Chapter 4

Structural Estimation of a Medical Insurance Principal-Agent Model with Moral Hazard

4.1 Introduction

Contract theory has been extremely important in the development of modern economic theory during the last thirty years. However, the increasing sophistication of the theory has not gone hand in hand with empirical validation of the models as (Salanié 1997) points out. Chiappori and Salanié (2000) offer us an up to date perspective of the literature that has tried to link econometrics and contract theory. Most of the existing works have used a reduced form approach.[1] Obtaining policy recommendations from reduced form models is usually a difficult task. Firstly, estimates of the raw economic parameters (technology and preferences) are usually necessary to find the policy function. Secondly, the parameters of reduced form models are a function of the raw parameters and agents constraints, which would in general shift in case a policy change takes place (Lucas 1976).

The purpose of this paper is to estimate the raw parameters of a principal-agent model of medical insurance that accommodates moral hazard. This will allow us to solve for the policy function (e.g., the optimal copayment) in presence of moral hazard. The

[1] Some exceptions are Ferrall and Shearer (1999), Margiotta and Miller (2000), Paarsch and Shearer (2000) and Biais *et al.* (1999).

framework also allows us to investigate how much the optimal copayment is influenced by departures from the fair premium hypothesis.

Our preliminary results give an estimate of optimal copayment of 2.1%, which is much lower than previously estimated in the literature. This result is consistent with the observation in Meza (1983) for which previous approaches overestimated welfare losses due to moral hazard. Furthermore, we find that slight departures from the fair premium hypothesis yield optimal copayments of the size found in the real world. This result suggest that policy makers should target more insurance companies efficiency rather than increasing consumer cost sharing.

Moral hazard in the use of medical services has been one of the most recurrent issues in health economics. Early references about the topic are Arrow (1963), Pauly (1968) and Zeckhauser (1970). The more widespread view is that

"When moral hazard is present, insurance that reduces risk will also cause larger expected losses. In medical care, these losses represent the consumption of units of medical care whose value to the consumer is less than their cost, because the insurance coverage reduce the user price below cost"(Pauly 1986).

An extensive literature has estimated elasticities of utilization of health care services with respect to insurance generosity.[2] However, this literature often lacks precise policy recommendations about consumer cost sharing levels. A step further is given by Feldstein (1973), Feldman and Dowd (1991), Buchanan *et al.* (1991) and Manning and Marquis (1996) that analyze the optimality of different insurance plans. Buchanan *et al.* (1991) found that the optimal copayment is 100% if the consumer faced a cap on expenditures of $1000 and no initial deductible. Manning and Marquis (1996) estimated the optimal copayment to be 45% in an insurance plan without a cap on expenditures. The papers mentioned above have extended the methodology initiated by Feldstein (1973) that considered that the first best level of utilization was the one corresponding to the absence of insurance. This attributes all extra medical consumption in presence of insurance to moral hazard. However a part of this increase might be due to income effects.[3] As a consequence, we may overestimate welfare losses attributable to moral hazard (Meza 1983).

Our paper differs from previous approaches in the methodology used. In order

[2]See, *inter alia,* Cameron *et al.* 1988, Manning *et al.* 1987, Coulson *et al.* 1995, Chiappori *et al.* 1998, Holly *et al.* 1998, Street *et al.* 1999 and Vera-Hernández 1999.

[3]Manning and Marquis (1996) found a positive influence of income on health care consumption.

to obtain the optimal copayment, we rely on first principles arising from the solution to a principal agent model, rather than comparing welfare losses between utilization for different levels of copayments and an *ad-hoc* value assigned to the first level of utilization: the one that corresponds to absence of insurance. The methodology used has provided us with a much lower estimate of optimal copayment than obtained by the previous literature. Moreover, simulation results suggest that real world copayments are more likely to occur because of unfair insurance premiums rather than consumer moral hazard. This has strong implications for public policy, since it implies that policy makers should target more insurance companies efficiency rather than establishing high rates of consumer cost sharing.

Consumer cost sharing is an important policy issue. On July 1993 in France, the social security cost sharing for ambulatory care and pharmaceutical goods was reduced in a 5% (Chiappori *et al.* 1998). In an attempt to control rising public health care costs, the Belgian government has raised coinsurance rates several times over the period 1986-1995 with a sharp increase in 1994 (Van de Voorde *et al.* 2000). In Spain, doctor visits are free but consumers are charged 40% for outpatient drugs charges. Ellis and McGuire (1993) argue that demand side cost-sharing might be desirable even in Health Maintenance Organization insurance plans.

The paper is organized as follows. In Section 4.2 we describe how the individual chooses to have treatment or not when suffering an illness episode by comparing out of pocket costs of treatment and the health penalty of not obtaining treatment. This follows the tradition health economics research in which the consumer is responsible for the contact decision, but not for the amount of treatment. In Section 4.3 we describe the data that we use, that comes from the RAND Health Insurance Experiment, that randomly assigned individuals to insurance plans. Consequently, for the estimation, we do not need to assume that observed contracts are optimal and we can condition on the insurance contract without the need of dealing with endogeneity issues as in the two previous papers of the thesis. Section 4.4 describes the econometric strategy followed to estimate the model. Section 4.5 gives the results of a preliminary analysis and Section 4.6 discuss the estimates of the structural parameters, and evaluates the suitability of the model. Section 4.7 sets up the principal-agent problem and explains the numerical technique to solve for the optimal copayment. In this section, we discuss in depth how our approach differs from the previous literature. The following section gives the directions

for improvement and the last section concludes.

4.2 The demand model

4.2.1 Individual decision problem

This section is devoted to modeling the individual's decision concerning whether to be treated or not when suffering an illness spell. This model is the basis for the estimation of the parameters of the principal-agent model. In our set-up, the consumer faces a specific insurance contract that will influence her decision. Given that in our dataset, individuals were randomly assigned to insurance contracts, we do not describe how individuals chose among different insurance choices.

Our model is inspired in the baseline model of Ma and Riordan (1997). At the end of this section, we will make specific the difference between their and our model. We build the model based on the following hypotheses: First, the individual decides whether to be treated or not but the amount of treatment is left for the doctor. This seems realistic given the informational asymmetry between doctor and patient. We will also assume that the doctor chooses treatment costs independently of individual's insurance coverage and income. This corresponds to the situation where the medical guideline that the doctor follows does not take into account individual economic characteristics but gives the most cost-effective treatment. Keeler and Rolph (1988) argue that some physicians may believe it improper to offer cheaper and lower quality care to patients with larger copayments[4]. Consequently, we will assume that treatment costs come from a given technological relation.

Second, the individual is rational and compares benefits and costs when she decides. This might be a strong assumption when one is dealing with severe illnesses for which the individual lacks experience and can hardly value the benefits and costs of the treatment. Furthermore, the treatment decision in the case of very severe illnesses might depend on long term effects that we are not ready to model. In the empirical application we will restrict the type of illness spells that we study in order to make our rationality assumption likely to hold.

We assume the individual is endowed with a health capital stock $\overline{s}$, and an income level y. An individual will be ill when she receives a penalty health shock of magnitude

[4]We will apply the model to relatively minor illnesses for which affordability will not be a problem.

$s > 0$. There will be a stochastic relation between s and $\overline{s}$, given by the conditional density function $f_{s|\overline{s}}$. If $s \leq 0$, the individual is not ill and therefore treatment is not demanded. We assume the support of s is $(-\infty, +\infty)$ .With this specification the illness process follows the same determinants as the health penalty process. We will justify this in terms of requirements for identification in the section devoted to the econometric specification.

The individual is under the coverage of an insurance contract that under premium p reimburses the cost of the treatment c, if the individual agrees to pay the quantity $D(c)$, where $D(.)$ is a non-negative function specified in the insurance contract. This function gives the cost sharing agreement between insurer and insured. The health penalty shock and cost of treatment variables will follow a joint density function conditional on the health stock, $\overline{s}$, and cost shifters $\overline{c}$ given by

$$g_{s,c|\overline{s},\overline{c}} = \begin{cases} 0 & if\ s \leq 0\ and\ c > 0 \\ f_{s|\overline{s}} & if\ s \leq 0\ and\ c = 0 \\ f_{s,c|\overline{s},\overline{c}} & if\ s > 0\ and\ c > 0. \end{cases} \tag{4.1}$$

The first part of the density function implies that costs cannot be positive when the individual is not ill ($s \leq 0$), since there is no need for treatment.

The timing of the model is as follows. First the individual receives a random draw of (s, c) from the joint density $g_{s,c|\overline{s},\overline{c}}$. When the individual suffers an illness spell ($s > 0$), she will obtain different utilities depending on her decision regarding treatment. We will assume that the individual knows the health penalty s and the cost c when she decides. We reckon that this might be a strong assumption but it is difficult to figure out an estimable demand model where the consumer knows neither the price nor the good she is buying. In the empirical application, we restrict the type of illness spells that we consider in order to make this assumption more plausible. We also assume that in case the treatment is obtained, there is perfect healing and the initial level of health is recuperated. As before, we think that restricting the types of illness spell that we consider might help to make this assumption more plausible. If treatment at cost c is obtained, the individual will have to pay the quantity $D(c)$. Specifically, the consumer's ex-post utility will be

$$\begin{array}{ll} U(y-p,s), & \text{if ill with health penalty } s \text{ but treatment is not obtained} \\ U(y-p-D(c),0), & \text{if ill and treatment is obtained with } D(c) \text{ as out of pocket payment} \\ U(y-p,0), & \text{if consumer is not ill.} \end{array} \tag{4.2}$$

We assume that $U(.,.)$ is increasing and concave in the first argument, while decreasing and convex in the second.

In what follows, we will describe the individual decision problem. Given that the set of actions is discrete (to have or not to have treatment), it is of interest to look for the health penalty threshold, that given a cost, leaves the individual indifferent between having treatment or not. For each cost c there is a unique value of $s = \widetilde{s}(c) \geq 0$ such that

$$U(y-p,\widetilde{s}(c)) = U(y-p-D(c),0).$$

Given c, the uniqueness comes from $U(y-p,0) > U(y-p-D(c),0)$ and the fact that the utility function is strictly decreasing in the second argument. Consequently we make the following definition.

Definition 1 *The health penalty threshold is the function* $\widetilde{s}(c)$ *such that* $U(y-p,\widetilde{s}(c)) = U(y-p-D(c),0)$.

Given a draw of (s,c) from the joint distribution above, the individual will decide to have treatment $(T=1)$ or not $(T=0)$, according to the following rule:

$$T = \begin{cases} 1 & if\ s > \widetilde{s}(c) \\ 0 & if\ s \leq 0\ or\ 0 < s < \widetilde{s}(c). \end{cases} \tag{4.3}$$

The intuition is very simple. The individual will decide not to have treatment $(T=0)$ either when she is not ill $(s<0)$ or when the health penalty shock does not offset the out of pocket monetary cost of the treatment $(0 < s < \widetilde{s}(c))$. Notice that $\widetilde{s}(c)$ depends on the cost sharing function $D(c)$. In particular, if the contract does not specify any out of pocket payment, $D(c) = 0$ for all c, then the individual will decide to have treatment when she is ill independently of the costs of the treatment.

As we said at the beginning of the section, the model is inspired in the baseline model by Ma and Riordan (1997). However, there are some differences. In our model, the illness probability and the health penalty shock are drawn from the same distribution, while in their model, both processes are modelled independently. We need this assumption in order to be able to identify the parameters in estimation. However, in our model, costs are random Ma and Riordan (1997) insurance model is for a specific illness and they assume that costs are fixed and do not vary with the health penalty shock.

4.2.2 Expected results on observed costs per episode

The structural model above allow us to analyze how the observed costs per episode vary with the copayment rates. To our knowledge, the literature has not discussed this issue using a theoretical demand model before. We will see that observed costs per episodes are decreasing in the generosity of insurance coverage, even if neither the patient nor the doctor decides to spend less in a treatment. To be specific, consider that $D(c) = kc$, where k is the copayment rate that is, the fraction of treatment costs that will have to be paid by the individual in case of having treatment. We give the following proposition.

Proposition 2 *For a given premium, p, and for every cost, c, given $D(c) = kc$, the health penalty threshold function $\widetilde{s}(c)$ is increasing in the copayment, k.*

This result directly comes from Definition 1. The greater is k, the smaller is the value of $U(y - p - kc, 0)$, and consequently, s must be greater for the equality in Definition 1 to hold. This means that the greater the copayment, the greater must be the health penalty in order to ask for treatment. Another way to look at this result is that given a health penalty, the greater the copayment the smaller must be the cost in order to seek treatment. That is, given a health penalty, those with a large copayment will ask for treatment only if the treatment is inexpensive enough. Consequently, observed costs are expected to be smaller for those individuals with greater copayments. Notice that this holds in a framework where nobody chooses treatment costs, but they are given by a technological relation. The key assumption is that individuals are able to anticipate treatment costs. We will see in Section 4.5 that the data verify this relation.

4.3 The Data

4.3.1 The experiment

The data that we use come from the RAND Health Insurance Experiment (HIE), a social experiment conducted between 1975 and 1982 in six different cities of USA. Families participating in the experiment were randomly assigned to one of fourteen different fee-for service health insurance plans. More information about the experiment can be obtained in Manning *et al.* (1987) and Newhouse *et al.* (1993). Keeler and Rolph (1988), Manning and Marquis (1996), Marquis and Holmer (1996), Deb and Trivedi (1999) and Guilleski and Mroz (2000) have previously used this data.

We would like to highlight two important characteristics of the dataset. First, insurance plans are exogenous to the individuals. Individuals did not choose the insurance plan they were enrolled, but they were randomly assigned to it. Participation incentives were paid to minimize the risk of attrition bias. Therefore the analyst does not need to face the endogeneity problem we have faced in the other two papers of the thesis, that is, the possibility that less healthy people, anticipating large medical expenditures, buy more generous insurance coverage.

Second, it gives information on illness episodes. Charges on claims from providers were grouped to create episodes of treatment. The grouping was based primarily on diagnosis, time since last charge for a related diagnosis, and information from the provider on treatment history. For each episode, the dataset contains information on total expenses, beginning and ending date, as well as the type of episode and the principal provider of the service. According to the type of the episode, they were classified as acute, chronic, chronic flare-up,[5] well-care and prenatal/maternity. The classification for providers include hospital inpatient, hospital outpatient, physician, dentist, pharmacy and nonpharmacy supplier. The classification of the provider is hierarchical, that is, outpatient services preceding or following a hospitalization were included in a hospital provider. Drugs and tests were part of the episode in which they were prescribed. This exhaustive information about episodes is important for us, since we are using a behavioral model that assumes that patients know the cost and severity of the treatment. This will be useful when we restrict the type of illness spells that we will use. More information on the way episodes were constructed can be obtained from Keeler and Rolph (1988) and references thereinafter.

[5] Temporary problem in a usually controlled condition.

Using information on episodes has an important advantage: one can separately study decisions to start episodes of treatment and decisions on the amount of treatment (cost of treatment). The decision to start an episode of treatment is mainly done by the patient, while the doctor, or the doctor jointly with the patient, will decide on the amount of treatment (or cost). Therefore, it is desirable to be able to study these two decisions separately.

The fee-for-service plans of the experiment had different levels of cost sharing that varied over two dimensions: the copayment rate (the percentage of the cost of each insurance claim that the individual pay out of her pocket) and an upper limit on annual out of pocket expenditures called Maximum Dollar Expenditure (MDE). Consequently, the family only paid according to the copayment if the total out of pocket expenditures had not exceed the MDE. The copayment rates were 0, 25, 50 or 95 percent. Depending on the plan, the MDE was 5,10, or 15 percent of the previous year's income, with a maximum of $1000.[6] In section 4.3.3, we will comment on the consequences of this MDE.

4.3.2 Our sample

In this subsection we will make specific the sample we have used in our estimation. It is important to restrict the sample in order to make it compatible with our structural model. We will begin to comment on the types of episodes used. In our structural model, the individual might receive a health penalty shock. This is not compatible with well care episodes nor with prenatal/maternity episodes. Nor are chronic treatments compatible with our model, since they are routine treatments. Therefore, the individual will predict them at the beginning of the contract year. We will not consider inpatient episodes since they are probably too severe and the individual could hardly decide on the basis of economic factors. We do not consider either dental episodes, since they were not usually covered by insurance contracts in those years and consumers showed an opportunistic behavior during the experiment. We do not analyze mental care episodes since they have different determinants and, even with generous insurance coverage, relatively little was spent on outpatient mental health care at the time the experiment was conducted.

Consequently, we have only considered acute and chronic-flare-up outpatient

[6] Apart from the ones mentioned above, there was a HMO plan, as well as an individual deductible plan that limited annual out-of-pocket outpatient expenditures to $150 per person, with a 95% of copayment rate for outpatient expenditures. These ones will not be used in this study.

episodes. Acute episodes (Newhouse *et al.* 1993) are defined by unforeseen and undeferrable treatment opportunities. From an economic point of view, spending on these episodes will occur only when the patient is temporarily sick. Consequently our model might be appropriate to deal with acute episodes. Chronic flare-up episodes are much as acute episodes, but they are caused by a chronic condition. Since we have not considered inpatient episodes, most of the episodes will be because of relatively minor conditions.

A subsample of individuals were enrolled in the experiment during five contract years (contracts had a duration of 365 days but not natural years, so they were called contract years), but most of them were enrolled only three contract years. We will use observations on the second contract year, so we are sure that individuals were familiar with the conditions of the insurance contract offered by the experiment. The third contract year will be used to evaluate the out-of-sample forecast capability of the model.

Our model does not allow for multiple spells. As we will see, the econometric implementation needs to account for sample selection in costs since we do not observe the treatment costs of those that were ill but did not seek treatment. Consequently, the model is already quite complicated in order to account for multiple spells. Therefore, our dependent variable will be a dichothomous variable that takes value 1 if the individual started to be treated by a new episode during the first month of the contract year.[7] As Table 4.1 indicates few people started more than one episode in the first month. As Guilleski (1998), in case of starting more than one episode in the month, we take the first one. Therefore, we only have one observation per person.

Table 4.2 indicates the duration of the episodes.[8] Most of them take as much one month. As Guilleski (1998) we decided to finally restrict ourselves to episodes that took at most one month. They are the most common ones and we think that constraining ourselves to not very severe episodes is important given that our model assumes that the individual fully takes into account economic factors when making her decision. Furthermore we do non trivial assumptions about individual's information about her health penalty and treatment costs, consequently we would like to the individual is as familiar as possible with the episodes.

[7]More specifically, during the second day of the contract-year and the following 30 days. Medical expenses in the new contract year that were generated by illness episodes that started in the preceeding contract year were recorded as starting the first day of the new contract year. We want to exclude them, since they are not decisions taken in the current contract year.

[8]After aplying the rest of filters that we used. These are described below.

We will only consider individuals that were in the experiment at any time of the first and second year.[9] Moreover, we only consider people older than 17, in order to consider episodes which are decided by the individual and not by their parents.[10] We do not consider either people that were retired or laid-off as well as self-employed people. They were very few in the sample and might differ much in their opportunity costs of having a treatment. Since we do not have information about this opportunity cost, we prefer not to include them in the analysis. We deleted 36 people with null income and 72 episodes that were recorded with a negative duration (recorded ending data was before recorded beginning date). Finally we deleted observations with missing values in relevant variables. Table 4.3 gives the description of the variables used, as well as the descriptive statistics of the observations used in the estimation. All the monetary variables are in 1973 dollars. We used monthly medical consumer price index to deflate treatment costs. Individuals participating in the experiment did not pay any insurance premium.

4.3.3 The maximum dollar expenditure and the copayment rate

In our estimation we will condition on the individual's copayment rate. Therefore, it is important that the copayment rate we condition on is the same one that the individual uses to decide on demand for treatment. As Keeler *et al.* (1977), Ellis (1986), Keeler and Rolph (1988) and Newhouse *et al.* (1993) have noted, the existence of a cap on out of pocket payments may make the effective marginal price and the nominal price differ (the copayment rate we condition on). If individuals were able to anticipate exceeding the cap with certain probability, then the effective marginal price will be smaller than the nominal one.

We have reasons to think that in our case this is a minor problem. First, according to Newhouse *et al.* (1993) "Few participants [in the HIE experiment] proved able to anticipate exceeding the MDE, which allowed us to ignore this factor and to obtain a much more tractable estimation problem". In addition, they show that the size of the remaining MDE was important in the decision to initiate hospital episodes, but not to initiate other episodes type. Consistent with the same idea, Keeler and Rolph (1988) showed that people participating in the experiment adopted a mixture of myopic and inflexible behavior. That is, if expenditures did not exceed the MDE, then they responded to current copayments.

[9] For the out-of-sample forecast, we take individuals that were in the experiment for the first three years

[10] Eligible individuals for the experiment were younger than 61.

Once out of pocket expenditures exceeded the MDE, people did not instantly adapt to the zero price of medical care, but they took some time to do so.

Our analysis is based on the first month and for episodes that took as much one month. Given the references above that apply to the same data that we use, families should be far from either exceeding the MDE or anticipating to exceed the MDE. Therefore, we feel confident on conditioning on the current copayment rate.

4.4 Econometric specification

4.4.1 Functional form assumptions

In this section, we will give specific functional forms to the demand model described above in order to carry out the estimation. We will start by the density function. The health penalty random variable, s, follows a normal distribution with mean $\overline{s} = x_s\beta_s$. That is,

$$s = x_s\beta_s + \varepsilon_s,$$

where x_s is a vector of covariates, β_s a conformable vector of parameters and ε_s is an unobservable component that follows a normal distribution with zero mean and variance σ_s^2. Consequently,

$$f_{s|\overline{s}} = \frac{1}{\sigma_s} * \phi(\frac{s - x_s\beta_s}{\sigma_s}),$$

where $\phi(.)$ is the standardized normal density function.

The cost per episode will be given by

$$\ln C = \alpha \ln(s+1) + x_c\beta_c + \varepsilon_c,$$

where x_c is a vector of covariates, β_c a conformable vector of parameters and ε_c is an unobservable component. Notice that the individual is ill only when $s > 0$, so the function above is well defined in the appropriate range. Inside the logarithm in the left hand side, we add one to the health penalty. This ensures that the cost increases with α for any $s > 0$. The functional form above is convenient since the predicted cost is always positive for any value of the parameters. All the previous studies cited above that estimated cost equations also used a logarithm transformation.

It will be assumed that $(\varepsilon_s, \varepsilon_c)$ follows a bivariate normal distribution with null means and variance covariance matrix given by:

$$\Sigma = \begin{bmatrix} \sigma_s^2 & \rho\sigma_s\sigma_c \\ \rho\sigma_s\sigma_c & \sigma_c^2 \end{bmatrix}.$$

By means of a change of variable we can obtain the joint density of (s, c), when both variables take positive values:

$$f_{s,c|x_s,x_c} = \frac{1}{c\sigma_s\sigma_c} * b(\frac{s - x_s\beta_s}{\sigma_s}, \frac{\ln c - \alpha \ln(s+1) - x_c\beta_c}{\sigma_c}; \rho), \tag{4.4}$$

where $b(.,.;\rho)$ is the standardized bivariate normal with correlation coefficient equal to ρ.

We will specify two types of utility functions, previously proposed by Ma and Riordan (1997). When the individual is ill with health penalty s, but does not have treatment, the utility function will be either,

$$\begin{aligned} U_I(y-p,s) &= U(y-p) - s, \text{ or} \\ U_{NI}(y-p,s) &= U(y-p-s). \end{aligned}$$

In case of having treatment the utilities will be

$$\begin{aligned} U_I(y-p-k*c,0) &= U(y-p-k*c), \text{ or} \\ U_{NI}(y-p-k*c,0) &= U(y-p-k*c) \text{ accordingly.} \end{aligned}$$

Using U_I, given a cost, the individual will have an episode of illness treated when the health penalty is larger than the health penalty threshold previously defined:

$$s > \widetilde{s}_I(c) = U(y-p) - U(y-p-k*c), \tag{4.5}$$

where k is the copayment rate. Notice that income, y directly influences $\widetilde{s_I}(c)$, therefore income will influence the decision to have treatment. In this case, we will say that there are income effects. Notice that in this case, income and health are not directly comparable, since the comparison is done by means of the marginal utility of income. Since the existence or not of income effects is a controversial issue in the literature, it is desirable to consider a specification where there are no income effects. The second utility function is useful for this purpose. For the second utility function the individual will have treatment when:

$$s > \widetilde{s}_{NI}(c) = k * c. \tag{4.6}$$

In this case, health and money are directly comparable and there are no income effects, that is, income does not influence individual demand for treatment.

We still must give a functional form for $U(.)$. The exponential utility function is very convenient for our purposes. Ferrall and Shearer (1999) comment that from a computational standpoint, the exponential utility is perhaps the only feasible functional form when solving the principal-agent problem[11]. Support to the exponential utility is given by Manning and Marquis (1996) and Marquis and Holmer (1996). Also using data from the RAND HIE experiment, they both argue in favor of the constant absolute risk aversion hypothesis. Therefore, based on both the complexities that arise using different specifications and the support found in favor of the constant absolute risk aversion hypothesis, we will use

$$U(z) = -\exp(-\theta z),$$

where θ stands for the constant absolute risk aversion coefficient. Marquis and Holmer (1996) found that θ did not vary with income or other demographics, so we will not parametrize it as a function of individual characteristics.

4.4.2 The likelihood function

Estimation will be done by maximum likelihood. Here we will give its formulation. Whether the individual had treatment or not gives us different information. For individuals with treatment, we can observe the cost of the treatment, however this is not the case for individuals without treatment. This is a sample selection issue that the model must accommodate.

The likelihood contribution for individual i if she had treatment ($T = 1$) at a cost $c > 0$, will be

$$L_{1,i} = \int_{\widetilde{s}_j(c_i)}^{+\infty} \frac{1}{c_i \sigma_s \sigma_c} * b(\frac{s - x_{si}\beta_s}{\sigma_s}, \frac{\ln c_i - \alpha \ln(s+1) - x_{ci}\beta_c}{\sigma_c}; \rho)\partial s, \qquad j = I, NI. \qquad (4.7)$$

The integral above is the probability of having treatment at cost c_i. This is the area of the density function (4.4) in the region at which it is optimal to have treatment. As given by (4.5) or (4.6), this occurs when the health penalty is large enough. We emphasize that, in general, the integration limit depends on the utility function used.

[11] Others authors that have used the exponential utility function are Towsend (1994), Mace (1991), Haubrich (1994) and Marigotta and Miller (2000).

The likelihood contribution for individual i if she did not have treatment ($T = 0$) will be

$$L_{0,i} = \Phi(-\frac{x_{si}\beta_s}{\sigma_s}) + \int_0^{+\infty} \int_0^{\widetilde{s}_j(c)} \frac{1}{c\sigma_s\sigma_c} * b(\frac{s - x_{si}\beta_s}{\sigma_s}, \frac{\ln c - \alpha \ln(s+1) - x_{ci}\beta_c}{\sigma_c}; \rho)\partial s \partial c, \qquad j = I, NI. \tag{4.8}$$

where $\Phi(.)$ denotes the cumulative distribution function of the standardized normal. This first term is the probability of not being ill, while the second is the probability of being ill but with a health penalty not large enough to offset the monetary out of pocket costs. The cost has to be integrated out, since it is not observed for those that did not have treatment.

Individuals with a copayment rate of zero, $(k = 0)$, do not pay anything for the treatment. In this case, $\widetilde{s}_j(c) = 0$ for all c. In this case, since the individual does not pay anything, the second term of (4.8) vanishes and the integral in (4.7) has a lower limit of zero. That is, the probability of having treatment is the probability of being ill. In this case, even if we use the utility function that exhibits income effects, income will not influence the demand decision. The model that does not exhibit income effects is robust to the choice of the functional form $U(.)$ since it does not enter in $\widetilde{s}_{NI}(c)$.

The log-likelihood function is given by

$$\ln L = \sum_{i=1}^{N} 1[T_i = 1] * \ln(L_{1,i}) + \sum_{i=1}^{N} 1[T_i = 0] * \ln(L_{0,i}).$$

Computation of the log-likelihood function requires numerical integration. The integration over costs in done using Hermite quadrature which is specially convenient to deal with integrands related to the normal density function (Judd 1998). To compute this integral, one needs to write the likelihood in terms of $\exp(-\xi^2)$ where ξ is the variable to integrate. We achieve this by doing the following change of variable:

$$\xi = \frac{\ln(c) - x_c\beta_c}{\sigma_c}.$$

As a result, we obtain integration limits that go from minus infinity to plus infinity, which is an additional requirement to apply Hermite quadrature. The integral over s is done by Legendre quadrature. This integration routine requires fixing the limits of integration to finite quantities. Therefore the plus infinity in (4.7) was replaced by the mean of s plus eight times its standard deviation. Since s is a normal random variable, this is

practically the same as integrating up to infinity. All the integration routines used are Gaussian quadrature techniques that outperform the alternative Newton-Cotes formula (Judd 1998). We used 7 points of quadrature in the Hermite integration and 40 in the Legendre[12].

4.4.3 Identification

First of all, we would like to comment on the restriction imposed in the theoretical model that both the illness process and the health penalty are generated by the same random variable s. In order to consider a more general formulation where illness and health penalty were generated by different processes we would need to observe a variable indicating that the individual was ill but decided not to have treatment. Since we only observe to have treatment or not, we have to constrain to a unique process that determines both illness and health penalty.

In order to consider more issues related to identification we will express the model in the following way:

$$\Pr(T=1|c) = \Pr(s > \widetilde{s}_j(c)), \quad j = I, NI \tag{4.9}$$

$$\ln c = \alpha \ln(s+1) + x_c\beta_c + \varepsilon_c, \ if \quad s > 0, \text{ where}$$

$$s = x_s\beta_s + \varepsilon_{s.}$$

$$U_j(y-p, \widetilde{s}_j(c)) = U_j(y-p-kc, 0). \tag{4.10}$$

Therefore it is a simultaneous equation model in which the dependent right hand side variables are not censored and therefore we do not require any restriction for coherency of the system. In order to obtain non-parametric identification, it is necessary to have at least one continuous restriction in both x_s, and x_c.

Another issue of concern is identification of the parameters of the discrete choice equation. For the model that does not exhibit income effects we have that equation (4.9) is

$$\Pr(T=1|c) = \Pr(\varepsilon_{s.} > \frac{kc - x_s\beta}{\sigma_s}),$$

where k is the copayment rate and therefore it is not a parameter to estimate, but a data we condition on. Consequently, if there are enough observations with $k > 0$ then, σ_s is

[12]The Hermite routine needs less points since it is designed to deal with normal related integrands as ours. The Legendre is a more general purpose routine so we have chosen more points for the Legendre than for the Hermite.

identified because the product kc changes among individuals. However, if there are not enough elements in the sample for which $k \neq 0$, this might result in poor identification. For the model that exhibits income effects, equation (4.9) is

$$\Pr(T = 1|c) = \Pr(\varepsilon_s > \frac{\exp(-\theta y) * (-1 + \exp(\theta kc)) - x_s\beta_s}{\sigma_s}). \tag{4.11}$$

In this case, the product kc is affected by the parameter θ and therefore the previous argument does not hold. Notice that since θ enters non linearly in the fraction, this might allow us to identify the parameters of the model. However, identification that relies on non linear restrictions can give poor results. This is the case in our model and we have decided to constrain one of the parameters. We use the value of θ estimated by Marquis and Holmer (1996) that used also data from the RAND HIE. However, we will not constrain θ directly, rather we will constrain for a value of σ_s that gives as a corresponding point estimate for θ equals to the estimated value by Marquis and Holmer (1996). With this strategy, all the observations in the sample will contribute to the identification of the parameters. If on the contrary, we constrained the value of θ directly, the observations with $k = 0$ will not contribute to the identification of the parameter because in that case the first term of the numerator of (4.11) vanishes.

4.5 Preliminary analysis

In this section we will analyze Tables 4.4 to 4.7 that show the results of a preliminary analysis of the data. Table 4.4 shows how frequency of episodes treated varies with copayment rates. Those that enjoy free care seek care more often than those that face cost-sharing contracts. Differences among copayment plans are not so clear. This is not strange since Newhouse *et al.* (1993) found that the largest decrease in use of outpatient services occurs between the free and the 0.25 copayment rate.[13] Table 4.5 shows the estimates of a standard probit model for TREAT as a dependent variable. The results on the copayment rates give us basically the same information that Table 4.4. The negative sign in the dummy variables for copayment 0.5 and 0.95 indicate a lower probability of having an episode treated but given the smaller sample size for these groups (see table 4.4), we do not obtain very precise estimates. Other significant variables are SOC and FEM. SOC measures the number of individual's social contacts. The magnitude of the

[13]Furthermore, in table 4.8 in Newhouse *et al.* (1993) the coefficients over the 0.5 and 0.95 copayment rates were very similar (-0.46 vs -0.49) in a Negative Binomial regression of number of acute episodes.

effect is surprising. We do not have a clear explanation for it but it seems interesting given that most studies lack information on this variable. DISEA is close to significance at the 10% level and it is highly correlated to GHEA. Their signs indicates that those with poor health have episodes treated with higher probability. AGE is not significant, what might be because we are considering only people older than 17 and younger than 62 and we have already conditioned in better health indicators than age.

Table 4.6 shows how the average of episode treatment cost for those treated varies with copayment rates. Table 4.7 shows a regression for $\ln(c)$ over covariates, including dummies for cost-sharing contracts. It seems clear that the higher the copayment rate, the lower the average observed cost. We anticipated this result in section (4.2.2) using our theoretical demand model. According to our model, the higher the copayment rate, the more inexpensive the treatment has to be in order to ask for treatment, given an illness episode of the same severity. Therefore, one expects to observe this decreasing pattern in costs per episode.

Using a different subsample, Keeler and Rolph (1988) also finds this negative pattern in cost per episodes but their coefficients are smaller (of a magnitude of -0.03 to -0.14). Their subsample consider all ages, and they pool together all the episodes of an individual in the second year. Therefore, they include all the episodes once MDE has been exceeded while we only consider the first episodes in the first month. Consequently, it is not strange that their coefficients are smaller since individuals with cost-sharing plans will pay nothing once the MDE has been exceeded.

Other significant variables in the regression are GHEA, DISEA and FEM. Ones with better health (GHEA variable) have smaller costs per episode. On the contrary, those with more diseases have smaller costs. We think that this is not implausible, since people with more diseases might have a longer relationship with their doctors and this might save diagnosis costs. Education years is close to be significant at the 5% level. Its negative coefficient might be indicating that those with more education might be more efficient producing health.

4.6 Structural estimation results

4.6.1 Model results

In this section we will comment on the results obtained in the estimation of the model developed in section (4.4). We will first comment on the model that exhibits income effects. Table 4.8 shows the parameter estimates and standard errors that have been computed using the sandwich form. We have excluded AGE from both equations since it did not turn out significant in any equation in the preliminary analysis shown above. EDUC is excluded from the health penalty equation since we do not think that education might influence health per se, and in addition it did not turn out significant in the previous simple probit analysis. The variable SOC is excluded from the second equation since it did not influence costs significantly in the preliminary analysis shown above and we do not have a priori reason to think that it might influence cost per episode.

The coefficients shown in the health penalty equation determines the mean of the s variable, that is, $\overline{s}$. Women and people with frequent social contacts have a bigger mean which gives a larger probability of illness and consequently, higher health penalties. The coefficients related to variables DISEA and GHEA give also expected signs, though since they are highly collinear, only DISEA is close to significance. The results reported are for a $\sigma_s = 2.75$ which gives a risk aversion coefficient of $\theta = 0.00306$ very close to the one estimated by Marquis and Holmer (1996) which is 0.00309 in 1973 dollars[14].

Regarding the coefficients of the cost equation, the ones related to covariates have the same sign that the OLS regression presented in the preliminary analysis. The point estimate of α which gives the relation between costs and health penalty have a positive sign and is smaller than one, which means that the relation between costs and health penalty is concave.[15] If we had a convex relation, cost would be very large when health penalty is large. As the rest of estimates in the cost equation, the estimates are not very precise. In fact, from a statistical point of view, we interpret this cost equation as a cost equation with additional unobserved heterogeneity. Given that the number of observations with episodes treated (and therefore positive costs) is small (201), we do not find strange that most of the variables do not turn out to be significant when additional unobserved heterogeneity is included. We think that the reduced sample size is also responsible for the large standard

[14] As we mentioned in the section devoted to identification, we have chosen σ_s in order to produce this estimate of risk aversion.

[15] The parameter α was estimated unrestricted.

error of the logarithm of risk aversion $\ln(\theta)$.[16] In fact, as we explained in section (4.4.3), only observations with non-zero copayment are relevant for the risk aversion coefficient. As shown in Table 4.4 almost half of our sample have zero copayment. Consequently, it is not surprising to find that the estimate is imprecise.

As we anticipated in section (4.4.3), the identification of the model that did not exhibit income effects relied only on those observations with copayment different from zero. We estimated the model, but we did not find significant variables even in the health penalty equation. Furthermore, the α turned out to be negative although not significant. Therefore we do not report the results here. However, we do not take this as favoring the income effects model, but just that the reduced sample size have not allowed us to estimate the model without income effects.

4.6.2 Model evaluation

The purpose of this section is to analyze the suitability of the structural model estimated above. We have computed the value of Andrews' (1988) goodness of fit test. This is a conditional moment test for the difference between the average estimated probability of TREAT=1 and the frequency of ones in the sample. We obtained a P-value of 0.52 that does not reject the structural model at standard confidence values.

Table 4.9 and 4.10 analyze the predictive capability of the model compared with the probit model both in sample and out of sample (we took third year observations). An observation was predicted as one when the predicted probability of TREAT=1 for this observation was greater than the frequency of TREAT=1 in the sample. The structural model does a slightly better job predicting TREAT=1 than the probit model, while the probit does it better predicting TREAT=0. The overall percentage both in sample and out of sample is slightly better for the probit model, which is not surprising since in the probit model the copayments were included as dummy variables, which gives a good deal of flexibility to the model. Overall we take the predictive capability of the structural model as reasonable, since the percentage of observations predicted out of sample is quite close to the one given by the probit and also close to the in-sample values.

In Table 4.11, we give the results of montecarlo simulations based on 50000 replications of the model using our sample. It is useful to compare the predicted frequencies

[16] In the estimation we used $\ln(\theta)$ as input in the likelihood function in order to preserve positiviness of θ. The standard error of θ is obtained using the delta method and it is shown below in the table.

of TREAT=1 with the real ones, as well as predicted costs by copayment with real ones. Regarding probabilities, the point estimates overestimate the frequency of TREAT=1 but the real values are within 1 standard deviation of the predictions. Moreover, the point estimates seem to reproduce the U-shape that appears in the data because the frequency of 1 for those with copayment 0.95 is higher than the one with copayment 0.5. Regarding costs, the point estimates of costs underpredict the real cost for those with zero copayment and overpredict the costs of those with 0.5 and 0.95 copayment rates. The point estimates seem to capture that the costs for those with 0.5 and 0.95 copayment rates are quite close, as well as the level of costs for those with 0.25 copayment rate, but do not capture the difference between the zero and 0.25 copayment rate, as well as 0.25 and 0.5. The real values are within 1.5 standard deviation of the predicted values. The standard deviation of the montecarlo series are large, because of the large standard deviations of the point estimates. We think that we need more precise estimates to give policy recommendations with confidence. The following sections are a hint of the power of the structural approach in giving policy recommendations.

4.7 Optimal contracts

4.7.1 Problems definition

Once one has estimated the parameters of the principal-agent problem, one can solve for the optimal contracts and therefore estimate them. This is possible thanks to the fact that we have estimated the parameters of the theoretical model in the previous section. In this section, we will define the optimal contracts: the first best, the second best and optimal copayment. We would like to highlight that, for the estimation, we did not need to assume that contracts were optimal. This was possible thanks to the experimental design of the data we used.

It is important to describe the timing of events as well as the contractibility assumptions. At $t = 1$, the contract is offered to the individual. The contract might specify the premium p, the cost sharing function $D(.)$, and, if possible, the region of the (s, c) space at which treatment will be covered by the insurance scheme. She accepts or rejects. If she accepts, at $t = 2$, she will obtain a draw of (s, c) from the joint distribution $g_{s,c|\overline{s},\overline{c}}$. If she is ill, $s > 0$, conditional on the contract, the individual will decide to have treatment or not. At this stage, the individual knows the realization of the two random

variables. In the First Best, which is useful as a benchmark case, the two random variables are contractible. However in the second best and the optimal copayment case, the only contractible variable is the cost, and the health penalty shock will be individual's private information. This informational asymmetry is the source of moral hazard[17].

In the first best problem, since both costs and health penalty shock are contractible, the insurer does not need to assume that the individual will behave optimally. That is, the contract offered to the individual will be a complete contingent plan on both s and c, that maximizes her expected utility conditional on the zero profit restriction for the insurance company, which comes from the assumption on competition.[18] The first best contract

$$c_{fb} = \{P, D(s,c), TC, NTC)\},$$

will determine the Premium p, the cost sharing function $D(s,c)$ and the sets TC and NTC. Since both variables are contractible, the cost sharing function might depend on both. The set TC will contain the pairs (s,c) for which treatment is covered by the insurance. NTC is the set of pairs (s,c), with $s > 0$, for which treatment is not covered, that is

$$NTC = \{(s,c) : \ s > 0, \ c > 0, \ (s,c) \notin TC\}.$$

We have not specified what means that the treatment is not covered by the insurance contract. If a realization of costs and health penalty shock (s,c) is not covered by the insurance contract, then it means that the individual will have to pay all the costs of the treatment, should he want a treatment. We anticipate the following idea: the insured will be better off not having treatment covered for illness spell for which the cost is high and the health penalty shock is negligible. Having treatment covered for such a spell will imply an important increase in the premium and a negligible benefit in health. The first

[17] We say this is a moral hazard problem since it is an informational asymmetry after the signature of the contract (Laffont-Tirole 1993, Macho and Pérez 1993). Since it is more an informational advantage than an action, other authors would call it hidden information.

[18] In fact, the problem could easily be re-specified to allow for a positive level of expected profits. By solving the problem for different level of profits one would obtain the utility-profit frontier.

best problem will be

$$\begin{aligned} \underset{\{P,D(C,s),TC,NTC)\}}{Max} \Pr(s \quad &< \quad 0|\overline{s}) * U(y-p,0) + \iint_{NTC} U(y-p,s) g_{s,c|\overline{s},\overline{c}}(s;c)\partial s \partial c \\ &+ \iint_{TC} U(y-p-D(s,c),0) g_{s,c|\overline{s},\overline{c}}(s;c)\partial s \partial c. \\ st \quad &: \quad P = \iint_{TC} c g_{s,c|\overline{s},\overline{c}}(s;c)\partial s \partial c. \end{aligned}$$

The objective function to maximize is individual's expected utility. Its first term relates to the utility of not being ill. The second relates to the utility of being ill but not having treatment and the third to the one corresponding to being ill and having treatment. In the second part, the individual will suffer the health penalty disutility while in the third she will have to pay $D(s,c)$ for having the treatment. The restriction of the problem is that the insurance contract is fair, in the sense that the premium will be equal to the expected costs. In this case, this is equivalent to the zero profit condition. We will make an assumption about the utility function:

Assumption. We assume that $\frac{\partial^2 U}{\partial y \partial s} = 0$, that is, the marginal utility of income does not change with the health penalty shock.

With this assumption, we can characterize the optimal solution for $D(s,c)$. It is given in the following proposition:

Proposition 3 *if* $\frac{\partial^2 U}{\partial y \partial s} = 0$, *then the optimal solution of* $D(s,c)$ *in the first best problem is* $D_{fb}(s,c) = 0$ *for all* c *and* s. *That is the optimal solution to the first best problem implies complete insurance.*

This comes from the standard result that if all the variables are contractible then optimal risk sharing applies. Since optimal risk sharing requires that marginal utility of income is the same across states, and marginal utility of income does not change with states then, the optimal solution is complete insurance.

Now we will simplify the problem. The optimal solution for TC, that is, TC_{fb} contains the pairs (s,c) such that it is efficient to give treatment. It is expected that for a given cost $c > 0$, if it is efficient to give treatment for $s_a > 0$, it will be efficient to give treatment for any other larger health penalty since the premium does not increase and utility increases. In the same way, if for a given cost $c > 0$, it is not efficient to give treatment for $s_b > 0$ then it will not be efficient either to give treatment for $s < s_b$, since the cost is given and the benefits of treatment are smaller. In order to simplify the

problem we define the function $\widehat{s}(c)$ as the minimum health penalty for which it is efficient to have treatment, given the cost c. More formally,

Definition 4 $\widehat{s}(c) = \min(\{s : (c,s) \in TC_{fb}\})$.

Using this last definition and the fact that $D(c,s) = 0$, we can simplify the first best problem as:

$$\begin{aligned} \underset{\{P,\widehat{s}(c)\}}{Max} \Pr(s \quad < \quad & 0|\overline{s}) * U(y-p,0) + \int_0^{+\infty}\int_0^{\widehat{s}(c)} U(y-p,s) g_{s,c|\overline{s},\overline{c}}(s;c)\partial s \partial c \\ & + \int_0^{+\infty}\int_{\widehat{s}(c)}^{+\infty} U(y-p,0) g_{s,c|\overline{s},\overline{c}}(s;c)\partial s \partial c. \\ st \quad : \quad & p = \int_0^{+\infty}\int_{\widehat{s}(c)}^{+\infty} c g_{s,c|\overline{s},\overline{c}}(s;c)\partial s \partial c. \end{aligned}$$

Solving the problem might be a difficult task. Solving the problem implies finding the function $\widehat{s}(c)$ that maximizes the objective function above. This is a difficult problem since the function appears in the integration limits. We have not found a standard mathematical analytical tool to deal with it. However, the problem itself gives us an intuition about the trade off faced by the agents. An increase in the availability of treatment will increase the expected utility because less health losses will be faced by the individual, but it will increase the premium and this will influence negatively the objective function. One possible way to deal with the problem above is to parametrize the function $\widehat{s}(c)$ using a enough flexible function and obtain the parameters of this parametrization that solve the problem above. This could give an approximation to the first best solution. We have tried with the exponential of a polynomial in $ln(c)$ but we have found difficulties in convergence. That is not surprising since the function is in the integration limits and we have used standard gradient methods to maximize the function. In the future, we would like to try with more robust maximization algorithms such as simulated annealing.

The second best problem assumes that the insurer can just contract on the costs but not on the health penalty shock. Moreover, since there is noise between these two random variables, the insurer cannot recuperate the health penalty from the cost. Consequently, in the second best problem no argument of the contract can depend on the health penalty shock. Therefore, the second best contract will be

$$c_{sb} = \{p, D(c)\},$$

where p is the premium and $D(c)$ the cost sharing function. At the second best, the insurer cannot design a region of (s,c) for which treatment is covered, since s is not contractible. On the contrary, the insurer will take into account that the insured will behave optimally according to his optimal decision rule (4.3). Therefore, if the treatment episode is c, the individual will demand treatment when $s > \widetilde{s}(c)$ where $\widetilde{s}(c)$ is implicitly defined by $U(y-p,\widetilde{s}(c)) = U(y-p-D(c),0)$. Notice that $\widetilde{s}(c)$ is different from $\widehat{s}(c)$. The first one corresponds to the ex-post optimal individual's response, while the second one correspond to the ex-ante optimal one, the first best.

The second best problem is defined by

$$\begin{aligned}
\underset{\{P,D(c)\}}{Max}\ \Pr(s < 0|\overline{s}) * U(y-p,0) + \int_0^{+\infty}\int_0^{\widetilde{s}(c)} U(y-p,s)g_{s,c|\overline{s},\overline{c}}(s,c)\partial s\partial c \quad (4.12)\\
+ \int_0^{+\infty}\int_{\widetilde{s}(c)}^{+\infty} U(y-p-D(c),0)g_{s,c|\overline{s}}(s,c)\partial s\partial c.\\
st \ : \ p = \int_0^{+\infty}\int_{\widehat{s}(c)}^{+\infty}(c-D(c))g_{s,c|\overline{s},\overline{c}}(s,c)\partial s\partial c. \quad (4.13)\\
U(y-P,\widetilde{s}(c)) = U(y-p-D(c),0). \quad (4.14)
\end{aligned}$$

Notice that the problem has two constraints. Now the function that appears on the integration limit is not the solution of the problem (as in the first best one), but is a constraint given by the optimal individual's behavior. The optimal copayment problem can be obtained using the second best problem by restricting to $D(c) = k * c$, where k is the copayment rate. In this case, we just need to solve for k and p.

4.7.2 Numerical solution of optimal copayment

One of the advantages of structural estimation is the possibility to solve for policy measures. This is possible because one recuperates the raw parameters of the theoretical model. In this case, we are interested in obtaining the optimal copayment.

Solving for the optimal copayment implies solving for p and k in (4.12) where $D(c) = k*c$. We have used the estimates previously obtained, the mean value of covariates and the model that includes income effects

$$U(y-p,s) = -\exp(-\theta * (y-p)) - s.$$

As a numerical technique, we have chosen a grid search on p and k. This is a robust algorithm which is likely to be successful. The grid for the premium goes from

zero to 4.37. This last value corresponds to zero copayment according to restriction (4.13) that is, the maximum possible fair premium. The grid for the copayment was from zero to one. For each premium level, we found within the grid for optimal copayment the most suitable one according to restriction (4.13).[19] The expected utility was computed using each premium of the grid and its associated copayment. The function $\widehat{s}(c)$ was obtained from restriction (4.14). We give the results for selected values of the grid in Table 4.12.

We found the optimal copayment to be 2.1% and a associated premium of 4.26. Therefore according to our results, the consumer should be responsible for only a very small part of treatment costs. This might be driven by the underprediction in our point estimate of cost for those with zero copayment. This makes cheaper than might be to have treatment and therefore the estimation of the optimal copayment is biased downwards. However, the following section might give new insights for our low optimal copayment.

4.7.3 Optimal copayment with unfair premium

Real life copayments are higher than the one we have estimated. However other factors different from moral hazard might be affecting real life copayments. Ex-ante asymmetry of information (adverse selection) might increase the copayment rate of some contracts in order to separate people with different ex-ante risk. Copayments might also be higher if insurance companies do not charge fair premiums (that is, if they charge premiums greater than the expected cost). Our model can incorporate this issue by substituting equation (4.13) in the optimal copayment problem (problem 4.12 with $D(c) = kc$) by

$$p = (1+\lambda)\int_0^{+\infty}\int_{\widehat{s}(c)}^{+\infty}(c-kc)g_{s,c|\overline{s},\overline{c}}(s,c)\partial s\partial c,$$

where λ is a loading factor that express by how much the premium exceeds the expected cost. In Table 4.13 we have tabulated the optimal copayments for $\lambda = 0, 0.05, 0.1, 0.15$. From the results it is clear that small loading factors can increase substantially the value of the optimal copayment. An optimal copayment value of 14.2% is obtained when the premium exceeds the expected cost in 5%. Notice also that the higher the loading factor, the smaller the expected utility of the optimal copayment, as one would expect. If these results were confirmed by more precise estimates, this would suggest that real life observed values of copayments are more a product of unfair premiums rather than inefficiencies arising from moral hazard. We think that this is a very interesting

[19] The increment of the grid for the premium was 0.01 and for the copayment 0.001.

result since it means that policy makers should target insurance companies rather than increasing consumer cost-sharing.

4.7.4 Differences with previous literature

We would like to comment on the difference between our approach to obtain the optimal copayment and the ones previously used in the literature. Previous papers (Feldman and Dowd, 1991, Buchanan *et al.* 1991, Manning and Marquis, 1996) assumed that the first level of utilization of health care services was the one corresponding to the absence of insurance. For each copayment smaller than one, they computed the welfare loss associated to a utilization larger than the assumed first best and the gains of facing less risk than in the case of no insurance. The difference between both of them constituted the net welfare gain associated to each copayment. The optimal copayment corresponded to the one with a higher welfare gain. So previous approaches did not set up a principal-agent model but obtained their conclusions from a demand function and a utility function to assess the gains from reducing risk. Consequently a key assumption of previous approaches is that the first level of utilization used in the analysis (the utilization level of no insurance) corresponded to the real first best level.[20]

However, this might not hold in presence of income effects. Ma and Riordan (1997) claim that in presence of income effects, the implementation of the first best level requires the consumer to be responsible for only a fraction of treatment costs because her marginal valuation of income rises once a deductible is paid out of her income. Therefore, the first level of utilization is greater than the one corresponding to the absence of insurance. Another way to look at this is that insurance transfers income from the healthy states to the sick states. Consequently, if health care is a normal good, more health care will be demanded under insurance than in the absence of insurance, even if all the variables are contractible. Consequently, in presence of income effects, previous approaches might have overestimated the welfare loss since the quantity they used as a first best level was smaller than the actual first best. We think that this argument is not new but previous empirical work has neglected it. In fact, when Ellis and McGuire (1993) explain demand side cost-sharing effects, they do not identify the first level of utilization with the utilization without insurance (Figure 2 in their paper). The same idea was already present in the usually neglected but very interesting paper by Meza (1983). In the abstract, one can

[20] We say assumption because it was not obtained from an explicit theoretical model.

read:

"With rare exceptions the provision of actuarially fair health insurance tends to substantially increase the demand for medical care by redistributing income from the healthy to the sick. This suggests that previous studies which attribute all the extra demand for medical care to moral hazard effects may overestimate the efficiency costs of health insurance".

Although the presence of income effects in health care demand is still an unsolved question, previous studies using data from the RAND HIE found positive and significant income effects. Manning and Marquis (1996) estimated income elasticity in 0.22. On the basis of this, it is not surprising that previous studies had found the optimal copayment to be quite large. Buchanan *et al.* (1991) found that the optimal copayment was 100% if the consumer faced a MDE of $1000 and no initial deductible.[21] It was necessary an initial $200 deductible to obtain an optimal copayment of 25%. Manning and Marquis (1996) estimated the optimal copayment to be 45% in an insurance plan without MDE. Feldman and Dowd (1991) found that the 95% HIE insurance plan with its corresponding MDE was superior to the free care plan. The papers above can be criticized on the basis of two considerations. Firstly, all except Manning and Marquis (1996) used the Marshallian measure of welfare loss that does not take into account income effects. Secondly, all of them took as a first level of utilization the health care consumption level attributed to absence of insurance.

Our paper is free of this criticism since our optimal copayment is not derived by comparing a first best level of utilization with the ones predicted by different levels of copayment, but by solving the optimal copayment problem. Therefore, although our estimation of the optimal copayment might be biased downwards due to poor fit of the cost equation, our estimate is consistent with the idea that in presence of income effects the optimal copayment should be smaller than previously estimated ones, since we are not exaggerating welfare losses. This exaggeration in previous studies is due to attributing the first best level of utilization to the health care consumption level in absence of insurance. Therefore, it should be clear the advantage of structural estimation, in that it allows us to solve for the policy function (the optimal copayment) in a theoretically rigorous way.

It must be said that our estimation also differs from previous ones in the kind of

[21] If a consumer faces a deductible of $200 means that she will fully have to pay the first $200 of treatment costs.

spells considered. Previous papers also used utilization and costs that came from chronic routine episodes and inpatient episodes. On one hand we think that chronic routine episodes should not be used in computing the optimal copayment, since they are perfectly predictable at the beginning of the year and therefore they do not involve any risk.[22,23] Certainly they do not fit our theoretical model of demand. On the other hand, inpatient episodes are surely more severe than acute outpatient episodes and in real world the cost sharing associated with them is much smaller. In fact, Newhouse *et al.* (1993) comment that very few hospitalizations were eligible. Therefore, the copayment rates associated with inpatient episodes would be smaller than with outpatient episodes. However, it would be difficult to maintain the assumptions of our model with severe episodes for which the consumer lacks experience about health penalties and costs.[24]

Finally, our approach also differs in the temporal span used. We have estimated a model for a month, not for a full year as the previous papers. It would not be hard to estimate a different structural model for a year, considering the cost as the measure of utilization. However, that would imply assuming that the individual chooses costs directly rather than having treatment or not, which seems less realistic. Furthermore, it would be necessary to make the assumption that larger costs imply larger utility for the individual, which seems to confuse the price of the input with the input itself. We prefer to model an episodic model where costs are seen as the price of the input (treatment) rather than the demanded good by itself. We hope that our approach could be useful to derive multi-episodic models in the future.

4.8 Extensions

We think that increasing the sample size should help to increase the precision of the estimates. Given that the out-of-sample capability of the model seems quite close to the in-sample one, we think that pooling observations of the second, third, and maybe fourth

[22] This would change if we took an ex-ante perspective, that is, to estimate the optimal copayment before nature gives the health stock. However all the previous studies, as ours, condition on current health status.

[23] In fact we are not aware of any theoretical model of optimal insurance that gives the optimal insurance contract when some people in the population knows they will certainly spend a given amount due to routine treatment. We suspect that in this case deductibles and different copayment rates are necesary to separate types. In the future we would like to pursue this investigation.

[24] We think that it is not only a feature of our model but of any economic model that assumes a demand curve. The demand curve must be derived from a utility function and being aware of the valuations of the goods is a necesarry prerrequisite.

and fifth contract years might be useful. In this case one has to deal with correlations in the error terms due to having the same individual in different years. Given that we have two unobservable variables, it might be time consuming but it could be worthy to do it. This would also help to have enough observations to reliably estimate the model that does not exhibit income effects. An alternative is to use a different dataset with a larger sample size and for which one could assume that endogeneity in insurance contract choice is not important.

Another worthwhile extension is to estimate a model with a utility function like $U(y - bs) - s$ which mixes the ones presented. The shortcoming of this specification is that there is not an explicit closed form solution for the health penalty threshold and one needs to use simulation methods. Before considering this, the problem of estimates precision needs to be solved.

4.9 Conclusions

The purpose of this paper is to fill the gap between empirical and theoretical work related to moral hazard in the utilization of health care services. As we have mentioned above, empirical work might have overestimated welfare losses due to moral hazard. This may be due to the fact that in the presence of income effects the optimal level of utilization is larger than previously used (Meza 1983). To do so, we have developed a demand model where the individual takes the decision whether to be treated or not when suffering an illness episode by comparing health benefits and monetary costs of the treatment. We have applied the model to minor episodes coming from the RAND Health Insurance Experiment that randomly assigned individuals to insurance plans. The estimates of the demand model allow us to solve for the optimal copayment of the principal-agent model which is derived from the assumption of non-contractability of the health penalty variable and the estimated demand model.

The model is not rejected statistically, but estimates are still imprecise. Preliminary results show an optimal copayment of 2.1%, much lower than previously estimated. This is consistent with the idea that in the presence of income effects welfare losses due to moral hazard should be smaller than previously estimated. We have found that slight departures from the fair premium hypothesis yield optimal copayments of the size observed in real life. If the results were confirmed, this might indicate that the conventional wisdom

that American families are overinsured must be revisited and more attention should be paid to the administrative costs and inefficiencies of insurance companies.

4.10 Tables

Table 4.1. Distribution of number of episodes that started in the first month and took as much one month

Number of episodes	Number of people
1	178
2	19
3	4

Table 4.2.Distribution of episodes duration in months

Number of months	Number of episodes
1	201
2	5
3	4
4	0
5	5
6	3
7	3
8	0
9	2
10	2
11	3
12	2

Table 4.3. Description of variables

Variable	**Mean**	**Std.**	**Description**
			Endogenous
Treat	0.141	0.348	=1 if treated by an episode that started in the first month, 0 on the contrary
Costs	34.092	62.444	episode treatment cost. 1973 dollars
			Exogenous
Copay	0.264	0.341	copayment rate: 0, 0.25, 0.5, 0.95
Inc	352.079	236.698	montly per-capita family income at enrollement. 1973 dollars
Fem	0.578	0.493	=1 if female
Ghea	0.706	0.147	general health index divided by 100. Bigger value indicates better health
Soc	0.697	0.247	index of social contacts divided by 100 Bigger value indicates more contacts
Dayton	0.157	0.364	=1 if residence is in Dayton
Seattle	0.245	0.430	=1 if residence is in Seattle
Fitchburg	0.132	0.339	=1 if residence is in Fitchburg
Franklin County	0.192	0.394	=1 if residence is in Franklin County
Charleston	0.119	0.324	=1 if residence is in Charleston
Georgetown County	0.151	0.358	=1 if residence is in Georgetown County
Pdn0	0.075	0.263	=1 if individual suffers drinking problems
Educ	1.231	0.301	Number of years of education divided by 10
Disea	1.247	0.890	Index for number of diseases divided by 10 Bigger value indicates more diseases
Age	3.675	1.161	age divided by 10

Table 4.4. Distribution of episodes treated

Copayment	Number of observations	Percentage of episodes treated
0	695	16%
0.25	378	12%
0.50	116	11%
0.95	236	14%
Total	1425	14.1%

Table 4.5. Probit regression for Treat

Vble	Coef	Std. Err	t-stat.
Constant	-1.342	0.384	-3.494
Ghea	-0.209	0.315	-0.662
Soc	0.410	0.180	2.274
Fem	0.234	0.096	2.439
Age	-0.024	0.040	-0.594
Inc*10^{-3}	0.075	0.191	0.394
Pdn0	0.230	0.162	1.415
Educ	0.003	0.153	0.021
Disea	0.096	0.056	1.713
Seattle	0.043	0.140	0.307
Fitchburg	-0.037	0.162	-0.230
Franklin County	-0.034	0.147	-0.233
Charleston	-0.091	0.168	-0.541
Georgetown County	-0.047	0.156	-0.302
Copay=0.25	-0.208	0.104	-2.007
Copay=0.5	-0.174	0.171	-1.023
Copay=0.95	-0.145	0.121	-1.200

Table 4.6. Average costs of those treated

Copayment	Average cost
0	41.35
0.25	32.13
0.5	18.22
0.95	18.08
Number of individuals with episode treated	201

Table 4.7. OLS regression over

$\ln(c)$ of those treated

Vble	Coef	Std. Err	t-stat
Constant	4.014	0.102	39.169
Disea	-0.128	0.015	-8.306
Seattle	0.245	0.038	6.474
Fitchburg	-0.335	0.042	-7.827
Franklin County	-0.118	0.039	-3.004
Charleston	-0.472	0.044	-10.698
Georgetown County	-0.137	0.041	-3.308
Ghea	-0.757	0.086	-8.784
Soc	-0.007	0.046	-0.170
Fem	-0.426	0.024	-17.184
Age	0.006	0.010	0.580
Pdn0	0.026	0.044	0.598
Educ	-0.073	0.039	-1.862
Copay=0.25	-0.032	0.027	-1.203
Copay=0.5	-0.471	0.043	-10.765
Copay=0.95	-0.492	0.032	-15.301
Sample Size	201		

Table 4.8 Structural model estimates

Vbles	Coef	Std.Dev	t-stat
Health penalty equation			
Constant	-4.033	0.805	-5.011
Ghea	-0.679	0.853	-0.797
Soc	1.169	0.538	2.174
Fem	0.605	0.256	2.369
Pdn0	0.566	0.457	1.239
Disea	0.239	0.141	1.7
Risk aversion			
$\ln(\theta)$	-5.787	3.07	-1.885
Cost equation			
Constant	2.694	2.021	1.333
Seattle	0.178	0.262	0.68
Fitchburg	-0.507	0.333	-1.525
Franklin County	-0.202	0.266	-0.761
Charleston	-0.522	0.34	-1.537
Georgetown County	-0.252	0.281	-0.898
Educ	-0.062	0.336	-0.186
Disea	-0.134	0.125	-1.076
Ghea	-0.841	0.519	-1.62
Fem	-0.431	0.283	-1.523
α	0.839	0.722	1.163
Variance covariance elements			
b	0.455	1.094	0.416
c	-0.964	0.349	-2.76

θ=0.00306, Std.Err.(θ) =0.00941 (Delta method).

$\rho = \frac{b}{\sqrt{b^2+c^2}}$ =0.426, Std.Err.(ρ) =0.939 (Delta method).

Table 4.9. In sample forecast (%)

	Probit		Structural	
Real\Predicted	0	1	0	1
0	48.21	37.68	46.52	39.58
1	5.54	8.56	5.47	8.63
Correctly predicted	56.77		54.94	

Table 4.10. Out of sample forecast (%)

	Probit		Structural	
Real\Predicted	0	1	0	1
0	48.79	36.93	46.29	39.43
1	6.64	7.64	6.5	7.79
Correctly predicted	56.42		54.07	

Table 4.11. Results of Montecarlo simulation

Copayment	Frequency of treatment		Costs of treatment	
	Predicted*	Real	Predicted*	Real
0	0.140 0.013	0.159	33.51 6.35	41.35
0.25	0.142 0.017	0.119	32.10 7.99	32.13
0.50	0.131 0.031	0.112	30.53 13.90	18.22
0.95	0.145 0.022	0.135	30.28 8.65	18.08

* Standard deviation of montecarlo series at the bottom row of each cell.

Table 4.12 Selected values of grid search for optimal copayment with actuarially fair premium

Premium	Copayment	Restriction slack	Expected utility
0.01	0.998	0.0012	-0.34514227
1	0.77	0.0013	-0.34468181
2	0.541	0.0016	-0.34438198
3	0.311	0.0003	-0.34420366
4	0.081	0.0009	-0.34412863
4.26	0.021	0.0021	-0.34412268
4.35	0.001	0.0007	-0.34412545

Table 4.13. Optimal copayments for different load factors

Load factor	Premium	Copayment	Expected utility
0	4.26	0.021	-0.34412268
0.05	3.92	0.142	-0.34433521
0.10	3.47	0.275	-0.34451679
0.15	3.07	0.387	-0.34467323

Bibliography

[1] Andrews, D.W.K. (1988), Chi-Square Diagnostic Tests for Econometric Models. Introduction and Applications, Journal of Econometrics, 37, 135-156.

[2] Arrow, K. (1963), Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review,* 53, 941-973.

[3] Biais, B., Bisière, C. and Décamps, J.P. (1999), A Structural Econometric Investigation of the Agency Theory of Financial Structure. GREMAQ, IDEI. Mimeo.

[4] Buchanan, J. L., Keeler, E., Rolph, J. and Holmer, M. (1991), Simulating Health Expenditures under Alternative Insurance Plans. *Management Science,* 37, 1067-1090.

[5] Cameron, A.C., Trivedi, P.K., Milne, F, and Pigott, J. (1988), A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia. *Review of Economic Studies,* 55, 85-106.

[6] Coulson, N.E., Terza, J., Neslulan, C.A. and Stuart, C.B. (1995), Estimating the Moral Hazard Effect of Supplemental Medical Insurance in the Demand for Prescription Drugs by the Elderly, *American Economic Review Papers and Proceedings*, 85, 122-126.

[7] Chiappori, P.A., Durand, F., and Geoffard., P-Y. (1998). Moral Hazard and the Demand for Physician Services: First Lessons from a French Natural Experiment. *European Economic Review,* 42, 499-511.

[8] Chiappori, P.A., Salanié, B. (2000), Testing Contract Theory: a Survey of Some Recent Work. Invited lecture to the World Congress of the Econometric Society.

[9] Deb, P. and Trivedi, P. (1999), The Structure of Demand for Health Care: Latent Class versus Two Part Models. Indiana University Purdue Univiersity. Mimeo.

[10] Feldman, R. and Dowd, B. (1991), A New Estimate of the Welfare Loss of Excess Health Insurance. *The American Economic Review, 81,1, 297-302.*

[11] Feldstein, M. (1973), The Welfare Loss of Excess Health Insurance. *Journal of Political Economy.* 81, 251-80.

[12] Ferrall , C. and Shearer, B. (1999), Incentives and Transaction Costs within the Firm: Estimating an Agency Model Using Payroll Records. *Review of Economic Studies, 66, 309-338.*

[13] Haubrich J. (1994), Risk Aversion, Performance Pay and the Principal-Agent Problem. *Journal of Political Economy.102, 258-276.*

[14] Holly, A., Gardiol, L., Domenighetti, G. and Bisig, B. (1998), An Econometric Model of Health Care Utilization and Health Insurance in Switzerland, *European Economic Review*, 42, 513-522.

[15] Judd, K. (1998), *Numerical methods in economics.* The MIT Press, Cambridge, Massachusetts.

[16] Keane, M. and Wolpin, K. (1997), Introduction to the JBES Special Issue on Structural Estimation in Applied Microeconomics. *Journal of Business and Economic Statistics, 15,2, 111-114.*

[17] Keeler, E., Newhouse, J., ans Phelps, C. (1977), Deductibles and the Demand for Medical Care Services: the Theory of a Consumer Facing a Variable Price Schedule under Uncertainty. *Econometrica.* 45, 641-656.

[18] Ellis, R.P (1986), Rational Behavior in the Presence of Coverage Ceilings and Deductibles, *RAND Journal of Economics*, 158-175.

[19] Ellis, R.P. and McGuire, T. G. (1993), Supply-Side and Demand-Side Cost Sharing in Health Care. *Journal of Economic Perspectives, 7,4, 135-151.*

[20] Guilleskie, D. and Mroz, T (2000), Estimating the Effects of Covariates on Health Expenditures. NBER working paper w7942.

[21] Laffont, J. and Tirole, J. (1993), *A theory of incentives in procurement and regulation.* The MIT Press. Cambridge, Massachusetts.

[22] Lucas, R. (1976), *Econometric policy evaluation: a critique.* Carnegie-Rochester conference series on public policy, 1, 19-46.

[23] Ma, A. C. and Riordan, M. H. (1997), Health Insurance, Moral Hazard and Managed Care. *Industry Studies Program discussion paper 80.* Boston University.

[24] Macho, I. and Pérez, D. (1994), *Introducción a la economía de la información.* Ariel, Barcelona.

[25] Manning, W.G., Newhouse, J. P. *et al.* (1987), Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. *The American Economic Review,* 77, 251-277.

[26] Manning, W. G. and Marquis, M. S. (1996), Health Insurance: the Tradeoff Between Risk Pooling and Moral Hazard. *Journal of Health Economics*, 15, 609-639.

[27] Manning, W. G. and Marquis, M. S. (2001), Health Insurance: Tradeoffs Revisited. *Journal of Health Economics,* 20, 289-293.

[28] Mace (1991), Full Insurance in the Presence of Aggregate Uncertainty. *Journal of Political Economy*, 99, 928-956.

[29] Marigotta and Miller (2000), Managerial Compensation and the Cost of Moral Hazard. *International Economic Review*, 41, 3, 669-719.

[30] Marquis, M.S. and Holmer, M.R. (1996), Alternative Models of Choice under Uncertainty and the Demand for Health Insurance. *Review of Economic and Statistics,* 78(3), 421-427.

[31] Meza, D (1983), Health Insurance and the Demand for Medical Care. *Journal of Health Economics, 2, 47-54.*

[32] Newhouse, J.P.et.al. (1993), *Free for all? Lessons from the Rand Health Insurance Experiment.* Harvard University Press, Cambridge, Massachusetts.

[33] Paarsch, H. and Shearer, B. (2000), Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records. *International Economic Review,* 41,1, 59-92.

[34] Pauly, M. (1968), The Economics of Moral Hazard. A Comment. *The American Economic Review,* 58, *531-537.*

[35] Pauly, M. (1986), Taxation, Health Insurance and Market Failure in the Medical Economy. *Journal of Economic Literature,* 24, *629-675.*

[36] Salanié, B. (1997). *The Economics of Contracts: a primer.* MIT Press, Cambridge, Massachusetts.

[37] Street, A. Jones, A. and Furuta, A (1999), Cost-Sharing and Pharmaceutical Utilization and Expenditure in Russia, *Journal of Health Economics*, 18, 459-472.

[38] Towsend (1994), Risk and Insurance in Village India. *Econometrica*, 62, 3, 539-591.

[39] Van de Voorde, C., Doorslaer., E. V., Shokkaert., E. (2000), Do Doctors Induce Demand to Offset Patient's response to Cost Sharing? Evidence from a Natural Experiment in Belgium. Ninth European Workshop of Econometrics and Health Economics. Tinbergen Institute. Amsterdam.

[40] Vera Hernández, A.M. (1999), Duplicate Coverage and Demand for Health Care. The Case of Catalonia, *Health Economics*, 8, 579-598.

[41] Zeckhauser, R. (1970), Medical Insurance: a Case Study of the Tradeoff Between Risk Spreading and Appropiate Incentives. *Journal of Economic Theory*, 2, 10-26.