$

subject to

$$pu(w-q_1) + (1-p)u(w-q_2) \geq \bar{u}$$
$$pq_1 + (1-p)q_2 \geq pL$$
$$[u(w-q_2) - u(w-q_1)] - [u(w-q_0+L) - u(w-q_0)] \geq 0.$$

In the same way as previously, we begin by writing the corresponding Lagrange function:

$$\begin{aligned} L(q_1, q_2, \lambda_1, \lambda_2, \lambda_3) = \; & B(q_1, q_2) - pL + \lambda_1 I_1(q_1, q_2) + \lambda_2[P(q_1, q_2) - \bar{u}] \\ & + \lambda_3[B(q_1, q_2) - pL]. \end{aligned}$$

The two first-order conditions are

$$\frac{\partial L}{\partial q_1} = (1+\lambda_3)\frac{\partial E\pi}{\partial q_1} + \lambda_1\frac{\partial I_1}{\partial q_1} + \lambda_2\frac{\partial P}{\partial q_1} = 0$$
$$\frac{\partial L}{\partial q_2} = (1+\lambda_3)\frac{\partial E\pi}{\partial q_2} + \lambda_1\frac{\partial I_1}{\partial q_2} + \lambda_2\frac{\partial P}{\partial q_2} = 0$$

and we must also bear in mind the restrictions

$$\lambda_i \geq 0.$$

Since $B(q)$ is continuous, and since the feasible set is compact and nonempty, the Weierstrass theorem guarantees that the problem has a global maximum. On the other hand, since the regularity conditions are satisfied, the global maximum must satisfy the Kuhn–Tucker conditions. As previously, we proceed in simple steps.

1. The insured's participation constraint must saturate, $\lambda_2 \neq 0$. To see this, note that if $\lambda_2 = 0$ then from the first-first order condition we would have

$$\frac{\partial L}{\partial q_1} = (1+\lambda_3)\frac{\partial B}{\partial q_1} + \lambda_1\frac{\partial I_1}{\partial q_1} = (1+\lambda_3)p + \lambda_1 u'(w-q_1) > 0.$$

Hence, in this case this first-order condition is not satisfied.

2. The incentive compatibility condition must saturate, $\lambda_1 \neq 0$. To see this, note that if $\lambda_1 = 0$ from the first-first order condition we would have

$$\frac{\partial L}{\partial q_1} = (1+\lambda_3)\frac{\partial B}{\partial q_1} + \lambda_2\frac{\partial P}{\partial q_1} = (1+\lambda_3)p - \lambda_2 p u'(w-q_1) = 0.$$

However, from the second first-order condition we would have

$$\frac{\partial L}{\partial q_2} = (1+\lambda_3)\frac{\partial B}{\partial q_2} + \lambda_2\frac{\partial P}{\partial q_2} = (1+\lambda_3)(1-p) - \lambda_2(1-p)u'(w-q_2) = 0$$

and so it would have to be true that

$$u'(w-q_1) = \frac{1+\lambda_3}{\lambda_2} = u'(w-q_2),$$

that is, $q_1 = q_2$ which can never satisfy the incentive compatibility condition.

3. If the incentive compatibility condition is saturated, $\lambda_1 \neq 0$, and if the insured's participation constraint is satisfied, $\lambda_2 \neq 0$, then the insurer's participation constraint must be slack, $\lambda_3 = 0$, since the intersection point between $I_1(q_1, q_2) = 0$ and $P(q_1, q_2) = \bar{u}$ cannot saturate $B(q_1, q_2) = pL$.

Therefore, there any solution that satisfies the Kuhn–Tucker conditions is characterized by the simultaneous solution to the two equations

$$I_1(q_1, q_2) = 0$$
$$P(q_1, q_2) = \bar{u}.$$

Let us denote any such point by $q^m = (q_1^m, q_2^m)$. However, as we have already noted, the contour implied by the equation $I_1(q_1, q_2) = 0$ is strictly increasing, and the contour implied by the equation $P(q_1, q_2) = \bar{u}$ is strictly decreasing, and so they can only have a single intersection. Finally, if we define q^* to be the point at which $P(q_1, q_2) = \bar{u}$ intersects the diagonal (clearly, q^* is the greatest premium that the individual would pay for full insurance in any given single period contract), that is $q^* \leftarrow u(w - q^*) = \bar{u}$, then it must always hold that $q_0 < q_1^m \leq L$, and $0 \leq q_2^m < q^*$.

Proof of Theorem 3: Take the approximate solution Equation (9) to be exact, and define

$$C(q_2, A) \equiv q_2 - A(y(q_2))\,\frac{(2pL - q_2)^2}{4p},$$

where $y(q_2) \equiv w + \frac{q_2(1-2p)}{2p}$. In this way, $C(\bar{q}_2, A) = 0$. From the implicit function theorem, after some simplifying, we have

$$\frac{d\bar{q}_2}{dA} = -\frac{\left(\frac{\partial C(\bar{q}_2, A)}{\partial A}\right)}{\left(\frac{\partial C(\bar{q}_2, A)}{\partial q_2}\right)} = \frac{(2pL - \bar{q}_2)^2}{4p - A'(y(\bar{q}_2))y'(\bar{q}_2)(2pL - \bar{q}_2)^2 + (2pL - \bar{q}_2)A(y(\bar{q}_2))}.$$

But since $y'(q) = \frac{1-2p}{2p}$ is nonnegative if $p \leq \frac{1}{2}$, then the entire denominator will certainly be positive if absolute risk aversion is nonincreasing (recall that $\bar{q}_2 \leq pL < 2pL$, so the third term on the denominator is strictly positive).

REFERENCES

Eeckhoudt, L., and C. Gollier, 1992, *Risk; Evaluation, Management and Sharing* (New York: Harvester Wheatsheaf).

Picard, P., 2000, Economic Analysis of Insurance Fraud, in: Dionne, G., ed., *Handbook of the Economics of Insurance* (Boston, MA: Kluwer Academic Publishers).

Crocker, K., and J. Morgan, 1998, Is Honesty the Best Policy? Curtailing Insurance Fraud Through Optimal Incentive Contracts, *Journal of Political Economy*, 106: 355-375.

Lemaire, J., 1985, *Automobile Insurance: Actuarial Models* (Boston, MA: Kluwer Academic Publishers).

Lemaire, J., 1995, *Bonus-Malus Systems in Automobile Insurance* (Boston, MA: Kluwer Academic Publishers).

Moreno, I., F. Vázquez, and R. Watt, 2002, Experience Rating as a Mechanism Against Insurance Fraud, paper presented to the meeting of the European Group of Risk and Insurance Economists, University of Nottingham.

Moreno, I., 2003, Un Modelo Bonus-Malus Para Eliminar el Fraude en los Seguros, Ph.D. Thesis, Univesidad Autónoma de Madrid.