Estimating Risk and Risk Aversion in the Automobile Insurance Market

Bruno Cesar Aurichio Ledo*
Caio Matteucci de Andrade Lopes**

Abstract
This study is based on the structural model proposed by Cohen and Einav (2007) to estimate the joint distribution of risk and risk aversion in the automobile insurance market. However, while they estimated the model for a single insurer in the Israeli market, we estimated the model by considering the top five insurers in the Brazilian market at the same time. This difference allowed us to capture the effect of competition on the joint distribution of risk and risk aversion. A counterfactual exercise also allowed us to verify that the insurer with the largest market share is able to implement the optimal contract, while others do not.

Keywords: Automobile insurance, risk, risk aversion, pricing
JEL Codes: D1, C10

1 Introduction
Cars are constantly exposed to the risk of collision, as well as the risks of theft or robbery due to their high added value. Therefore, it is natural for people to have an interest in demanding mechanisms to share such risks: insurance contracts.

A typical auto insurance contract is the one in which the consumer pays a premium to the insurer in all possible states of nature and in return receives indemnities from the insurer only in states of nature where losses occur. Thus, the insurance contract serves to transfer wealth between different states of nature in order to reduce their variance.

Submitted on 27 February 2018; Reviewed on 21 December 2018
* University of São Paulo (USP).
** Federal University of Paraná (UFPR).
bruno@fearp.usp.br caiomdealopes@gmail.com
Consumers demand insurance for two main reasons: first, because they do not like to lose wealth; second, because they do not like uncertainty. In other words, both risk and risk aversion play an important role in the consumer’s decision to purchase an insurance contract. In light of this, the maximization of expected profit on the part of the insurers depends on the available information on risk and risk preferences of the consumers.

This is where the story gets complicated. The first issue stems from the high heterogeneity among consumers. That is, even if insurers had complete information on risk and risk aversion, it would still be difficult to design contracts that would perfectly separate heterogeneous types. The second difficulty stems from the fact that insurers cannot even perceive such consumer characteristics (asymmetric information). Therefore, insurers face a complex problem of maximizing profit in the presence of unobservable heterogeneities in two important dimensions.

In this context, the present study aims to estimate the joint distribution of risk and risk aversion in the auto insurance market. To do so, we will apply the structural model proposed by Cohen and Einav (2007), which will allow us to identify the joint distribution of these two latent variables from the chosen coverage (deductible choice) and the ex-post risk of consumers (number of claims reported).

While the Cohen and Einav (2007) article used data from a single insurer in the Israeli market, this study estimated the model for multiple insurers at the same time. This fact allowed us to capture the effect of competition on the joint distribution of risk and risk aversion.

The results closely resemble those obtained by Cohen and Einav (2007): (i) large heterogeneity in risk preferences across individuals and (ii) a positive correlation between unobserved risk and unobserved risk aversion. From these results, we performed counterfactual exercises and found that, in the presence of competition, the largest market share insurer would be implementing the optimal contract, while the others would not.

The text is organized in 7 sections, including this Introduction. In section 2, a review of the literature on each of the two topics are presented: the estimation of risk aversion and information asymmetry. Section 3 presents the database and its main descriptive statistics. In section 4, we present the theoretical and empirical models. Section 6 presents the main results of the estimation. Section 7 presents counterfactual exercises on the profitability of insurers for different levels of the deductible. Finally, the section 8 presents conclusions.
2 Literature review

The model developed by Cohen and Einav (2007) assumes that consumer preferences can be represented by a von Neumann–Morgenstern expected utility function. Thus, the model directly debates with all literature of choice under uncertainty that uses the expected utility theory (EU).

Within EU modeling, a parametric functional form for utility is usually adopted. Differently from the most common functional forms - such as constant absolute risk aversion (CARA), constant relative risk aversion (CRRA) and the hyperbolic absolute aversion family (HARA)—Cohen and Einav (2007) introduced an approach based on a Taylor expansion of the second order. This approach, later referred to as the negligible third derivative by Barseghyan, Molinari, O’Donoghue, and Teitelbaum (2018), was an important step in research aimed at assessing agent risk preferences.

In this sense, the model of this article approaches the literature whose main objective is to measure risk aversion of individuals. Initially, the scarcity of data on real markets led researchers to develop controlled experiments in order to measure risk aversion (Kachelmeier & Shehata, 1992; Smith & Walker, 1993). Other articles used data from TV programs (Gertner, 1993; Jullien & Salanié, 2000; Metrick, 1995) or hypothetical questionnaires (Barsky, Juster, Kimball, & Shapiro, 1997; Evans & Viscusi, 1991; Viscusi & Evans, 1990), all aiming to measure risk aversion of individuals.

The first study to use real market data to estimate risk aversion was Cicchetti and Dubin (1994). The authors used individual information on fixed telephone insurance purchases to estimate risk preferences. Later, Saha (1997) and Chetty (2006) estimated agent risk aversion by turning to firm production decisions and labor supply respectively. Sydnor (2010) found high values for the absolute risk aversion coefficient from household insurance data in the United States.

Similar to the approach of Cohen and Einav (2007), Barseghyan, Prince, and Teitelbaum (2011) and Barseghyan, Molinari, O’Donoghue, and Teitelbaum (2013) also estimated the preferences under risk using information on households who held auto or home policies. In the first study, the authors proposed a test to verify the stability of preferences from a single database with information about the two markets (cars and houses). In the second, the authors included, through a probability distortion function, additional sources of risk aversion to the methodology of Cohen and Einav (2007). Both articles used the negligible third derivative technique.

In a more recent article, Paravisini, Rappoport, and Ravina (2016) estimated...
the risk aversion of financial market investors. This article emphasized the importance of the unobserved heterogeneity of the parameter of risk aversion in the financial decisions of individuals. Similarly, the model proposed by Cohen and Einav (2007) admits unobserved heterogeneity ex ante the signature of the contract, both on risk aversion and on risk. For this reason, this study also discusses the empirical literature on adverse selection in insurance markets.

In that regard, the article of Chiappori and Salanié (2000) stands out. For parametric and semi-parametric methods, the authors tested and found no empirical evidence of the existence of asymmetric information in the auto insurance market. This approach consists of testing, after controlling for observable variables, the correlation between the level of coverage and the empirical risk of the insured. This procedure is very widespread in the literature (Cawley & Philipson, 1996; Cohen, 2005; Dionne, Michaud, & Dahchour, 2013; Dionne & Vanasse, 1992; Finkelstein & McGarry, 2006; Finkelstein & Poterba, 2004; Puelz & Snow, 1994).

However, it is worth noting that the objective of our study is somewhat different since we intend to estimate the distribution of preference under risk, and, at the same time, to accommodate a possible mechanism of adverse selection through a structural model. Basically, the hypotheses of this method is that the importance of the adverse selection as a result of the unobserved heterogeneities can be evaluated in relation to two aspects: risk and risk aversion. This approach resembles Cardon and Hendel (2001), which model health insurance choices and also consider two dimensions of unobserved heterogeneity.

Finally, it is important to emphasize that our article differs from Cohen and Einav (2007) because it includes a database with several insurers, which allows us to estimate the effect of competition on consumer preferences and also their risk.

3 Data

The database used in this study was prepared by the Superintendence of Private Insurance (SUSEP). Created in 1966 as an autarchy linked to the Brazilian Ministry of Finance, this agency is responsible for the control and oversight of insurance, open private pension, capitalization, and reinsurance markets.

In this study, the data refer to the period from July 1 to December 31, 2010, and are in accordance with the regiment established by the SUSEP Directive n. 360 of February 2008. First, we restricted our analysis to the metropolitan
region of São Paulo, which has the largest participation in the Brazilian auto insurance market (23.8%).

In addition, the sample included only personal policies and non-commercial vehicles (national or imported), with comprehensive coverage.\(^1\) Lastly, we kept only contracts with a bonus class equal to zero in the sample, since these are likely to be new insurers, and therefore without any switching cost.

Originally, the database has three types of deductibles: low, regular and high. Due to the methodology used in this study, the high deductible was considered regular. According to Cohen and Einav (2007), this approach does not cause any kind of bias, since, according to the structural model developed, the individual who chose the high deductible would opt for the regular one instead of the low deductible if only these two options were available. The policies that have been endorsed, that is, any changes during their term, and the collective policies, were disregarded. Finally, we only verified policies for new cars, used for a year or less. At the end of these filters, 89,407 observations remained in the sample.

### 3.1 Individual Characteristics

The variables present in the database can be divided into three groups. First, the covariates, which represent the observed individual characteristics, and car attributes and insurance company chosen. Next, we have the variables related to the contract, that is, premiums and deductibles. Finally, in case of a claim, the number of times the insurance was triggered in each contract is known.

Table 1 is comprised of information on the characteristics of policyholders and vehicles. Insured variables are limited to the gender,\(^2\) age and postal code (ZIP code) of the insured. In the sample, women account for 45\% of insured persons and the average age is 42 years. The base also contains dummy variables for the first three digits of the vehicle usage code. This variable aims to capture the effects of the region of circulation of the vehicle on the risk of an accident. We have a total of 67 dummies, and for this reason, the descriptive analysis of these variables will not be included in this table.

The characteristics of the car are comprised of the assured importance (or simply the value of the vehicle), the age of the car (time of use), dummies indicating whether the car is flex-fuel, has 4 doors and has a turbocharged engine. Also, there is information about the power of the engine (displacement),

---

\(^1\) Comprehensive Coverage is the most used in auto insurance, offering coverage on collision, fire, and theft. It is also broad, as it covers cases of partial and total loss of the vehicle.

\(^2\) The gender variable assumes value 1 for women and 0 for men.
Table 1. Summary Statistics – Covariates.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>S.D.</th>
<th>Min.</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insured</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.452</td>
<td>0.498</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Age</td>
<td>42.012</td>
<td>13.726</td>
<td>18</td>
<td>90</td>
</tr>
<tr>
<td>ZIP Code</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Car attributes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>38610.933</td>
<td>15324.014</td>
<td>2000</td>
<td>325076</td>
</tr>
<tr>
<td>Car age</td>
<td>0.446</td>
<td>0.497</td>
<td>−1</td>
<td>1</td>
</tr>
<tr>
<td>Flex-fuel</td>
<td>0.848</td>
<td>0.359</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4 doors</td>
<td>0.695</td>
<td>0.461</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Engine size (cc)</td>
<td>2.102</td>
<td>1.32</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Turbo</td>
<td>0.002</td>
<td>0.04</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Car usage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>weekends</td>
<td>0.336</td>
<td>0.472</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>daily</td>
<td>0.623</td>
<td>0.485</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>work</td>
<td>0.039</td>
<td>0.194</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Car manufacturer</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Insurance company</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurer 1</td>
<td>0.672</td>
<td>0.469</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Insurer 2</td>
<td>0.225</td>
<td>0.418</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Insurer 3</td>
<td>0.085</td>
<td>0.279</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Insurer 4</td>
<td>0.015</td>
<td>0.121</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Insurer 5</td>
<td>0.003</td>
<td>0.051</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Notes: 89,407 observations in the sample. *Value relative to the insured amount of the bodywork.*
about the use made of the vehicle (weekends, daily or for work) and also the car manufacturer. The last part of Table 1 shows the dummies referring to insurers. Insurance company 1, which has the largest market share, is present in 67.2% of the sample. Further details can be seen in Figure 1.

![Figure 1. The insurance companies participation in sample.](image)

### 3.2 Premium and Deductible

After knowing the vector of individual characteristics and the car, $x_i$, the insurer offers a menu with two contract options. Among them, the regular deductible is the most chosen among the insured and relatively similar among insurers. The low deductible value corresponds to 50% of the regular deductible value, so a low deductible contract offers greater coverage (because it compensates for a greater range of losses).

Given the coverage, the total premium observed by the insured is composed of the sum of the insurer’s premium plus the broker’s commission. Because the insurer cannot choose for itself the premium difference between its contracts, we had to estimate this difference in premiums from the choices made by policyholders. To do so, we regressed the logarithm of the total insurance premium against a dummy variable and a set of control variables, that is:

$$\ln p_i = \theta F_i + x_i'\phi + \eta_i,$$

where $p_i$ represents the total premium of the chosen contract. The dummy $F_i$ represents the type of deductible chosen so that it assumes value one for the
low deductible. The control variables, $x'_i$, consist of information relevant to the pricing of contracts by insurers. We use the log model since the estimated value of parameter $\theta$ already reflects an approximation of the effect of changing the chosen deductible—from regular to low—on the value of the premium.

The regression of the equation (1) was done for the five insurers with the largest participation in the period. As previously mentioned, Figure 1 shows their market share in the selected database and was calculated by the frequency of each insurer’s contracts in the sub-sample. Table 2 presents the estimated coefficient of dummy type for each of the five insurers analyzed.  

Table 2. Estimation of premium variation in relation to deductible choice.

<table>
<thead>
<tr>
<th>Insurer 1</th>
<th>Insurer 2</th>
<th>Insurer 3</th>
<th>Insurer 4</th>
<th>Insurer 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>0.0302*</td>
<td>0.0312*</td>
<td>0.2063*</td>
<td>0.1591*</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>(0.0015)</td>
<td>(0.0056)</td>
<td>(0.0489)</td>
<td>(0.0126)</td>
</tr>
</tbody>
</table>

Notes: Standard deviations in parentheses. *Significant at the 5% confidence level.

The results found indicate that low deductible contracts are between 3% and 21% more expensive than regular deductible ones, depending on the insurer. Thereby, one can calculate the premium-deductible values observed by policyholders before making their choices.

The descriptive analysis concerning these data can be observed in Table 3. It is worth mentioning that the variable Claims per semester represents the number of times the insurance was triggered divided by the exposure time of the policy. That is, if the policy remained active throughout the analyzed period (second half of 2010), the exposure was 1. Similarly, if the contract lasted for only one day in the semester, the exposure was $1/181$, where 181 is the number of days in the second half of 2010. From the evaluation of this variable, it can be observed that the semiannual rate of activation of insurance, in relation to all individuals, was 0.023. When we focus on just one type of deductible, the data show that the drive per semester was roughly the same for the two types of coverage.

According to Table 3, on average, an insured who chooses the reduced-deductible contract pays R$58 ($\Delta p$) more, but saves R$1,008.97 ($\Delta d$) in case of loss—ratio $\Delta p/\Delta d = 0.61$. If the insured is risk-neutral, we can say that

---

3 It is important to note that another insurer was among the largest market participants, but the estimate of its $\theta$ coefficient presented a non-significant value. This result can be explained by the fact that the insurer in question is a financial institution (bank) and therefore offers other products on the market, which may influence its policy-pricing strategy.
### Table 3. Summary Statistics – Menu, Choices and Claims.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>S.D.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deductible</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1008.975</td>
<td>419.17</td>
<td>192.5</td>
<td>10703.5</td>
<td>89407</td>
</tr>
<tr>
<td>Regular</td>
<td>2017.951</td>
<td>838.341</td>
<td>385</td>
<td>21407</td>
<td>89407</td>
</tr>
<tr>
<td><strong>Premium</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1278.371</td>
<td>652.442</td>
<td>6.191</td>
<td>14380.326</td>
<td>89407</td>
</tr>
<tr>
<td>Regular</td>
<td>1220.425</td>
<td>625.717</td>
<td>6</td>
<td>13952</td>
<td>89407</td>
</tr>
<tr>
<td>Δp/Δd</td>
<td>0.061</td>
<td>0.075</td>
<td>0</td>
<td>1.233</td>
<td>89407</td>
</tr>
<tr>
<td><strong>Deductible</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Deductible</td>
<td>0.165</td>
<td>0.371</td>
<td>0</td>
<td>1</td>
<td>89407</td>
</tr>
<tr>
<td>Regular Deductible</td>
<td>0.835</td>
<td>0.371</td>
<td>0</td>
<td>1</td>
<td>89407</td>
</tr>
<tr>
<td><strong>Claims</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both</td>
<td>0.023</td>
<td>0.155</td>
<td>0</td>
<td>3</td>
<td>89407</td>
</tr>
<tr>
<td>Low</td>
<td>0.023</td>
<td>0.161</td>
<td>0</td>
<td>3</td>
<td>14769</td>
</tr>
<tr>
<td>Regular</td>
<td>0.023</td>
<td>0.154</td>
<td>0</td>
<td>3</td>
<td>74638</td>
</tr>
<tr>
<td><strong>Claims per semester</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both</td>
<td>0.039</td>
<td>0.713</td>
<td>0</td>
<td>181</td>
<td>89407</td>
</tr>
<tr>
<td>Low</td>
<td>0.036</td>
<td>0.342</td>
<td>0</td>
<td>16.455</td>
<td>14769</td>
</tr>
<tr>
<td>Regular</td>
<td>0.039</td>
<td>0.765</td>
<td>0</td>
<td>181</td>
<td>74638</td>
</tr>
</tbody>
</table>

*The average and standard deviation of the claims per semester are weighted by the duration of the policy in order to adjust for variations in the exposure period.*
he will only choose the reduced-deductible contract if his probability of loss is greater than 0.61. However, the average risk is only 0.023. So, this contract is not actuarially fair. Still, 16% of people chose it, indicating that risk aversion can play a very important role in this decision. Hence the importance of the deductible choice model (in the presence of unobservable risk aversion and unobserved risk) developed in the next section.

4 Deductible choice model

The theoretical model developed by Cohen and Einav (2007) is based on the idea of the indifferent agent between two contracts. These correspond to a pair premium and deductible, such that \((p_i^h, d_i^h)\) and \((p_i^l, d_i^l)\) represent the contracts for the individual \(i\) of regular and low deductibles, respectively. In addition, let \(w_i\) be the wealth of the individual \(i\) and \(u_i(w)\) its corresponding utility function of the vNM type. The contract time is represented by \(t_i\). Assume that the insurance is triggered according to a Poisson distribution with an annual rate, \(\lambda_i\). In other words, \(\lambda_i\) is the risk inherent in each individual and, by hypothesis, self-knowledge. It is also assumed that \(\lambda_i\) is independent of the franchise choice, i.e. there is no moral hazard. Lastly, the latter hypothesis states that in the event of an accident, the indemnity paid must be over \(d_i^h\). For the rest of this section, the \(i\) subscript will be omitted for convenience.

This model establishes that both the premium and the risk are proportional to the contract time. Cohen and Einav (2007) explain that this approach has three advantages. The first is that it helps to deal with canceled contracts, or those that last for a shorter period of time. The second refers to the deductible choice to be independent of the long-term uncertainties, enabling individuals to focus on short-term risk preferences. The third advantage results from an analytical and computational convenience.

The expected utility that the individual obtains from the choice of contract \((p,d)\) is given by:

\[
v(p,d) \equiv (1 - \lambda t)u(w - pt) + (\lambda t)u(w - pt - d).
\] (2)

With this, one can characterize the set of parameters that makes the individual indifferent between regular and low deductible contract. This allows one to set a lower (upper) limit for the level of risk aversion of individuals who chose low (regular) deductible for a given \(\lambda\). Applying the limit in relation to \(t\) and using
the L’Hospital rule, we observe:

\[
\lambda = \lim_{t \to 0} \frac{\frac{1}{t} (u(w - ph^t)) - u(w - pl^t)}{(p^f - ph^t)u'(w) - u(w - d^f) - u(w - pl^t - dh^t)}.
\]

Rearranging,

\[
(p^f - ph^t)u'(w) = \lambda (u(w - d^f) - u(w - d^h)).
\]  

The expression (4) has a simple interpretation: the right side represents the expected utility gain per unit time of choosing the low franchise, while the left side equals the cost of that choice per unit. In order for the individual to be indifferent between the two contracts, the expected gains must be equal to the costs. The absolute risk aversion coefficient of the indifferent individual can be calculated from the hypothesis that the third derivative of the utility vNM is not very large. Thus, by applying a Taylor expansion in the two terms on the right-hand side of the equation (4) one obtains, in a general way,

\[
\frac{p^f - ph^t}{\lambda} u'(w) \approx (d^h - d^f) u'(w) - \frac{1}{2} (d^h - d^f)(d^h + d^f) u''(w).
\]

Renaming variables, \(\Delta d \equiv d^h - d^f > 0\), \(\Delta p \equiv p^f - ph^t > 0\) e \(\dd \equiv \frac{1}{2} (d^h + d^f)\), we got:

\[
\frac{\Delta p}{\lambda \Delta d} u'(w) \approx u'(w) - \dd u''(w)
\]  
or  
\[
\frac{u''(w)}{u'(w)} \approx \frac{\Delta p}{\Delta d} - \frac{1}{\dd},
\]

where \(r\) is the absolute risk aversion coefficient given to wealth \(w\). Thus, the equation (7) defines the indifference set that relates, from the premium and franchise data, the risk variables and absolute risk aversion coefficient \((r^*(\lambda), \lambda)\). Both are specific to individuals, since they depend on the choice of the menu of contracts, which varies among the insured. In this way, it is an individual \(i\) represented by a pair \((r_i, \lambda_i)\), which is offered a menu of contracts \(((p_i^h, d_i^h), (p_i^f, d_i^f))\), then the latter will choose the low franchise agreement if and only if its coefficient of aversion to the absolute risk satisfies \(r_i > r_i^*(\lambda)\).
5 Econometric Model

The objective is to estimate the joint distribution between risk and the absolute risk aversion coefficient, \((\lambda_i, r_i)\), in the insured population, conditional on the observed variables. For this, it is assumed that \((\lambda_i, r_i)\) follows a bivariate lognormal distribution, so that

\[
\begin{align*}
\ln \lambda_i &= x_i' \beta + \epsilon_i, \\
\ln r_i &= x_i' \gamma + \nu_i,
\end{align*}
\]

with

\[
\begin{pmatrix}
\epsilon_i \\
\nu_i
\end{pmatrix}
\overset{iid}{\sim} N\left(0, \begin{bmatrix}
\sigma_{\lambda}^2 & \rho \sigma_{\lambda} \sigma_{r} \\
\rho \sigma_{\lambda} \sigma_{r} & \sigma_{r}^2
\end{bmatrix}\right),
\]

where \(r\) and \(\lambda\) are unobserved latent variables. Therefore, in order to estimate the joint distribution of such variables, it is necessary to define the relation of these variables with those observed. First, it is assumed that the number of claims performed by the individual \(i\) results from a Poisson distribution, such that

\[
\text{claims}_i \sim \text{Poisson}(\lambda_i, t_i),
\]

where \(t_i\) is insurance contract time. The absolute risk aversion coefficient is related to the choice of deduction—low or regular—through the theoretical model, so that the individual will choose the contract with greater coverage, and therefore lower deductibility, if its coefficient of aversion to the absolute risk is greater than the limit stipulated by the theoretical approach. This is,

\[
\begin{align*}
\Pr(\text{choice}_i = 1) &= \Pr \left( r_i > \frac{\Delta p_i}{\lambda_i \Delta d_i} - 1 \right) \\
&= \Pr \left( \exp(x_i' \gamma + \nu_i) > \frac{\exp(x_i' \beta + \epsilon_i) \Delta d_i}{\Delta d_i} - 1 \right). 
\end{align*}
\]

The equation (12) considers that the individual chooses the contract with the highest coverage—reduced franchise—only if \(r_i > r_i^*(\lambda_i)\) is defined by the equation (7). Since there is unobserved heterogeneity in \(\lambda_i\), i.e., \(\epsilon_i \neq 0\), this model relies on factors not observed by the insurer to explain the individual risk. In other words, this approach admits adverse selection. Otherwise, the equation (12) would be reduced to a Probit, since the risk would be perfectly estimated by means of the observed variables, \(\hat{\lambda}(x_i)\).
Estimating Risk and Risk Aversion in the Automobile Insurance Market

The likelihood function of the model described in this section is represented by

\[ L(\text{claims}_i, \text{choice}_i | \theta) = \Pr(\text{claims}_i, \text{choice}_i | \lambda_i, r_i)\Pr(\lambda_i, r_i | \theta), \]

where \( \theta \) is the vector of parameters to be estimated. However, estimation via maximum likelihood is not trivial. Due to the existence of unobserved heterogeneity in the risk and also in the aversion to the risk, the estimation becomes a computationally painful process, since it is necessary to realize the integration in relation to the two dimensions. In contrast, the Gibbs sampling, which uses Monte Carlo method via Markov Chains (MCMC), is very compelling for this case. Cohen and Einav (2007) argue that this methodology is ideal for this case because it allows the increase of latent variable data (Tanner & Wong, 1987). Thus, we can simulate \((\lambda_i, r_i)\) and then treat such simulations as part of the data. In addition, the log-normality hypothesis implies that \(F(\ln(\lambda_i) | r_i)\) and \(F(\ln(r_i) | \lambda_i)\) follow a normal distribution, which contributes to decrease computational effort.

The Gibbs sampling methodology is described in Appendix A. His basic intuition is to regress the conditional 8 and 9 equations to \(\lambda_i\) and \(r_i\) for each individual. To obtain random observations about \((\lambda_i, r_i)\), several iterations are performed. Conditional to \(\lambda_i\), the posterior distribution of \(\ln(r_i)\) follows a truncated normal distribution, where the truncation point depends on the menu offered to the insured and its direction (if the distribution is above or below the truncation point) comes from the choice of the franchise. Collecting a sample from the posterior distribution of \(\ln(\lambda_i)\) conditional to \(r_i\) is more complicated since there are two truncation points. The first one stems from the adverse selection (similar to \(r_i\)) and the second one is due to the hypothesis of the equation (11) on the distribution of the number of actuations carried out, which brings additional information about the distribution to posteriori of \(\lambda_i\). To obtain a sample on the unknown distribution of \(\lambda_i\), the “sliced sampler” (Damlen, Wakefield, & Walker, 1999) will be used.

The results presented in this study are the result of 100,000 iterations of the Gibbs sampling. Since the estimate of latent variable distribution depends on an initial kick and converges after a number of iterations, the first 10,000 iterations were discarded (“burn-in”).

5.1 Model with Competition

The main difference between our papers and Cohen and Einav (2007) is that we have included dummies per insurer in equations of risk and risk aversion.
Bruno C.A.Ledo and Caio M. de Andrade Lopes

This approach will be consistent only if the allocation of insurees among different insurers is not correlated with their risk and risk aversion.

If the insurer’s information sets are homogeneous, with common information about risk and risk aversion of insurees, then differences between their contract menus will depend on other factors (such as pricing technology). Therefore, the allocation of insurees among different insurers will not correlate with their risk or risk aversion.

It is worth remembering that the allocation of insurees among different contracts depends on their risk and risk aversion. However, what we are suggesting here is that the allocation of insurees among different insurers is independent of risk and risk aversion.

One anecdotal evidence that supports this argument is that insurees almost do not switch insurance companies, despite the fact that their risk is in constant changing. In our sample, for example, more than 95% of insurees remained under the same insurer when renewing their contracts.

6 Results

This section presents our main contribution. In estimating the model with multiple insurers, we allowed it to capture the effect of competition on the distribution of risk and risk aversion. As we shall see, this approach brought qualitatively different results from those obtained in the previous section.

6.1 Reduced-Form Estimation

An initial idea regarding the level of absolute risk aversion can be obtained through the non-conditional mean values of $\Delta p$, $\Delta d$, $\lambda$ and $\bar{d}$. By replacing these values in the equation (7), we obtain a risk aversion coefficient of 0.0006, which can be interpreted as a point of average indifference. Since 16.5% of policyholders have chosen low deductible policies, 16.5% of policyholders have a risk aversion coefficient greater than 0.0006.

Table 4 presents the results of 3 econometric models in reduced form. Column (1) reports the results of the model whose dependent variable is the number of claims reported in the policy. This model is directly related to the equation (8). The results suggest that older insureds are less likely to report a claim.

---

4 The non-conditional value of $\lambda$ was computed by maximum likelihood, using the data on claims and duration of the policies.
Table 4. Reduced Form.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Poisson regression(^a)</th>
<th>Probit regression(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dep. var: number of claims</td>
<td>Dep. var: 1 if low deductible</td>
</tr>
<tr>
<td></td>
<td>Coef. (S.D.)</td>
<td>IRR(^b) (1)</td>
</tr>
<tr>
<td>Constant</td>
<td>(-2.3259 (2.0261)^*)</td>
<td>–</td>
</tr>
<tr>
<td>In (\hat{\lambda})</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Insured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.0468 (0.0466)</td>
<td>1.0479</td>
</tr>
<tr>
<td>Age</td>
<td>(-0.0488 (0.0112)^*)</td>
<td>0.9523</td>
</tr>
<tr>
<td>Age2</td>
<td>0.0004 (0.0001)^*</td>
<td>1.0004</td>
</tr>
<tr>
<td>ZIP code(^d)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Car attributes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log (Value)</td>
<td>0.0000 (0.1960)</td>
<td>0.9999</td>
</tr>
<tr>
<td>Car age</td>
<td>(-0.0230 (0.0464))</td>
<td>0.97725</td>
</tr>
<tr>
<td>Flex-fuel</td>
<td>0.3041 (0.1036)^*</td>
<td>1.35547</td>
</tr>
<tr>
<td>4 doors</td>
<td>0.0041 (0.0603)</td>
<td>1.004182</td>
</tr>
<tr>
<td>1200cc</td>
<td>Omitted</td>
<td>–</td>
</tr>
<tr>
<td>1400cc</td>
<td>0.2134 (0.10013)^*</td>
<td>1.237903</td>
</tr>
<tr>
<td>1600cc</td>
<td>0.0635 (0.0897)</td>
<td>1.065639</td>
</tr>
<tr>
<td>1800cc</td>
<td>0.1822 (0.1632)</td>
<td>1.199932</td>
</tr>
<tr>
<td>2000cc</td>
<td>0.1822 (0.1632)</td>
<td>1.32028</td>
</tr>
<tr>
<td>2000+ cc</td>
<td>0.1843 (0.3440)</td>
<td>1.202434</td>
</tr>
<tr>
<td>Turbo</td>
<td>0.4613 (0.6343)</td>
<td>1.586227</td>
</tr>
<tr>
<td>Weekends</td>
<td>Omitted</td>
<td>–</td>
</tr>
<tr>
<td>Daily</td>
<td>(-0.2981 (0.0720)^*)</td>
<td>0.7421569</td>
</tr>
<tr>
<td>Work</td>
<td>(-0.0335 (0.1168))</td>
<td>0.96700</td>
</tr>
<tr>
<td>Manufacturer(^e)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Insurance company</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurer 1</td>
<td>Omitted</td>
<td>–</td>
</tr>
<tr>
<td>Insurer 2</td>
<td>(-18.9835 (470.3895))</td>
<td>0.0000</td>
</tr>
<tr>
<td>Insurer 3</td>
<td>(-0.1903 (0.0832)^*)</td>
<td>0.8267</td>
</tr>
<tr>
<td>Insurer 4</td>
<td>(-0.4977 (0.2362)^*)</td>
<td>0.6078</td>
</tr>
<tr>
<td>Insurer 5</td>
<td>0.2345 (0.3190)</td>
<td>1.2643</td>
</tr>
</tbody>
</table>

Notes: Standard deviations in parentheses.  
\(^a\)Significant at 5%.  
\(^b\)Estimation by maximum likelihood considering variations in the time of exposure of contracts.  
\(^c\)Two different regressions were performed. The first, probit (column 2) used an additional regressor (with \(\hat{\lambda}\) predicted from column 1),  
\[ \ln\left( \frac{\Delta p_i}{\left(\hat{\lambda}(x_i)\Delta d_i\right)} - 1 \right) / \Delta d_i \]  
which enables a structural interpretation and a comparison between these coefficients and that of the main model. However, it is worth noting that this regression does not allow heterogeneities unobserved in risk. Column 3 reports the marginal effects of the probit without the additional regressor, making it possible to compare the results of the two columns.  
\(^d\)ZIP code variables were used as controls, but the results were not reported.  
\(^e\)The car manufacturer variables were used as control, but the result was not reported.
Column (2) presents probit estimates in which the dependent variable assumes value 1 if the insured chooses the reduced deductible. The variable \( \ln \left( \frac{\Delta p_i}{(\hat{\lambda}(x_i)\Delta d_i) - 1} / d_i \right) \) was used as regressor, and the estimate of \( \hat{\lambda}(x_i) \) was obtained by computing the predicted value of column (1) regression.

The results present in the table are in line with the estimates of the main model, with women and older insured being the most risk-averse. Column (2) also reports values of the marginal effects of the regressors on the probability of choosing a low deductible policy, as well as column (3), albeit it does not use \( \ln r_i \) as a regressor of its corresponding probit, thus not having an interpretation.^[5]

### 6.2 Bayesian Estimation

For the estimation of the model, we performed the normalization of the control variables \( X \), so the interpretation must be made in terms of standard deviations. This approach makes the scale of the regressors irrelevant, which allows high comparability between the coefficients.

The results of the model are included in Table 5. Taking into account only the variables of the risk equation, the probability of an accident is higher for the female sex (about 22% higher). In relation to the age group, it is worth noting that older insured persons have a lower risk of an accident. This result may be associated with the level of driving experience. The other control variables—insured importance, which acts as a proxy for the car value, year of the vehicle model, and engine power—had a significant effect on the probability of an accident. The dummies related to the first three digits of the postal address explain the risk, since the region where the vehicle circulates interferes with the probability of an accident.

When focusing on the risk aversion equation we notice some interesting results. The gender coefficient suggests that women’s absolute risk aversion is not statistically different from that of men. This result is in disagreement with that obtained by Cohen and Einav (2007) who found a coefficient 20% higher for women. The age variable shows that the older the insured, the higher their risk aversion.

In analyzing the insurer’s dummies, we detect significant differences in risk and risk aversion in the insurers’ portfolios. This result suggests that pricing algorithms may be heterogeneous among insurers.

[^5]: Column (3) uses all insureds at the base but admits that the choice of coverage does not depend on price, risk, and risk aversion.
### Table 5. Bayesian estimation.

| Variables | \( \ln \lambda \) & \( \ln r \) | Other Results |
|-----------|-----------------|-----------------|-----------------|
| **Insured** | | | variance-covariance matrix |
| Constant | \(-5.0518 (0.0343)^*\) | \(-12.5913 (0.2381)^*\) |
| Female | 0.2192 (0.0182)^* | 0.0799 (0.0646) |
| Age | \(-0.0345 (0.0044)^*\) | 0.2016 (0.0160)^* |
| Age2 | 0.0003 (0.0000)^* | \(-0.0016 (0.0002)^*\) |
| ZIP code\(^a\) | – | – |
| **Car attributes** | | | Unconditional statistics |
| Log (Value) | 1.0350 (0.0712)^* | \(-2.4015 (0.2301)\) |
| Car Age | 0.0437 (0.0195)^* | 1.1434 (0.0779) |
| Flex | 0.1249 (0.0327)^* | – |
| 4 Doors | \(-0.1208 (0.0227)^*\) | 0.3935 (0.0919)^* |
| 1200-cc Omitted | \(-0.0457 (0.0368)\) | 0.5989 (0.1390) |
| 1400 cc | \(-0.0921 (0.0350)^*\) | 1.2053 (0.1234) |
| 1600 cc | \(-0.3546 (0.0654)^*\) | 1.9207 (0.2106)^* |
| 1800 cc | \(-0.2791 (0.0641)^*\) | 2.2296 (0.2051)^* |
| 2000 cc | \(-0.9354 (0.1318)^*\) | 3.5166 (0.5119)^* |
| 2000+ cc Turbo Omitted | \(-0.2735 (0.2572)\) | 0.5867 (0.9089) |
| weekends daily work Omitted | 0.2792 (0.0321) | 6.5131 (1.5244)^* | 0.3311 (0.0554) * | 6.8762 (1.5275)^* |
| **Insurance company** | Omitted | Omitted |
| Insurer 1 | Omitted | Omitted |
| Insurer 2 | \(-1.4291 (0.1174)^*\) | 13.1609 (1.4620)^* |
| Insurer 3 | 0.6263 (0.0658)^* | 1.8362 (0.3814)^* |
| Insurer 4 | 1.1450 (0.1640)^* | 6.7018 (2.6941)^* |
| Insurer 5 | 0.9965 (0.1341)^* | \(-19.4148 (13.8198)\) |

Notes: Standard deviations in parentheses, obtained from samples of the posterior distribution.

*Significant at 5%.

\(^a\) We chose not to report the coefficients of 67 ZIP code dummies.

\(^b\) We chose not to report the coefficients of 17 Car manufacturers dummies.
At each iteration of the Gibbs sampling, the mean and standard deviation of the random sampling of $\lambda_i$ and $r_i$ were computed, as well as the correlation between these variables. Table 5 reports the averages and standard deviations of the quantities computed at each iteration of the Gibbs sampling. Thus, these estimates are not conditional to the observable characteristics of the insured, i.e., it is not possible to obtain them directly from the parameters. The average of the absolute risk aversion coefficient is 0.0051 and the median is very close to zero.

In addition, a high unobserved heterogeneity of risk aversion ($\sigma_r$) and risk ($\sigma_\lambda$) is observed, but the former is higher than the latter. This result corroborates that obtained by Cohen and Einav (2007). The authors argue that this by-product of the estimation indicates that the unobserved heterogeneity of risk aversion is a more important source of selection than risk.

Finally, Table 5 also points to a positive correlation of 0.471 between unobserved risk aversion and unobserved risk, $\rho$, and an unconditional correlation of $-0.031$ between $\lambda$ and $r$. At first, this second result is intuitive since it is natural to assume that individuals who are more risk averse will have more cautious attitudes and therefore are less risky. Finkelstein and McGarry (2006) observed this same evidence analyzing the American health insurance market. However, Cohen and Einav (2007) themselves list reasons why the correlation signal is ambiguous.

First, in the auto insurance market, agent risk does not depend only on precautionary attitudes or inherent insured factors, such as the innate ability to drive, but also on the interaction of drivers’ driving habits. Second, there are unobserved factors that may be positively correlated to the two dimensions (risk and aversion). For example, if more risk-averse individuals drive more often, then they are at greater risk. Finally, the non-conditional correlation between the variables is sensitive to the hypothesis performed on the distribution of the claims and is subject to the covariates’ coefficients. That is, the coefficients of the same covariate can affect risk and risk aversion in the same direction, so even if $\rho$ is negative, the correlation $(\lambda, r)$ may not be.

Thus, the results presented in this paper point to a positive correlation for unobserved factors, $\rho$, and negative for the two variables of interest $(\lambda, r)$. In relation to the dispersion of the absolute risk aversion coefficient, Table 6 reports the interpretation of the different percentiles of the nonconditional estimation of risk aversion. The interpretation was obtained for a utility function of the quadratic type, $\{u(w) = w - bw^2\}$, which holds the hypothesis that the

<table>
<thead>
<tr>
<th>Specification</th>
<th>Absolute risk aversion (R$)</th>
<th>Interpretation (R$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimating by average</td>
<td>$6 \times 10^{-4}$</td>
<td>94.33</td>
</tr>
<tr>
<td>Bayesian inference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$5.1 \times 10^{-3}$</td>
<td>64.04</td>
</tr>
<tr>
<td>25° percentile</td>
<td>$2.9 \times 10^{-7}$</td>
<td>99.99</td>
</tr>
<tr>
<td>Median</td>
<td>$3.6 \times 10^{-6}$</td>
<td>99.96</td>
</tr>
<tr>
<td>75° percentile</td>
<td>$8.4 \times 10^{-5}$</td>
<td>99.16</td>
</tr>
<tr>
<td>90° percentile</td>
<td>$2.7 \times 10^{-3}$</td>
<td>78.23</td>
</tr>
<tr>
<td>95° percentile</td>
<td>$1.3 \times 10^{-2}$</td>
<td>29.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Specification</th>
<th>Absolute risk aversion (US$)</th>
<th>Interpretation (US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen e Einav</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$6.7 \times 10^{-3}$</td>
<td>56.05</td>
</tr>
<tr>
<td>25° percentile</td>
<td>$2.3 \times 10^{-6}$</td>
<td>99.98</td>
</tr>
<tr>
<td>Median</td>
<td>$2.6 \times 10^{-5}$</td>
<td>99.74</td>
</tr>
<tr>
<td>75° percentile</td>
<td>$2.9 \times 10^{-4}$</td>
<td>97.14</td>
</tr>
<tr>
<td>90° percentile</td>
<td>$2.7 \times 10^{-3}$</td>
<td>78.34</td>
</tr>
<tr>
<td>95° percentile</td>
<td>$9.9 \times 10^{-3}$</td>
<td>49.73</td>
</tr>
</tbody>
</table>

Notes: a For the interpretation of Absolute Risk Aversion (ARA) we calculate \( x: u(w) = \frac{1}{2} u(w + 100) + \frac{1}{2} u(w - x) \). Thus, for different levels of ARA we reported values to \( x \), it means that the individual is indifferent between bet in the lottery having 50% chances of win R$100.00 and 50% chances of lose an \( x \) amount. For this interpretation, we assume the quadratic utility of individual as \( \{u(w) = w - bw^2\} \). Also, it is important to note that this interpretation does not depends of individual level of wealth, because the analysis begins from coefficient of absolute risk aversion.

Thus, an insured with the average absolute risk aversion coefficient will be indifferent between whether or not to participate in a lottery in which he has a 50% chance of winning R$100.00 and a 50% chance of losing R$64.03. Cohen and Einav (2007) have found that the average individual is indifferent to participating (or not) in a 50-50 lottery in which he earns US$ 100.00 or loses US$ 56.05. Although the comparison is not immediate due to the difference in the unit of measure (US$ and R$), in both results the median individual is almost risk neutral, although the average insured has a considerable degree of aversion to lotteries of this magnitude.

7 Counterfactual Analysis of Profits

This section addresses what would happen to the profitability of insurers if they changed their contract menu.
We know that each insurer offers a menu of contracts of the type \((p^h, d^h)\) and \((p^L, d^L)\), such that \(d^L < d^h\) and \(p^L > p^h\). To make the counterfactual, we will vary the regular deductible \((d^h)\) value while maintaining the premium \((p^h\) and \(p^L)\) and the low deductible \((d^L)\) ones fixed. The question is: what is the impact of this variation on the profits of insurers?

Suppose that each insured person can be represented by a random draw \((\lambda_i, r_i)\) of the following joint distribution:

\[
\begin{pmatrix}
\ln \lambda_i \\
\ln r_i
\end{pmatrix}
\sim \mathcal{N}
\begin{pmatrix}
x_i' \hat{\beta} \\
x_i' \hat{\gamma}
\end{pmatrix}
\begin{bmatrix}
\hat{\sigma}_\lambda^2 & \hat{\rho} \hat{\sigma}_\lambda \hat{\sigma}_r \\
\hat{\rho} \hat{\sigma}_\lambda \hat{\sigma}_r & \hat{\sigma}_r^2
\end{bmatrix}.
\]

(13)

After drawing your type of risk and risk aversion, the insured visits the insurers and faces the menu of contracts of each company. Then the insured chooses the contract that maximizes their expected utility. Let \(\pi_0\) be the expected profit by the chosen insurer if the insured chooses the contract with regular deductible \((p^h, d^h)\). Assuming the chosen insurer is risk-neutral, your expected profit may be written as

\[
\max_{\Delta d, \Delta p} \pi_0 + \Pr(r_i > r^*_i(\lambda_i; \Delta d, \Delta p)) \left\{ \Delta p \right. \\
- \Delta d \cdot \mathbb{E}[\lambda_i | r_i > r^*_i(\lambda_i; \Delta d, \Delta p)]
\}
\]

(14)

where \(\Delta d = d^h - d^l\), \(\Delta p = p^l - p^h\), and \(r^*_i(\lambda_i; \Delta d, \Delta p)\) from equation (7).

Equation (14) reflects two important tradeoffs that the insurer faces. The first one is related to demand. The greater the difference in the premiums, the greater the revenue of the insurer with the sale of the contract \((p^l, d^l)\), however, the probability of this contract being sold will decrease—measured by the term \(\Pr(r_i > r^*_i(\lambda_i; \Delta d, \Delta p))\). The second tradeoff is related to the problem of adverse selection. As the difference in prices increases, those who continue to demand the contract \((p^l, d^l)\) will be those with greater risks—measured by the term \(\mathbb{E}[\lambda_i | r_i > r^*_i(\lambda_i; \Delta d, \Delta p)]\). The magnitude and signal of these effects depend on the relative heterogeneity of \(\lambda_i\) e \(r_i\), as well as the sign of the correlation between them.

To illustrate, Figure 2 presents the histogram of 100,000 random draws of the joint distribution of risk and risk aversion related to the insurer with the highest market share on the market.

We simulated 100,000 draws and calculated the expected profit of each insurer (equation (14)) at different levels of \(\Delta d\) while keeping \(\Delta p\) unchanged.
Then we calculated the average profit for each level of $\Delta d$. The result of this counterfactual exercise is in Figure 3. The first thing we notice is that the leading insurer is the only one who is able to implement the optimum contract. That is, even if it changes the difference between the deductible values in the contracts menu, profit will remain unchanged (and equal to zero).

Insurer 2, which holds the second largest market share, operates very close to the optimum contract, but with a small loss per policy. While other insurers do not implement the optimum contract, they can reduce the difference between their deductible and increase their profit (or minimize their loss).

The results of Figure 3 are qualitatively very different from those found by Cohen and Einav (2007). Here, in the face of competition, insurers would see their profit margins greatly reduced and only the leading insurer would be implementing their optimum contract.

8 Concluding Remarks

The present study applied the methodology proposed by Cohen and Einav (2007) in order to estimate the distribution of absolute risk aversion from data on automobile insurance market. Considering multiple insurers, our model captured the effect of competition on such a distribution.

The results suggest that the mean risk aversion is 0.0051. In addition, we identified high heterogeneity in both risk and risk aversion. These estimates are
similar to those obtained by Cohen and Einav (2007), including the sign of the correlation between risk and risk aversion. Regarding covariates, it is perceived that women are more risk averse than men and that the risk aversion coefficient is higher for older insureds. Also, women and younger people are more likely to be in a car collision. Furthermore, we have ascertained a high amount of heterogeneity in the insurer’s client portfolio.

The interpretation of results becomes clearer when a functional form is defined for the utility of individuals. We assume, then, a utility of the quadratic type \( u(w) = w - bw^2 \), since its third derivative assumes zero value. This hypothesis was made in the theoretical model and allowed an analytical relation between risk and risk aversion. The results show that an insured with the average (0.0051) risk aversion coefficient will be indifferent between whether or not to participate in a lottery in which he has a 50% chance of winning R$100.00 and a 50% chance of losing R$64.04.

Finally, we perform counterfactual exercises for the profit of insurers. The results indicate that the insurer with the largest market share is the only one that can implement the optimal contract, that is, it is the only one that maximizes profit. Other insurers could change their contracts, particularly by reducing the value of the regular deductible (ceteris paribus), and would still increase their profitability.
References


Appendix A  Gibbs Sampler

Following Cohen and Einav (2007), this appendix will demonstrate the Gibbs Sampling algorithm (GS), which was used to estimate the model proposed in this study. This method was proposed by Geman and Geman (1984) and became popular among statisticians after 1990, after causing a significant impact on the development and practical applications of Bayesian Statistics.

One of the main advantages of GS is that it enables data augmentation of latent variables. In the present context, this approach allows us to augment risk aversion and risk type for each individual, i.e., \( \{\lambda_i, r_i\}_{i=1}^n \) are treated as additional parameters.

Therefore, the model can be written as follows:

\[
\ln \lambda_i = x_i' \beta + \varepsilon_i, \quad \ln r_i = x_i' \gamma + v_i,
\]

\[
\text{choice}_i = \begin{cases} 
1 & \text{if } r_i > r_i^*(\lambda_i), \\
0 & \text{if } r_i < r_i^*(\lambda_i), 
\end{cases}
\]

\[
\text{claims}_i \sim \text{Poisson}(\lambda_i, t_i),
\]

\[
\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_\lambda & \rho \sigma_\lambda \sigma_r \\ \rho \sigma_\lambda \sigma_r & \sigma^2_r \end{bmatrix} \right),
\]

\[
\delta \equiv \begin{bmatrix} \beta \\ \gamma \end{bmatrix}, \quad \Sigma \equiv \begin{bmatrix} \sigma^2_\lambda & \rho \sigma_\lambda \sigma_r \\ \rho \sigma_\lambda \sigma_r & \sigma^2_r \end{bmatrix}, \quad X \equiv \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.
\]


\[
y \equiv \begin{bmatrix} \lambda \\ r \end{bmatrix}, \quad u_i \equiv \begin{bmatrix} \varepsilon_i \\ v_i \end{bmatrix}.
\]

The set of parameters that we intend to estimate, that is, the posterior distribution, is given by \( \theta = \{ \delta, \Sigma, \{ u_i \}_{i=1}^n \} \). The prior distribution specifies that \( \{ \delta, \Sigma \} \) are independent of \( \{ u_i \}_{i=1}^n \), and \( \{ \delta, \Sigma \} \) has a diffuse conventional prior distribution. We adopted a hierarchical prior distribution for \( \{ u_i \}_{i=1}^n \):

\[
\{ u_i \}_{i=1}^n \bigg| \Sigma \sim \mathcal{N}(0, \Sigma) \\
\Sigma^{-1} \sim \text{Wishart}_2(a, Q).
\]

Then, conditional on all other parameters, we have:

\[
\Sigma^{-1} \bigg| \delta, \{ u_i \}_{i=1}^n \sim \text{Whishart}_2 \left( a + n - k, \left( Q^{-1} + \sum_i u_i u_i' \right)^{-1} \right),
\]

\[
\delta \bigg| \Sigma, \{ u_i \}_{i=1}^n \sim \mathcal{N} \left( (X'X)^{-1}(X'y), \Sigma \otimes (X'X)^{-1} \right).
\]

For \( \Sigma^{-1} \) also a diffuse conventional prior distribution is used such that \( a = 0 \) and \( Q^{-1} = 0 \).

GS is less trivial in cases involving sampling of the conditional distribution of the augmented parameters, \( \{ u_i \}_{i=1}^n \). All individuals are independent of each other, so that, conditional on the other parameters, there is no dependence on increased data of the other individuals. Thus, we need only describe the conditional probability of \( u_i \).

Note that, conditional on \( \delta \), we have: \( \varepsilon_i = \ln \lambda_i - x_i'\beta \) and \( v_i = \ln r_i - x_i'\gamma \). Therefore, one can only focus on the posterior distribution of \( \lambda_i \) and \( r_i \). These posterior distributions are

\[
Pr(r_i \mid \gamma, \beta, \Sigma, \lambda_i, \text{data}) \propto \begin{cases} 
\phi \left[ \ln r_i, x_i'\gamma + \rho \frac{\sigma_r}{\sigma_\lambda} (\ln \lambda_i - x_i'\beta), \sqrt{\sigma_r^2(1 - \rho^2)} \right] & \text{se } \text{choice}_i = I(r_i < r_i^*(\lambda_i)), \\
0 & \text{se } \text{choice}_i \neq I(r_i < r_i^*(\lambda_i)),
\end{cases}
\]

and

\[
Pr(\lambda_i \mid \gamma, \beta, \Sigma, \lambda_i, \text{data})
\]
Estimating Risk and Risk Aversion in the Automobile Insurance Market

\[ p(\lambda_i, \text{claims}, t_i) \phi \left[ \ln \lambda_i, x_i' \beta \right] \left[ + \rho \frac{\sigma_T}{\sigma_r} (\ln r_i - x_i' \gamma) \sqrt{\sigma^2_T (1 - \rho^2)} \right] \quad \text{se } \text{choice}_i = I(r_i < r_i^*(\lambda_i)), \]

\[ 0 \quad \text{se } \text{choice}_i \neq I(r_i < r_i^*(\lambda_i)), \]

where \( p(x, \text{claims}, t) = \lambda^\text{claims} \exp(-xt) \) is proportional to the function probability density of the Poisson distribution, and

\[ \phi(x, \mu, \sigma) = \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\} \]

is proportional to the normal probability density function and \( I(\cdot) \) is an indicator function.

The posterior distribution to \( \ln r_i \) is a truncated normal, for which we use a simple sampling of the “inverse cumulative density function” Devroye (1986). A posterior distribution to \( \ln \lambda_i \) is less trivial. We use a “sliced sampler” to rewrite Damlen et al. (1999):

\[ \Pr(\lambda_i) = b_0(\lambda_i) b_1(\lambda_i) b_2(\lambda_i), \]

where

\[ b_0(\lambda_i) \text{ is a truncated normal,} \]

\[ b_1(\ln \lambda_i) = \lambda_i^\text{claims} = (\exp(\ln \lambda_i))^{\text{claims}_i}, \]

\[ b_2(\ln \lambda_i) = \exp(-\lambda_i t_i) = \exp(-t_i \exp(\ln \lambda_i)). \]

The data augmentation will be performed through two additional variables, \( u^1_i \) and \( u^2_i \), which are uniformly distributed in \([0, b_1(\lambda_i)]\) and \([0, b_2(\lambda_i)]\), respectively. And the probability will be

\[ \Pr(\lambda_i, u^1_i, u^2_i) = b_0(\lambda_i) b_1(\lambda_i) b_2(\lambda_i) \]

\[ \left[ I \left( 0 \leq u^1_i \leq b_1(\lambda_i)/b_1(\lambda_i) \right) \right] \left[ I \left( 0 \leq u^2_i \leq b_2(\lambda_i)/b_2(\lambda_i) \right) \right] \]

\[ = b_0(\lambda_i) I \left( 0 \leq u^1_i \leq b_1(\lambda_i) \right) I \left( 0 \leq u^2_i \leq b_2(\lambda_i) \right), \]

whereas \( b_1(\cdot) \) and \( b_2(\cdot) \) are conditional monotonic functions in \( u^1_i \) and \( u^2_i \), so that \( b_1^{-1}(u^1_i) = (\ln u^1_i)/\text{claims}_i \) is the lower bound of \( \ln \lambda_i \) (for \( \text{claims}_i > 0 \)) and \( b_2^{-1}(u^2_i) = \ln(-\ln u^2_i) - \ln t_i \) is the upper limit of \( \ln \lambda_i \). Then, the sampling of \( \lambda_i \) from a truncated normal occurs after the modification of the limits according to \( u^1_i \) and \( u^2_i \).
Appendix B  The premium-deductible pair variation

Table 7. The premium-deductible pair for each insurer company.

<table>
<thead>
<tr>
<th>Insurers</th>
<th>$p^l$</th>
<th>$p^h$</th>
<th>$\Delta p$</th>
<th>$d^l$</th>
<th>$d^h$</th>
<th>$\Delta d$</th>
<th>Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1353.76</td>
<td>1313.43</td>
<td>40.32</td>
<td>994.22</td>
<td>1988.44</td>
<td>994.22</td>
<td>60115</td>
</tr>
<tr>
<td>2</td>
<td>1042.38</td>
<td>1010.26</td>
<td>32.12</td>
<td>1090.32</td>
<td>2180.65</td>
<td>1090.32</td>
<td>20136</td>
</tr>
<tr>
<td>3</td>
<td>1263</td>
<td>1027.49</td>
<td>235.50</td>
<td>896.13</td>
<td>1792.27</td>
<td>896.13</td>
<td>7608</td>
</tr>
<tr>
<td>4</td>
<td>1513.34</td>
<td>1290.69</td>
<td>222.64</td>
<td>1099.37</td>
<td>2198.74</td>
<td>1099.37</td>
<td>1318</td>
</tr>
<tr>
<td>5</td>
<td>1395.56</td>
<td>1288.01</td>
<td>107.54</td>
<td>958.04</td>
<td>1916.08</td>
<td>958.04</td>
<td>230</td>
</tr>
</tbody>
</table>