Note: This is a translated version of the original thesis in Portuguese (ABNT format) that was submitted to the examination board

Enrico Peceguini Ruggieri

Asymmetric Information in the Brazilian Supplementary Health Market: An Econometric Approach

São Paulo, SP, Brazil

June 26, 2024

Note: This is a translated version of the original thesis in Portuguese (ABNT format) that was submitted to the examination board

Enrico Peceguini Ruggieri

Asymmetric Information in the Brazilian Supplementary Health Market: An Econometric Approach

Monograph presented to the Department of Economics of FEA-USP for completion of the Bachelor's Degree in Economics.

University of São Paulo - USP

School of Economics, Business Administration, Accounting and Actuarial Science

Bachelor's Degree in Economics

Advisor: Professor Denise Cavallini Cyrillo, PhD

São Paulo, SP, Brazil

June 26, 2024

Note: This is a translated version of the original thesis in Portuguese (ABNT format) that was submitted to the examination board

Enrico Peceguini Ruggieri

Asymmetric Information in the Brazilian Supplementary Health Market: An Econometric Approach/ Enrico Peceguini Ruggieri. – São Paulo, SP, Brazil, June 26, 2024

Note: This is a translated version of the original thesis in Portuguese (ABNT format) that was submitted to the examination board-

41 p. : il. (some color.) ; 30 cm.

Advisor: Professor Denise Cavallini Cyrillo, PhD

Monograph – University of São Paulo - USP
School of Economics, Business Administration, Accounting and Actuarial Science
Bachelor's Degree in Economics, June 26, 2024
Note: This is a translated version of the original thesis in Portuguese (ABNT format) that was submitted to the examination board.

1. Risk and Uncertainty. 2. Health Economics

JEL: D82, G22, I11

Acknowledgements

To USP, for being such a significant space for me, present in so many moments of my personal and academic life.

To my advisor Denise, for the attentive guidance, for the patience with my comings and goings, and for the receptiveness to my ideas, even when they didn't follow the most orthodox paths.

To my college colleagues, for the support in difficult moments and for the sincere friendships that I didn't expect to find when entering college at a more mature stage of my life.

To my sister Mari, who always makes me think about how it's possible to love so much someone so different from yourself.

To Veri, who knows all my flaws and responds to them only with affection, for the enormous support throughout these four years.

To my mother, for teaching me to always seek more and to demand the best of myself.

To my father, whose intelligence has always inspired me to study and to engage in true curiosity for knowledge.

My sincere thank you to all who walked by my side so that I could get here.

“Viver é muito perigoso,
Porque aprender a viver é que o viver mesmo.
Travessia perigosa, mas é a da vida.
Sertão que se alteia e se abaixa.
O mais difícil não é ser bom e proceder honesto;
Dificultoso, mesmo, é saber definido o que quer,
e ter o poder de ir até o rabo da palavra.”
(João Guimarães Rosa)

Abstract

A partir da motivação do fenômeno da Seleção Adversa, este trabalho explora os desafios do mercado de saúde suplementar brasileiro e sua susceptibilidade ao problema da Informação Assimétrica. Por meio da aplicação do teste de correlação positiva (CHIAPPORI; SALANIE, 2000) aos dados da Pesquisa Nacional de Saúde de 2019, o trabalho busca contribuir com a bibliografia que estuda a presença dessa falha informacional.

Como contribuição original, propõe-se também uma atenção especial às extensões do teste, realizando uma revisão metodológica dessa abordagem econométrica para verificação da Informação Assimétrica. Em especial, são exploradas as extensões de variáveis não-utilizadas de Finkelstein and Poterba (2014) e o modelo multinomial de Kim et al. (2009).

Encontram-se evidências estatisticamente significativas da presença de informação assimétrica nesse mercado. Também são identificadas variáveis, com significância estatística, que contribuem para esse problema informacional por não serem utilizadas pelas companhias de seguro na diferenciação dos prêmios. Além disso, uma Regressão Binomial Negativa revela resultados semelhantes, utilizando um modelo mais complexo que permite a diferenciação entre tipos de cobertura.

Palavras-chaves: Economia da Saúde. Informação Assimétrica. Risco e Seguro.

Códigos JEL: D82, G22, I11.

Abstract

Motivated by the phenomenon of Adverse Selection, this work explores the challenges of the Brazilian supplementary health market and its susceptibility to the Asymmetric Information problem. By applying the positive correlation test (CHIAPPORI; SALANIE, 2000) to the 2019 National Health Survey data, the study aims to contribute to the literature examining the presence of this informational failure.

As an original contribution, special attention is also proposed for extensions of the test, conducting a methodological review of this econometric approach for assessing Asymmetric Information. In particular, the unused variable extensions by Finkelstein and Poterba (2014) and the multinomial model by Kim et al. (2009) are explored.

Statistically significant evidence of asymmetric information presence in this market is found. Additionally, variables are identified, with statistical significance, that contribute to this informational problem by not being utilized by insurance companies in premium differentiation. Furthermore, a Negative Binomial Regression reveals similar results, using a more complex model that allows differentiation between coverage types.

Keywords: Health Economics, Asymmetric Information, Risk and Insurance.

JEL codes: D82, G22, I11.

List of Figures

Figure 1 – Brazilian Consumer Price Index: General index and sub-indices related to Health Plans 11
Figure 2 – Adverse Selection Effect Hypothesis in the Health Insurance Market . . 14
Figure 3 – Average Claims by Age Group, with Additional Brackets 30

List of Tables

Table 1 – Probit Regression Results of Coverage for the canonical model 27
Table 2 – Probit Regression Results of the Risk of Loss for the canonical model . . 27
Table 3 – Probit Regression Results of Coverage - smoker variable 29
Table 4 – Probit Regression Results of Risk of Loss - smoker variable 29
Table 5 – Probit Regression Results for Coverage - diabetes and additional age groups 31
Table 6 – Probit Regression Results for Risk of Loss - diabetes and additional age groups 31
Table 7 – Multinomial Model - Ordered Probit Regression of Coverage 33
Table 8 – Multinomial Model - Negative Binomial Regression of Risk of Loss using calculated residuals as regressors 34

List of abbreviations and acronyms

ANS	Agência Nacional de Saúde (National Health Agency)
CPI	Comissão Parlamentar de Inquérito (Parliamentary Commission of Inquiry)
IBGE	Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics)
IPCA	Índice de Preços ao Consumidor Amplo (Broad Consumer Price Index)
MP	Ministério Público (Public Prosecutor's Office)
PNAD	Pesquisa Nacional de Amostra Domiciliar (National Household Sample Survey)
PNS	Pesquisa Nacional de Saúde (National Health Survey)
QMLE	Quasi-Maximum Likelihood Estimator
SIDRA	Sistema IBGE de Recuperação Automática (IBGE Automatic Data Retrieval System)

Contents

Resumo 5

Abstract 6

1 INTRODUCTION 11
1.1 The Brazilian Supplementary Health Market: Context and Perspectives 11
1.2 The Problem of Asymmetric Information in Insurance Markets . . . 13
1.2.1 Adverse Selection and Moral Hazard 13
1.2.2 Risk Profile and Risk Aversion 15

2 LITERATURE REVIEW 16

3 MATERIAL AND METHOD 19
3.1 Positive Correlation Test: The Canonical Model 20
3.2 Extensions of the Positive Correlation Test 21
3.2.1 Heterogeneity of Individual Preferences and Unused Variables 21
3.2.2 Multinomial Measures of Coverage and Risk of Loss 23

4 RESULTS 26
4.1 PNS 2019: An Overview 26
4.2 Positive Correlation Test 26
4.2.1 The Canonical Model 26
4.2.2 Preference Heterogeneity and Unused Variables 28
4.2.3 Testing Other Unused Variables 30
4.2.4 A multinomial model proposal for Coverage 32
4.3 Moral Hazard vs. Adverse Selection: a brief discussion 34

Conclusion 36

REFERENCES 38

APPENDIX A – R AND PYTHON CODE 41

1 Introduction

1.1 The Brazilian Supplementary Health Market: Context and Perspectives

The supplementary health sector plays a crucial role in Brazil's healthcare system, offering private insurance alternatives to public services. It currently covers 50.5 million beneficiaries, representing nearly a quarter of the Brazilian population. The sector, however, faces serious setbacks, where a high degree of consumer dissatisfaction coexists with financial difficulties of the insurance companies. At the same time that plan readjustments, both for individual and group policies, consistently exceed the average inflation measured by the Brazilian CPI (OCKé-REIS; FIUZA; COIMBRA, 2019), companies report consecutive losses in their financial results.

The sector is regulated by the National Supplementary Health Agency (ANS), established by Law 9656/98, which defines and oversees readjustments, as well as the list of health procedures to be covered by operators. Regarding the possibility of premium differentiation by insurers, the case of Brazilian supplementary health stands out for its notably restrictive legislation in this aspect. ANS Normative Summary 27/2015 states that

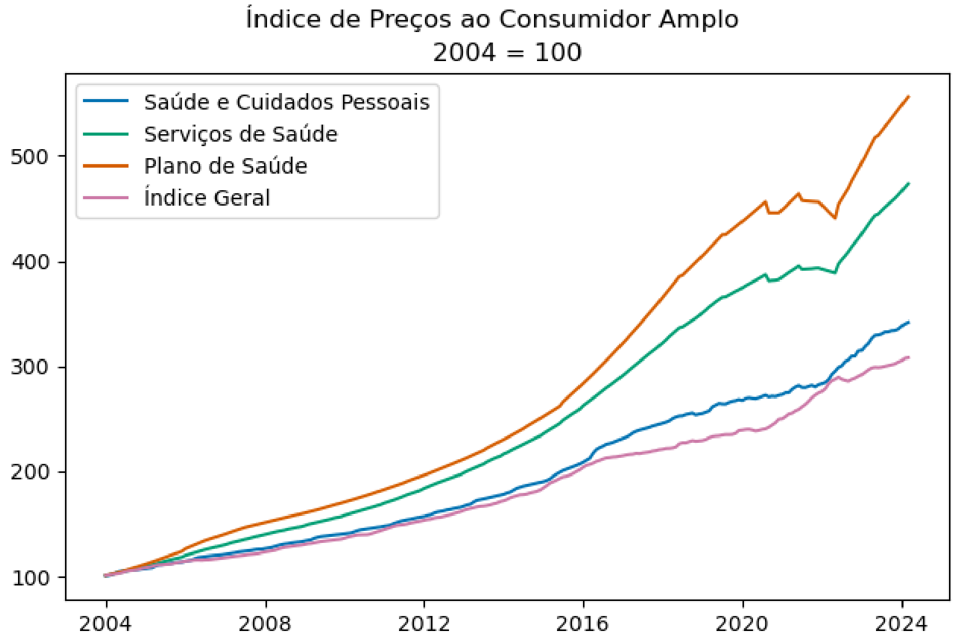

Figure 1 – Brazilian Consumer Price Index: General index and sub-indices related to Health Plans

Own elaboration based on data from SIDRA IBGE "the practice of risk selection by health plan companies in contracting any type of private health care plan is disallowed" (ANS, 2015). Currently, only premium differentiation in 10 age groups is allowed, with the Elderly Statute preventing any distinction after 59 years of age. Judicialization also imposes additional difficulties to providers, who find themselves obliged to fund very expensive medications, even when these are not included in the ANS list (Conselho Nacional de Justiça, 2024).

At the same time, the sector faces widespread consumer dissatisfaction. The number of complaints registered with the regulatory agency is growing rapidly (PIOVEZAN, 2023), with grievances about denied treatments and worsening coverage being the most frequent ones. Moreover, the cost paid for these services demands an increasingly larger portion of families' budgets. As illustrated in Figure 1, the sub-indices for health plans and services have consistently outpaced the general inflation index, highlighting the sector's rising relative costs.

A recent controversy in Brazil's supplementary health sector has drawn widespread public attention: the unilateral cancellation of health plans by insurers. This issue has disproportionately impacted vulnerable groups, including the elderly, individuals with pre-existing conditions, and those undergoing treatment for severe illnesses (MUGNATTO, 2024). In response, Legislative Assemblies and Public Prosecutors' Offices across various states have launched investigations, while the National Supplementary Health Agency (ANS) has stepped up its efforts to fulfill its regulatory role. These cancellations pose a significant threat, leaving already fragile individuals at risk during their most vulnerable moments. This situation undermines the fundamental purpose of financial protection against unexpected health crises, a key principle of health systems as established by the World Health Organization (WHO) (WORLD HEALTH ORGANIZATION, 2014).

It is from this difficult conjunction of factors that the motivation for this work emerges. As the increasing costs of insurance policies are passed on to consumers, the incentive increases for healthy individuals to leave the market and for individuals with higher risk to choose to contract insurance, increasingly raising the claims ratio of the plans. The existence of Asymmetric Information in this market, therefore, becomes quite a reasonable hypothesis.

A methodology frequently employed by the literature for evaluating Asymmetric Information in the supplementary health market is the positive correlation test, initially proposed by Chiappori and Salanie (2000). The original contribution of this work consists in applying this methodology to the most recent National Health Survey (2019). Furthermore, we propose a special focus on the extensions to the econometric test, so far little explored by the literature.

1.2 The Problem of Asymmetric Information in Insurance Markets

Given the rising claims ratio, the substantial increase in the real cost of insurance policies, and demographic factors tied to Brazil's aging population, the key question that emerges is whether asymmetric information exists in the supplementary health market.

This phenomenon is quite frequent in insurance markets and is theoretically divided between two market failures: Adverse Selection and Moral Hazard. While the problem of adverse selection is seen as an *ex-ante* informational problem, moral hazard is the *ex-post* informational problem. In the case of the insurance market, Adverse Selection would motivate the individual with the higher risk of loss to adhere more to coverage (or to adhere to a more comprehensive policy). Moral Hazard, in turn, would be the individual adopting a riskier posture because they are covered by insurance. The effect of Asymmetric Information can be extremely deleterious for insurance markets, leading to a disadvantageous selection of individuals, increasingly raising premiums and claims ratios, and potentially leading to a collapse of this market.

1.2.1 Adverse Selection and Moral Hazard

The theme of adverse selection has been the subject of intense study in recent decades. This phenomenon was initially described in the famous article *The Market for Lemons* (AKERLOF, 1970), and occurs when individuals use private information *ex-ante* to choose whether or not to participate in a market. A few years later, Rotschild and Stiglitz (1976) applied this concept to the insurance market, when individuals have private information about their risk that influences the choice of whether or not to contract insurance.

The concept of Adverse Selection in the supplementary health market is quite intuitive to understand. Let's assume a set of individuals where each has a given willingness to pay for a health insurance policy. Generally, an individual with poorer health will derive greater expected utility from the health plan, increasing their willingness to pay for coverage. Thus, the willingness to pay is also greater according to the individual's health status *ex-ante* to the decision about whether or not to join the plan.

If this health status is unknown to the insurer (or if there is some other impediment to premium differentiation), it is unequivocal to note how this phenomenon can lead to the disadvantageous selection. If there is an increase in the price of the policy, individuals with a lower willingness to pay will have a greater incentive to leave the plan. That is, there will be exclusion of individuals with better health status, which will tend to increase the claims ratio of the plan, consequently leading to an increase in the premium. Thus, a "snowball effect" is established which, in the limit, causes an increasingly lower adherence of healthy individuals and an increasingly higher adherence of individuals with increasingly

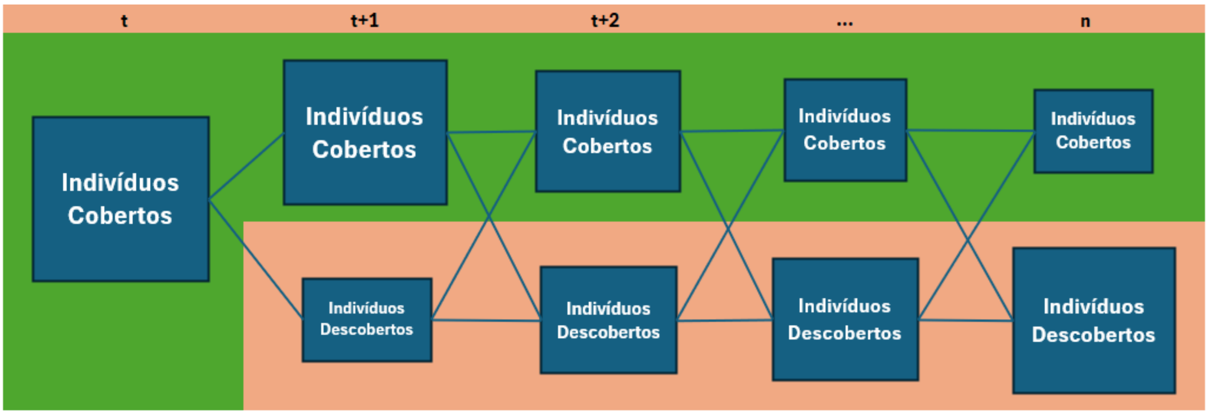

Figure 2 – Adverse Selection Effect Hypothesis in the Health Insurance Market

Source: Own elaboration based on Winssen, Kleef and Ven (2018)

worse health status.

To illustrate this problem, a noteworthy study is the article that, in fact, initially motivated this research. Using a very detailed database regarding the supplementary health market in the Netherlands, Winssen, Kleef and Ven (2018) performed an iterative simulation, repeated for 25 periods, analyzing the entry and exit of individuals and re-estimating the claims ratio and premium paid in each period. Figure 2 represents the effect of Adverse Selection in a supplementary health market as modeled by the authors. The conclusion of the article was that further premium differentiation could reduce the problem of adverse selection and avoid the collapse of this market.

Moral Hazard, on the other hand, as discussed earlier, is the *ex-post* informational problem. In this phenomenon, the economic agent adopts riskier behavior (or increases the claims ratio) after the decision to join the insurance.

In the case of health insurance, this concept may seem, at first, unreasonable. It would be illogical to assume, for example, that an individual would become sick more often because they are covered by a policy. For elective procedures, however, the hypothesis is more feasible, as the economic agent may undergo unnecessary procedures because they are covered by a plan, a possibility that motivates the existence of co-participation in various policies.

Some authors criticize this idea, however, arguing that the supposed "waste" attributed to moral hazard is, in fact, a reflection of the under-utilization of health services by individuals without coverage, unable to afford the high costs of this care (ROBERTSON et al., 2020).

Given the limited clarity surrounding Moral Hazard in supplementary health, this work is motivated by a specific focus on Adverse Selection in this market. It is important to emphasize, however, that the positive correlation test and its extensions are not capable of differentiating these two phenomena of Asymmetric Information. The methodology is robust in identifying whether there is private information from consumers about their risk, but not about the specific nature of this information.

> Both adverse selection and moral hazard can generate a positive correlation between insurance coverage and claims, but these are two very different forms of asymmetric information with very different implications for public policy. With adverse selection, individuals who have private information that they are at higher risk self-select into the insurance market, generating the positive correlation between insurance coverage and observed claims. As already discussed, the government has several potential welfare-improving policy tools to possibly address such selection. With moral hazard, individuals are identical before they purchase insurance, but have incentives to behave differently after. Those with greater coverage have less incentive to take actions that reduce their expected costs, which will generate a relationship between insurance coverage and observed claims. (EINAV; FINKELSTEIN, 2011)

Despite this caveat, in the results section of this work, a brief discussion will be presented, also based on the article by Finkelstein and Poterba (2014), which proposes a differentiation of an eminently qualitative nature between the two effects of the Asymmetric Information phenomena.

In order to be faithful to the methodological rigor of the test employed, and recognizing the limitation of this approach to differentiate these two phenomena, the term *Asymmetric Information* will be predominantly used throughout this work.

1.2.2 Risk Profile and Risk Aversion

A final key distinction for this introduction is to differentiate between private information from varying risk profiles and private information from the heterogeneity of individual preferences. *Ceteris paribus*, individuals with higher risk aversion are more likely to seek insurance policies, or opt for more comprehensive coverage.

The problem that emerges, therefore, is whether the positive correlation identified will come from this heterogeneity in individuals' risk aversion or from different risk profiles. The differentiation is important because it can invalidate the result of the econometric test, and will be discussed further in the methodology section of this work.

2 Literature Review

A few years after the publication of the influential article “The Market for Lemons” by Akerlof (1970), the phenomenon of Asymmetric Information was explored within the insurance market, with a notable contribution from Rotschild and Stiglitz (1976). This seminal paper gave rise to the Rothschild-Stiglitz model, which describes the occurrence of adverse selection in this domain and would later justify awarding the Nobel Prize to the second author. Since then, asymmetric information in insurance markets has been extensively researched. Arrow (1963), in turn, made a significant contribution with his work on Moral Hazard in insurance markets.

In the early years following the publication of these foundational articles, the debate primarily revolved around theoretical aspects of the asymmetric information phenomenon in insurance markets. However, in recent years, the focus has shifted toward empirical investigation. Researchers now leverage extensive datasets and apply econometric and quantitative methods to empirically examine whether asymmetric information is present in these markets.

Since the publication of Chiappori and Salanie (2000), the positive correlation test has become a canonical model for detecting asymmetric information in insurance markets. This test estimates two reduced-form equations: one for insurance coverage and another for the risk of loss. Derived from the Rothschild-Stiglitz model, the test is based on the hypothesis that agents more susceptible to the risk of loss will opt for contracts with more comprehensive coverage. Specifically, the test involves modeling insurance coverage as a function of the individual’s observable exogenous variables and, similarly, modeling the risk of loss using these same variables. A positive correlation in the residuals of these regressions provides evidence for the presence of asymmetric information. A formal presentation of this methodology will be provided in the following chapter of this monograph.

Although Chiappori and Salanie (2000) originally applied this methodology to study the automotive insurance market in France, it has since been widely used across various markets. In a comprehensive methodological review, Cohen and Siegelmann (2010) examined studies testing for asymmetric information across different insurance types and found that the method developed by Chiappori and Salanié remains one of the most commonly employed approaches for this purpose.

More specifically in the health market, some articles also stand out for their relevance in investigating this phenomenon. Cutler and Zeckhauser (2000) conduct an extensive literature review and discover that the vast majority of articles published until then found strong evidence of the existence of asymmetric information in the health insurance market. Among the thirty articles that the two authors aggregate, published over 26 years, twenty-five find the occurrence of asymmetric information and three point to an ambiguous effect. The methodology for evaluating this asymmetric information phenomenon in the health market is quite varied, but the results, in general, predominantly point to the occurrence of the informational problem. An important exception is the widely cited work of Cardon and Hendel (2001) which, using a two-stage structural model, finds no evidence of asymmetric information, with most of the variance in the risk of loss being explained by observable variables.

Several studies over the past two decades have applied the positive correlation test to Brazil's supplementary health market. Alves (2004) and Resende and Zeidan (2010) tested the asymmetric information hypothesis in the Brazilian supplementary health market using data from the PNAD Health 2003 (current PNS). The first work found evidence of asymmetric information, while the second obtained different results. These two authors, however, seem to have incurred some deviations from the canonical model by using a large number of exogenous variables not used by the insurance company in the matrix of observable variables. As discussed previously, Brazilian legislation imposes a series of restrictions on the possibilities of premium differentiation that insurers can perform. The inclusion, therefore, of a large number of exogenous variables in the matrix seems to diverge from what the bibliography on this subject suggests. In a subsequent effort, Fonseca (2017) repeated the procedure of these authors, using data from the PNS 2013. By employing a more robust methodology, closer to the canonical model, he also found evidence of asymmetric information.

Other methods have also been applied to investigate informational issues within Brazil's supplementary health system. For instance, Sá (2012) provides a literature review covering various studies that utilize data from the 1998 and 2003 PNAD Health surveys, encompassing not only the method proposed by Chiappori and Salanié but also other econometric approaches, such as the DiD estimator used by Nishijima, Postali and Fava (2011). Among the articles reviewed by Sá (2012), five examine the presence of moral hazard and all find evidence supporting this market failure. Regarding adverse selection, seven articles attempt to detect this issue, though only two report significant evidence.

More recently, some efforts have been made in an attempt to correct some of the limitations of the positive correlation test. Cutler, Finkelstein and McGarry (2008) suggest the inclusion of variables that capture individual characteristics related to risk aversion. The idea of this methodology would be to control the model for factors that may generate some type of endogeneity in the models due to the heterogeneity of individual preferences.

Finkelstein and Poterba (2014), in turn, propose the addition of another vector W_i incorporating the observable variables not used by the insurance company. This test of "unused observables" would seek to address the problem of variables that can generate asymmetric information both from private information of different profiles and from problems of heterogeneity in individual parameters that can affect the demand for insurance. When applying this model to the annuity market in the United Kingdom, the authors identified that the disregard of the place of residence (the unused observable) generates the same type of inefficiency that arises when contractors hold private information about mortality risk. The place of residence would later be included as a determinant variable of the premium by companies in the UK. A brief application of this extension to the Brazilian case, using the PNS of 2013, is suggested by Fonseca (2017), but the author uses this method only as a robustness test in the final section of his results.

Also within the scope of extensions of the positive correlation test, the importance of the article by Kim et al. (2009) stands out, which justifiably problematizes the dichotomous characteristic of positive correlation tests. In practical terms, using probit models implies that the dependent variable is constrained by binary outcomes for both insurance coverage and the risk of loss. To deal with this limitation, the authors use an ordered probit on the coverage variable, in order to separate policies with greater coverage from more limited policies. By applying this methodology in the automobile insurance market, the authors find strong evidence of the presence of asymmetric information.

With the same motivation, Dardanoni, Forcina and Donni (2018) proposed a multivariate model to deal with the problem of the multidimensional nature of private information. This extension of the positive correlation test is based on a flexible class of regressions that analyzes the association between coverage and more than one variable referring to the risk of loss. The authors apply this model to a database of the Health and Retirement Service that studies Medigap plans (a type of supplement to Medicare in the USA), and find evidence that there is a positive and statistically significant correlation between risk and coverage, thus rejecting the null hypothesis of symmetric information.

3 Material and Method

To investigate asymmetric information in the Brazilian supplementary health market, the main database used was the National Health Survey, conducted by IBGE in 2019. This dataset includes microdata on 293,726 individuals across 108,475 households, providing access to information on respondents' health status, participation in the supplementary health system, income, and healthcare usage.

In this study, the positive correlation test was applied to verify the presence of asymmetric information, with a focus on two key extensions developed later. The first extension addresses individual preference heterogeneity by introducing a variable related to behavior, allowing the model to control for risk aversion. The second extension investigates whether certain variables, which insurers do not or cannot use to differentiate premiums, contribute to informational asymmetry.

Finally, a multinomial model was also employed, enabling analysis of different types of coverage as an ordered outcome, and incorporating a discrete quantitative variable as the dependent variable in the Risk of Loss equation.

3.1 Positive Correlation Test: The Canonical Model

The methodology for investigating asymmetric information employs the positive correlation test, as briefly introduced in the literature review. This econometric test consists of two equations, with X_i representing a matrix of each individual i's exogenous variables used by insurers to determine premiums.

The first equation models insurance coverage as a function of the variables in X_i, while the second equation relates the risk of loss to these same variables. This approach tests the hypothesis, derived from the Rothschild-Stiglitz model, that individuals at greater risk of loss are more likely to obtain coverage or choose more comprehensive contracts. If there is correlation in the estimated equations distributions, it suggests that unobservable variables (contained in ϵ and η) are influencing both the choice of coverage and the likelihood of a claim. If all private information were perfectly captured within X_i, these distributions, and thus the residuals, would be uncorrelated.

Let Ω and Π be vectors of parameters, we can write:

$$Coverage_i = X_i\Omega + \epsilon_i \tag{3.1}$$

$$Loss_i = X_i\Pi + \eta_i \tag{3.2}$$

In the case of the PNS, the two models will be a regression of the exogenous variables used (age groups) on the response variable if those individuals have insurance or not, and a regression of the same exogenous variables on the response variable of how much these individuals used the plan.

These models are estimated by means of two independent probits and, if rejected $H_0 : Cov(\hat{\epsilon}, \hat{\eta}) = 0$, the null hypothesis of symmetric information is rejected. The authors suggest, for verification of this hypothesis, the statistic W, where w_i is the number of days that individual i would be covered by the insurance:

$$W = \frac{\left(\sum_{i=1}^{n} w_i \hat{\epsilon_i} \hat{\eta_i}\right)^2}{\sum_{i=1}^{n} w_i^2 \hat{\epsilon_i}^2 \hat{\eta_i}^2} \tag{3.3}$$

In this case, given the cross-sectional nature of the data and the periodicity of the PNS, we can treat w_i as constant. This allows us to factor it out of both the numerator and denominator of the summation, thereby excluding it from the calculation of the statistic. Under the null hypothesis of no correlation among the residuals, W follows an asymptotic chi-square distribution with 1 degree of freedom. In practical terms, for a significance level of 5%, the W_{critical} will be equal to 3.84. In addition to the W statistic originally used by the authors, the existence of correlation between the vectors of the estimated residuals can be supported by other correlation coefficients, as well as by the statistical significance of these coefficients, such as the Pearson, Spearman, and Kendall Tau coefficients (HEUVEL; ZHAN, 2022).

As previously mentioned, Alves (2004) and Resende and Zeidan (2010), who applied this test to the Brazilian context, deviated from the canonical model by including a series of exogenous variables in X_i that, legally, cannot be used by insurance companies to determine premiums. This practice diverges from the standard approach suggested in the literature and is not consistent with the theoretical framework of the model. Therefore, in this study, we restrict the exogenous variable matrix X_i to only include categorical variables representing the 10 age groups that Brazilian legislation permits for premium differentiation.

3.2 Extensions of the Positive Correlation Test

3.2.1 Heterogeneity of Individual Preferences and Unused Variables

Building on the works of the aforementioned authors, this project proposes an extended focus on the positive correlation test, particularly addressing the heterogeneity of individual preferences and the inclusion of variables not utilized by the insurance company in the regression models.

A potential limitation of the canonical model is its inability to differentiate between private information related to different risk profiles and individual differences in risk aversion. Formally, when residuals from the canonical model are denoted by ϵ and η, and both issues—variation in risk profiles (captured by X_1) and differences in risk aversion (X_2)—are present, we can define them as follows:

$$\epsilon_i = X_{1,i}\omega_1 + X_{2,i}\omega_2 + \epsilon'_i \tag{3.4}$$

$$\eta_i = X_{1,i}\pi_1 + X_{2,i}\pi_2 + \eta'_i \tag{3.5}$$

Thus, if there is some type of heterogeneity in preferences not taken into account by the model, it is possible that the result of the positive correlation test leads to an erroneous conclusion. Finkelstein and Poterba (2014), for example, point out the possibility of the occurrence of Type II Error (does not reject H_0 of symmetric information, but H_0 is false) in a situation where risk aversion is positively correlated with *Coverage*, but negatively correlated with *Loss*.

The first extension that seeks to address this problem is the one proposed by Cutler, Finkelstein and McGarry (2008), and consists of including variables that capture individual characteristics related to risk aversion. The idea of this methodology is to control the model for factors that may generate some type of endogeneity in the models due to differences in individual preferences. It would be reasonable to assume, for example, that a individual who smokes would have a lower aversion to health risk. Thus, the utility gain of this individual in contracting insurance would be lower than that of a non-smoker individual. The variable *smoker*, in this case, could be modeled as a behavior. Using the authors' indicator function notation, we can define:

$$\mathbb{1}Coverage_i = \beta_0 + \beta_1 \cdot Behavior_i + X_i\Omega + \epsilon_i \tag{3.6}$$

$$Loss_i = \alpha_0 + \alpha_1 \cdot Behavior_i + X_i\Pi + \eta_i \tag{3.7}$$

The second extension of the positive correlation test is the one proposed in Finkelstein and Poterba (2014), being the test of "unused observables" mentioned in the literature review. A matrix W_i of exogenous variables not used by insurers is added to the canonical model.

In practice, we can observe that this model is quite similar to the first extension:

$$Coverage_i = X_i\Omega + \alpha W_i + \epsilon_i \tag{3.8}$$

$$Loss_i = X_i\Pi + \delta W_i + \eta_i \tag{3.9}$$

This model also addresses the issue of preference heterogeneity but introduces some distinctions from the previous approach. Firstly, it is more general, as the variable W can influence both X_1 (risk profile) and X_2 (risk aversion).

Additionally, in this case, the W statistic is not employed. The authors propose a test based on examining the coefficients of variables excluded from the regressions. Rejecting the hypothesis that $\alpha = 0$ and $\delta = 0$ is equivalent to rejecting the null hypothesis of symmetric information. If both α and δ are found to be significantly different from zero, this indicates a correlation between a variable unused by the insurer and both the risk of loss and insurance coverage.

Moreover, this analysis enables the exploration of various candidates for W_i, which can help identify which variables, neglected by insurers, may cause asymmetric information. In the Brazilian context, this is particularly interesting given the legal constraints on premium differentiation and the detailed nature of the available databases, which offer several candidate variables. A brief application of this extension to the Brazilian market, using data from the 2013 PNS, is presented by Fonseca (2017), but the author only applies this method as a robustness check in the final section of his results.

The proposed econometric methodology aims to contribute to the broader discussion by assessing the presence (or absence) of asymmetric information in the Brazilian supplementary health insurance market. It is crucial to highlight, however, that the positive correlation test has a limitation: it does not directly differentiate between adverse selection and moral hazard. Therefore, distinguishing between the *ex-ante* and *ex-post* informational problems must be approached either through a qualitative analysis or by incorporating additional empirical evidence, as suggested by Finkelstein and Poterba (2014).

Furthermore, the method has limitations when applied to situations where the dependent variable is ordinal, as noted by both Kim et al. (2009) and Dardanoni, Forcina and Donni (2018). In the following section, we will focus on the approach proposed by the former authors to address this limitation.

3.2.2 Multinomial Measures of Coverage and Risk of Loss

As discussed in the canonical model, the Coverage equation is defined by the function:

$$Coverage_i = X_i\Omega + \epsilon_i \tag{3.10}$$

The positive correlation test discussed thus far has a significant limitation: estimating Coverage and Risk of Loss using probit models requires the dependent variables to be binary. However, Kim et al. (2009) propose a solution by combining an ordered probit for the Coverage equation with a negative binomial regression for the Risk of Loss equation, which allows for the use of an ordinal variable in the first case and a discrete quantitative variable in the second.

Thus, to begin the model estimation, we first define ${Coverage_i}^m$ as:

$${Coverage_i}^m = \begin{cases} 0 & \text{if } {Coverage_i}^{m*} < \mu_1 \\ 1 & \text{if } \mu_1 \leq {Coverage_i}^{m*} < \mu_2 \\ 2 & \text{if } \mu_2 \leq {Coverage_i}^{m*} \end{cases}$$

where ${Coverage_i}^{m*}$ is a latent variable and μ_1 and μ_2 are thresholds for the three categories defined above.

Using a multinomial ordered dependent variable of ${Coverage_i}^m \in \{0, 1, 2\}$, the estimation of residuals becomes more complex than in the canonical model.

First, the authors propose grouping the three choices of $Coverage_i{}^m \in \{0,1,2\}$ into two sets of ordered choices. We therefore define $Coverage_i{}^1 = 0$ if $Coverage_i{}^m = 0$, and $Coverage_i{}^1 = 1$ if $Coverage_i{}^m \in \{1,2\}$. Similarly, let $Coverage_i{}^2 = 0$ if $Coverage_i{}^m \in \{0,1\}$ and $Coverage_i{}^1 = 1$ if $Coverage_i{}^m = 2$.

From there, it becomes more intuitive to define the predicted probabilities for each grouping. Based on the work of Kim et al. (2009), once again, we define the generalized residuals as:

$$\hat{\epsilon}_i{}^{m_1} = \frac{\phi(X_i\Omega - \hat{\mu_1})}{\Phi(X_i\Omega - \hat{\mu_1})(1 - \Phi(X_i\Omega - \hat{\mu_1}))}[Coverage_i{}^1 - \Phi(X_i\Omega)] \tag{3.11}$$

$$\hat{\epsilon}_i{}^{m_2} = \frac{\phi(X_i\Omega - \hat{\mu_2})}{\Phi(X_i\Omega - \hat{\mu_2})(1 - \Phi(X_i\Omega - \hat{\mu_2}))}[Coverage_i{}^2 - \Phi(X_i\Omega)] \tag{3.12}$$

In practical terms, $\hat{\mu_1}$ and $\hat{\mu_2}$ are estimated thresholds for the categories observed by the ordered probit model. In the generalized residuals equations, $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution functions of the standard normal distribution, respectively, which allows them to be easily calculated computationally.

In this extension of the model, a definition of the Risk of Loss is proposed as a function such that:

$$Loss_i = g(X_i, \hat{\epsilon_i}) + \eta_i \tag{3.13}$$

where $Loss_i$ is a discrete quantitative variable, and no longer a binary variable.

In this case, the two residuals obtained (3.11 and 3.12) are included in the Loss equation as regressors, with the estimated regression coefficient of $\hat{\epsilon}_i{}^{m_1}$ capturing the effect of information asymmetry in the choice between no coverage ($Coverage_i{}^m = 0$) and intermediate coverage ($Coverage_i{}^m = 1$), and that of $\hat{\epsilon}_i{}^{m_2}$ capturing the effect of information asymmetry in the choice between intermediate coverage ($Coverage_i{}^m = 1$) and full coverage ($Coverage_i{}^m = 2$).

There are several possible ways to estimate this equation. The most elementary way to model a count variable would be using a Poisson Regression. In this model, for the Risk of Loss equation, from Wooldridge (2003), we can define the log-likelihood function as:

$$\mathcal{L}(\Pi) = \sum_{i=1}^{n} \ell_i(\Pi) = \sum_{i=1}^{n}(L_i x_i \Pi - exp(x_i\Pi)) \tag{3.14}$$

The estimators can then be obtained by maximizing the log-likelihood function.

It should be noted, however, that Poisson Regression assumes a very restrictive hypothesis that all moments of the distribution are equal to the mean. In particular, this entails assuming that $E(L|X) = Var(L|X)$. To deal with this limitation, a quasi-maximum likelihood analysis can be used, easily executed through econometric programming languages. Particularly, in the case of QMLE (quasi-maximum likelihood estimation), it is assumed that:

$$Var(L|X) = \varphi E(L|X) \tag{3.15}$$

where $\varphi > 0$ is an unknown parameter (WOOLDRIDGE, 2003).

An even more general solution, used by the article of Kim et al. (2009), is the Negative Binomial Regression. The definition of this model assumes that the first and second moments are given, respectively, by:

$$E(L|X) = \mu \tag{3.16}$$

$$Var(L|X) = \mu + \alpha\mu^2 \tag{3.17}$$

where α is the overdispersion parameter (HILBE, 2011). Thus, while the variance of the Poisson QMLE model is defined by a linear function of μ, the variance of the Negative Binomial Regression is a quadratic function of μ. In this work, we will adopt the Negative Binomial Regression, approaching the methodology employed by the authors.

4 Results

4.1 PNS 2019: An Overview

As mentioned in the previous section, the 2019 PNS dataset contains 293,726 observations of individuals interviewed across 108,475 households.

Of this sample, 14,344 individuals did not respond to the question: "Do you have any private health insurance plan, from a company or public agency?". For the purposes of regression analysis, these observations will be treated as missing data and excluded from the study.

Regarding the same question, 58,597 individuals (26.5%) responded affirmatively, indicating they had some form of health insurance coverage. This percentage aligns with data from the ANS (National Supplementary Health Agency) for the same period. According to the ANS, a significant proportion of plans are corporate collective plans (approximately 70%). This information is available in the PNS microdata and will be incorporated into one of the extensions of the econometric tests, providing motivation for using a multinomial model.

4.2 Positive Correlation Test

4.2.1 The Canonical Model

The implementation of the positive correlation test on the 2019 PNS data, as well as the extensions of the test, was primarily conducted using the *statsmodels* library in Python. The matrix X was segmented based on age groups, in accordance with legal restrictions, and divided into the following age ranges: 0-18 years, 19-23 years, 24-28 years, 29-33 years, 34-38 years, 39-43 years, 44-48 years, 49-53 years, 54-59 years, and 59 years or older. Two probit regressions were then performed, both demonstrating statistical significance.

It is important to emphasize the binomial nature of the canonical model. As such, it is necessary to construct the variables $Coverage$ and $Loss$. The coverage *dummy* variable takes the value of 1 for individuals who answered "yes" to the question "Do you have any private health insurance plan, from a company or public agency?" and 0 otherwise. The $Loss$ variable follows the threshold defined by Fonseca (2017), which considers individuals who answered the question "How many times did you consult a doctor in the last 12 months?" with a number of 3 or more to be claimants. It is evident that this threshold is

Table 1 – Probit Regression Results of Coverage for the canonical model

Dep. Variable:	coverage_plan	No. Observations:	279382
converged:	True	LL-Null:	-1.4349e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

variable	coef	std err	z	P>|z|
Intercept	-0.9381	0.006	-170.010	0.000
C(age_group)[T.19-23]	-0.0597	0.012	-5.168	0.000
C(age_group)[T.24-28]	0.0421	0.012	3.634	0.000
C(age_group)[T.29-33]	0.1615	0.011	14.187	0.000
C(age_group)[T.34-38]	0.2212	0.011	20.167	0.000
C(age_group)[T.39-43]	0.1965	0.011	17.704	0.000
C(age_group)[T.44-48]	0.1701	0.012	14.693	0.000
C(age_group)[T.49-53]	0.1795	0.012	15.311	0.000
C(age_group)[T.54-59]	0.2143	0.011	19.134	0.000
C(age_group)[T.59+]	0.2747	0.008	32.829	0.000

Table 2 – Probit Regression Results of the Risk of Loss for the canonical model

Dep. Variable:	Risk of Loss	No. Observations:	279382
converged:	True	LL-Null:	-1.8223e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

coef	std err	z	P>|z|	
Intercept	-0.4558	0.005	-93.656	0.000
C(age_group)[T.19-23]	-0.2007	0.010	-19.398	0.000
C(age_group)[T.24-28]	-0.1082	0.011	-10.299	0.000
C(age_group)[T.29-33]	-0.0356	0.011	-3.389	0.001
C(age_group)[T.34-38]	-0.0063	0.010	-0.615	0.539
C(age_group)[T.39-43]	0.0134	0.010	1.309	0.191
C(age_group)[T.44-48]	0.0679	0.011	6.429	0.000
C(age_group)[T.49-53]	0.1648	0.011	15.508	0.000
C(age_group)[T.54-59]	0.2593	0.010	25.543	0.000
C(age_group)[T.59+]	0.4653	0.008	61.431	0.000

somewhat arbitrary and provides limited insight into individuals' actual claims behavior, a point that will be further explored in the section on multinomial modeling later in this work.

Following the methodology, an analysis of the correlation between the residuals from the two probit regressions was conducted, revealing significant correlation coefficients.

The W statistic, as proposed by Chiappori and Salanie (2000), was computed using the residual vectors and yielded a value of $W = 6092.7$, which is far above the critical value of 3.84. Supporting this finding, the Pearson correlation coefficient was 0.1585, the Spearman coefficient was 0.2644, and Kendall's Tau was 0.2160. All these coefficients were accompanied by low p-values, indicating a statistically significant correlation among the regression residuals.

The results from these regressions, along with their statistical significance, provide strong evidence for the presence of asymmetric information in the Brazilian supplementary health insurance market.

4.2.2 Preference Heterogeneity and Unused Variables

To examine the two extensions of the positive correlation test, we propose the inclusion of the *smoker* variable. This variable will be derived from the more detailed section of the PNS, which contains a sample of 90,846 individuals. For the purpose of constructing the *dummy* variable, a smoker will be defined as an individual who responded affirmatively to either "Do you currently smoke any tobacco product?" or "In the past, did you smoke any tobacco product daily?".

For the first extension, we will have the following model:

$$\mathbb{1}Coverage_i = \beta_0 + \beta_1 \cdot Smoker_i + X_i\Gamma + \epsilon_i \tag{4.1}$$

$$Loss_i = \alpha_0 + \alpha_1 \cdot Smoker_i + X_i\Pi + \eta_i \tag{4.2}$$

The hypothesis here would be that an individual who smokes would be less risk-averse regarding health and would attribute less utility to a health plan, for example. The idea of including the smoker variable, therefore, would be to control the model for different risk aversions of individuals.

Repeating the above regressions with this model, we arrive at a $W = 1657.7$, which allows us to conclude that the phenomenon of Asymmetric Information remains statistically significant.

Recalling the definition of the two extensions in section 3.2.1, we can use, more generally, the same regression within the context of the second extension. Thus, we will consider the smoker variable as an unused variable, a first candidate among a series of variables that the granularity of PNS 2019 allows us to test. Furthermore, the choice of this second extension also implies not defining *smoker* solely as a *proxy* for individual risk aversion, but rather as a variable that can contribute both to preference heterogeneity and to private information arising from different risk profiles.

$$Coverage_i = X_i\Gamma + \alpha smoker_i + \epsilon_i \tag{4.3}$$

$$Loss_i = X_i\Pi + \delta smoker_i + \eta_i \tag{4.4}$$

In this case, the W statistic will not be used. For this model, rejecting $\{\alpha = 0, \delta = 0\}$ is tantamount to rejecting the null hypothesis of symmetric information.

Table 3 – Probit Regression Results of Coverage - smoker variable

Dep. Variable:	coverage_plan	No. Observations:	90846
converged:	True	LL-Null:	-48593.
Covariance Type:	nonrobust	LLR p-value:	1.456e-230

variable	coef	std err	z	P>|z|
Intercept	-1.0401	0.032	-32.426	0.000
C(age_group)[T.19-23]	0.0486	0.038	1.275	0.202
C(age_group)[T.24-28]	0.1564	0.037	4.267	0.000
C(age_group)[T.29-33]	0.3094	0.036	8.682	0.000
C(age_group)[T.34-38]	0.3692	0.035	10.508	0.000
C(age_group)[T.39-43]	0.3763	0.035	10.682	0.000
C(age_group)[T.44-48]	0.3270	0.036	9.175	0.000
C(age_group)[T.49-53]	0.3122	0.036	8.726	0.000
C(age_group)[T.54-59]	0.3630	0.035	10.334	0.000
C(age_group)[T.59+]	0.4235	0.033	12.733	0.000
smoking_status	-0.3620	0.015	-23.566	0.000

Table 4 – Probit Regression Results of Risk of Loss - smoker variable

Dep. Variable:	Risk of Loss	No. Observations:	90846
converged:	True	LL-Null:	-61082.
Covariance Type:	nonrobust	LLR p-value:	0.000

variable	coef	std err	z	P>|z|
Intercept	-0.7042	0.029	-24.613	0.000
C(age_group)[T.19-23]	0.2163	0.034	6.428	0.000
C(age_group)[T.24-28]	0.2637	0.033	8.073	0.000
C(age_group)[T.29-33]	0.3070	0.032	9.589	0.000
C(age_group)[T.34-38]	0.3160	0.032	9.994	0.000
C(age_group)[T.39-43]	0.3320	0.032	10.471	0.000
C(age_group)[T.44-48]	0.4041	0.032	12.645	0.000
C(age_group)[T.49-53]	0.4895	0.032	15.301	0.000
C(age_group)[T.54-59]	0.5853	0.031	18.620	0.000
C(age_group)[T.59+]	0.7559	0.030	25.413	0.000
smoking_status	-0.2072	0.013	-15.887	0.000

The results obtained allow us to reject $\{\alpha = 0, \delta = 0\}$ for any usual significance level, so that we can identify the *smoker* variable as a source of Asymmetric Information.

4.2.3 Testing Other Unused Variables

As discussed previously, a possible application of the second extension of the positive correlation test is the verification of other unused variables by insurance companies as potential sources of Asymmetric Information.

For the choice of candidate variables, individual characteristics that are not used for premium determination for institutional reasons will be chosen. Brazilian legislation prohibits risk selection and, to address this limitation, we will test the diagnosis of diabetes - a metabolic syndrome associated with a series of chronic complications - as a candidate variable. Additionally, within the institutional context of the Brazilian supplementary health market, the inability to adjust premiums after 60 years of age is also highly relevant, which will be addressed by including additional age brackets to capture this increased risk of loss. In Figure 3, a significant increase in average claims is observed in age groups starting at 60 years.

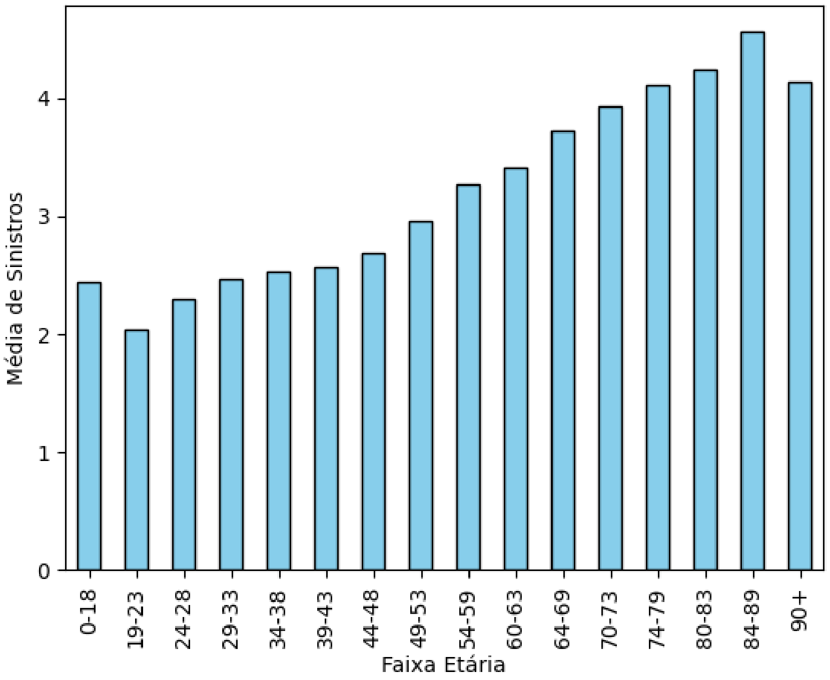

Figure 3 – Average Claims by Age Group, with Additional Brackets

Own elaboration based on PNS 2019 data

Using the PNS 2019 data, we can now test in the matrix X_i the diabetic variables, as well as age variables that differentiate from 60 years onwards. The sample space for this first variable is 84,073 respondents to the question “Has any doctor ever diagnosed you with diabetes?”, to which 7,374 interviewees (8.77%) responded positively. For the other age groups, we defined the categorical variables from the same variable used in the canonical model, now with a greater number of age groups.

Table 5 – Probit Regression Results for Coverage - diabetes and additional age groups

Dep. Variable:	coverage_plan	No. Observations:	84073
converged:	True	LL-Null:	-46357.
Covariance Type:	nonrobust	LLR p-value:	1.194e-73

variable	coef	std err	z	P>|z|
Intercept	-0.9314	0.035	-26.263	0.000
C(age_group)[T.19-23]	-0.0183	0.042	-0.438	0.661
C(age_group)[T.24-28]	0.0683	0.040	1.709	0.087
C(age_group)[T.29-33]	0.2221	0.039	5.709	0.000
C(age_group)[T.34-38]	0.2790	0.038	7.266	0.000
C(age_group)[T.39-43]	0.2819	0.038	7.330	0.000
C(age_group)[T.44-48]	0.2187	0.039	5.629	0.000
C(age_group)[T.49-53]	0.1944	0.039	4.991	0.000
C(age_group)[T.54-59]	0.2356	0.038	6.146	0.000
C(age_group)[T.60-63]	0.2387	0.040	5.949	0.000
C(age_group)[T.64-69]	0.2656	0.039	6.792	0.000
C(age_group)[T.70-73]	0.2950	0.042	7.005	0.000
C(age_group)[T.74-79]	0.3389	0.042	8.137	0.000
C(age_group)[T.80-83]	0.4007	0.048	8.264	0.000
C(age_group)[T.84-89]	0.4315	0.051	8.418	0.000
C(age_group)[T.90+]	0.4897	0.066	7.389	0.000
diabetes_status	-0.0294	0.017	-1.697	0.090

Table 6 – Probit Regression Results for Risk of Loss - diabetes and additional age groups

Dep. Variable:	Risk of Loss	No. Observations:	84073
converged:	True	LL-Null:	-57130.
Covariance Type:	nonrobust	LLR p-value:	0.000

variable	coef	std err	z	P>|z|
Intercept	-0.6105	0.032	-18.904	0.000
C(age_group)[T.19-23]	0.1744	0.038	4.642	0.000
C(age_group)[T.24-28]	0.1955	0.036	5.385	0.000
C(age_group)[T.29-33]	0.2271	0.036	6.381	0.000
C(age_group)[T.34-38]	0.2344	0.035	6.662	0.000
C(age_group)[T.39-43]	0.2308	0.035	6.548	0.000
C(age_group)[T.44-48]	0.3005	0.035	8.473	0.000
C(age_group)[T.49-53]	0.3607	0.035	10.163	0.000
C(age_group)[T.54-59]	0.4268	0.035	12.207	0.000
C(age_group)[T.60-63]	0.4635	0.037	12.655	0.000
C(age_group)[T.64-69]	0.5127	0.036	14.353	0.000
C(age_group)[T.70-73]	0.5769	0.039	14.931	0.000
C(age_group)[T.74-79]	0.6291	0.038	16.417	0.000
C(age_group)[T.80-83]	0.7247	0.045	15.980	0.000
C(age_group)[T.84-89]	0.7284	0.048	15.114	0.000
C(age_group)[T.90+]	0.7507	0.064	11.817	0.000
diabetes_status	0.5471	0.016	33.613	0.000

We can note that the *diabetic* variable does not pass the unused variables test. It is statistically significant for the Risk of Loss regression (being diabetic is correlated with higher health plan utilization), but it is not statistically significant for the Coverage regression. Therefore, it is not possible to reject the null hypothesis of symmetric information for this variable.

The age group divisions beyond what is currently allowed by legislation are statistically significant for both regressions and, therefore, according to the extension of Finkelstein and Poterba (2014), are a source of asymmetric information.

4.2.4 A multinomial model proposal for Coverage

As previously discussed, Kim et al. (2009) propose a multinomial model for insurance coverage to address the possibility of the insured choosing from a menu of optional coverages. In the authors' article, the analyzed market is automotive insurance.

In this work, we propose applying these authors' model to the PNS 2019 case. In Brazilian health insurance, it is known that a significant portion of policies is paid by the employer. Another portion, however, is entirely paid by the individual. This peculiarity of supplementary health insurance is potentially a problem for verifying Asymmetric Information, given that there will be a difference in the individual's utility whether they are responsible for paying the policy or not, for the same coverage. If the differentiation between policy types is not made, the positive correlation test result may lead to misleading conclusions, since the decision to join the plan did not necessarily come from the individual themselves.

To mitigate this problem, supplementary health coverage is modeled using Kim et al. (2009)'s multinomial proposal. With ${Coverage_i}^m$ being the variable representing different types of coverage, we can define ${Coverage_i}^m = 0$ for individuals without coverage, ${Coverage_i}^m = 1$ for individuals with coverage not directly paid by them, and ${Coverage_i}^m = 2$ for individuals whose coverage is paid by themselves.

To construct this variable, we will use the PNS 2019 question "Who pays the monthly fee for this health plan," and we will assign the value 2 to the variable only for those who answered "Only the policyholder or another household resident."

With this variable constructed, it is possible to estimate an ordered probit model using the same independent variables as the canonical model. The model implementation is done using the R language.

Table 7 – Multinomial Model - Ordered Probit Regression of Coverage

variable	coef	std error	t
C(age_group)[T.19-23]	0.5828	0.02101	27.74
C(age_group)[T.24-28]	0.9986	0.01882	53.06
C(age_group)[T.29-33]	1.1975	0.01806	66.30
C(age_group)[T.34-38]	1.2542	0.01757	71.40
C(age_group)[T.39-43]	1.2347	0.01762	70.07
C(age_group)[T.44-48]	1.2010	0.01801	66.69
C(age_group)[T.49-53]	1.2073	0.01804	66.94
C(age_group)[T.54-58]	1.2580	0.01741	72.27
C(age_group)[T.59+]	1.3495	0.01502	89.84

intercept	coef	std err	t
0|1	2.1203	0.0133	159.2695
1|2	2.5655	0.0137	187.3893

The t-values allow us to accept that there is statistical significance in the obtained results. Following the model proposed by the authors, from the intercepts estimated by the Ordered Probit, we can calculate the generalized residuals from the following equations:

$$\hat{\epsilon_i}^{m_1} = \frac{\phi(X_i\Omega - \hat{\mu_1})}{\Phi(X_i\Omega - \hat{\mu_1})(1 - \Phi(X_i\Omega - \hat{\mu_1}))}[Coverage_i{}^1 - \Phi(X_i\Omega)] \tag{4.5}$$

$$\hat{\epsilon_i}^{m_2} = \frac{\phi(X_i\Omega - \hat{\mu_2})}{\Phi(X_i\Omega - \hat{\mu_2})(1 - \Phi(X_i\Omega - \hat{\mu_2}))}[Coverage_i{}^2 - \Phi(X_i\Omega)] \tag{4.6}$$

being $\hat{\mu_1} = 2.1203$ and $\hat{\mu_2} = 2.5655$.

As discussed in the Methodology section, to allow the use of a discrete quantitative variable in the Risk of Loss equation, we will use a Negative Binomial Regression, a more general model than the Poisson Regression.

Table 8 – Multinomial Model - Negative Binomial Regression of Risk of Loss using calculated residuals as regressors

variable	**coef**	**std err**	**z**	**Pr(>|z|)**
(Intercept)	1.147558	0.004013	285.976	< 2e-16 ***
C(age_group)[T.19-23]	-0.048826	0.008397	-5.815	6.07e-09 ***
C(age_group)[T.24-28]	0.036746	0.008464	4.342	1.41e-05 ***
C(age_group)[T.29-33]	0.064822	0.008414	7.704	1.32e-14 ***
C(age_group)[T.34-38]	0.068681	0.008127	8.451	< 2e-16 ***
C(age_group)[T.39-43]	0.064389	0.008131	7.919	2.39e-15 ***
C(age_group)[T.44-48]	0.090464	0.008327	10.864	< 2e-16 ***
C(age_group)[T.49-53]	0.153246	0.008277	18.515	< 2e-16 ***
C(age_group)[T.54-59]	0.216001	0.007809	27.661	< 2e-16 ***
C(age_group)[T.59]+	0.302227	0.005808	52.034	< 2e-16 ***
calculated_residual_1	0.072102	0.001561	46.177	< 2e-16 ***
calculated_residual_2	-0.005298	0.002449	-2.163	0.0305 *

We can verify, once again, that there is statistical significance, for a significance level $\alpha = 5\%$, for both residuals used as regressors. The model employed was not only suitable for providing a more comprehensive analysis of Coverage and Risk of Loss but also reaffirmed the findings from previous regressions, highlighting the presence of Asymmetric Information in the Brazilian supplementary health market.

4.3 Moral Hazard vs. Adverse Selection: a brief discussion

As discussed previously, the positive correlation test is robust in identifying private information about individuals' risk types and, therefore, robust for verifying Asymmetric Information in insurance markets. However, it is unable to verify whether this asymmetry is associated with the phenomenon of Adverse Selection or Moral Hazard.

> The unused observables test, like the positive correlation test, is a test for asymmetric information. Without additional information, rejecting the null hypothesis of symmetric information **does not offer evidence on the question of whether asymmetric information results from moral hazard or from selection**. In some cases, such additional information may be available. For example, when a researcher has evidence suggesting that an unused observable variable is correlated with the risk of loss even among individuals who have identical insurance coverage, then finding that individuals with certain values of the unused observable select more insurance coverage supports the presence of selection based on ex-ante private information rather than moral hazard based on ex-post private information. (FINKELSTEIN; POTERBA, 2014)

The authors, in the text above, argue that having additional information that allows concluding that an unused variable is correlated with the risk of loss, even in individuals with exactly the same insurance coverage, would provide evidence of *ex-ante* private information, that is, the phenomenon of adverse selection. In the case of unused variables with statistical significance discussed in previous results, this correlation is quite trivial. The positive correlation between health service utilization and age, as well as between health service utilization and smoking, is undisputed (KALSETH; HALVORSEN, 2020; SIMONS et al., 2023). According to the authors of the unused variables test, this would be evidence in favor of the *ex-ante* nature of private information identified by the econometric test.

It is important to emphasize, however, that given this work's focus on the econometric approach and, especially, given the eminently qualitative nature of the differentiation proposed above, we recognize this as a limitation of this monograph.

Conclusion

The evidence found in this study corroborates the hypothesis of the presence of Asymmetric Information in the supplementary health market in Brazil. The variables *smoker* and the additional age groups are identified as sources of this market failure, according to the results of the positive correlation test and its extensions. These variables, not used for premium differentiation, prove to be relevant to this informational problem.

Interestingly, the variable *diabetic* did not yield similar results. While it showed statistical significance in the loss equation, this was not the case for coverage. This outcome suggests that private information regarding the risk status of diabetics does not influence decisions about policy enrollment. Thus, we do not reject the null hypothesis of symmetric information for this variable.

A multinomial model using Negative Binomial Regression further confirmed the presence of asymmetric information, as evidenced by the statistical significance of both calculated residuals used as regressors. This approach is promising because it allows for more flexible modeling, enabling the distinction between different types of coverage and the evaluation of claims variables beyond a binary framework.

This study recognizes the challenge of distinguishing between adverse selection and moral hazard, a well-known limitation of the positive correlation test. Future research could expand upon the qualitative argument presented by Finkelstein and Poterba (2014) in subsection 4.3 by developing an empirical approach to further clarify this distinction.

While the PNS dataset provides an extensive range of variables, only a select few were tested here to keep the study within scope. Further research could explore additional variables or incorporate data from other sources to better capture the dimensions of asymmetric information.

The extension of the econometric model to incorporate multiple coverage types is presented here as an original contribution within the context of the PNS. This approach offers significant potential for future research, as it can be applied to various types of insurance beyond the supplementary health market.

The findings here, along with the challenges facing Brazil's supplementary health market, bring to mind the iterative simulations by Winssen, Kleef and Ven (2018), which illustrate the potential collapse of markets plagued by adverse selection. Would expanded options for premium differentiation help mitigate this market failure? To what degree can the current crisis in supplementary health be attributed to asymmetric information, and to what extent should insurers be allowed to assess the risk of insureds?

While these questions fall well outside the scope of this study, they underscore the broader significance of the issue, touching upon economic theories, ethical considerations, and societal values. We hope this study contributes to these discussions, as the demographic, labor, and technological shifts underway in Brazil will likely make such questions ever more pressing, prompting Brazilian society to grapple with increasingly complex decisions.

References

AKERLOF, G. A. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, v. 84(3), p. 488–500, 1970. Citado 2 vezes nas páginas 13 and 16.

ALVES, S. L. Estimando seleção adversa em planos. *Revista Economia*, 2004. Citado 2 vezes nas páginas 17 and 21.

ANS. Súmula normativa 27/2015 da agência nacional de saúde suplementar. v. 11 de junho de 2015, 2015. Citado na página 12.

ARROW, K. J. Uncertainty and the welfare economics of medical care. *The American Economic Review*, American Economic Association, v. 53, n. 5, p. 941–973, 1963. ISSN 00028282. Available at: <http://www.jstor.org/stable/1812044>. Citado na página 16.

CARDON, J. H.; HENDEL, I. Asymmetric information in health insurance: Evidence from the national medical expenditure survey. *The RAND Journal of Economics*, v. 32(3), p. 408–427, 2001. Citado na página 17.

CHIAPPORI, P.-A.; SALANIE, B. Testing for asymmetric information in insurance markets. *Journal of Political Economy*, v. 108(1), p. 56–78, 2000. Available at: <https://doi.org/10.1086/262111>. Citado 5 vezes nas páginas 5, 6, 12, 16, and 27.

COHEN, A.; SIEGELMANN, P. Testing for adverse selection in insurance markets. *Journal of Risk and Insurance*, v. 77(1), p. 39–84, 2010. Citado na página 16.

Conselho Nacional de Justiça. *Saúde suplementar pontua impacto de processos judiciais para equilíbrio do setor*. 2024. Acessado em: 5 de junho de 2024. Available at: <cnj.jus.br/saude-suplementar-pontua-impacto-de-processos-judiciais-para-equilibrio-do-setor/>. Citado na página 12.

CUTLER, D. M.; FINKELSTEIN, A.; MCGARRY, K. Preference heterogeneity and insurance markets: Explaining a puzzle of insurance. *The American Economic Review*, v. 98(2), p. 158–161, 2008. Citado 2 vezes nas páginas 17 and 22.

CUTLER, D. M.; ZECKHAUSER, R. J. The anatomy of health insurance. *Handbook of Health Economics*, v. 1, p. 563–643, 2000. Citado na página 16.

DARDANONI, V.; FORCINA, A.; DONNI, P. L. Testing for asymmetric information in insurance markets: A multivariate ordered regression approach. *The Journal of Risk and Insurance*, v. 85(1), p. 107–125, 2018. Citado 2 vezes nas páginas 18 and 23.

EINAV, L.; FINKELSTEIN, A. Selection in insurance markets: theory and empirics in pictures. *Journal of Economic Perspectives*, v. 25, n. 1, p. 115–138, Winter 2011. Citado na página 15.

FINKELSTEIN, A.; POTERBA, J. Testing for asymmetric information using “unused observables” in insurance markets: Evidence from the u.k. annuity market. *Journal of Risk and Insurance*, v. 81(4), p. 709–734, 2014. Citado 10 vezes nas páginas 5, 6, 15, 17, 21, 22, 23, 32, 34, and 36.

FONSECA, R. B. d. A. Informational frictions in the brazilian health insurance market. *Dissertação de Mestrado submetida à EPGE*, 2017. Citado 4 vezes nas páginas 17, 18, 23, and 26.

HEUVEL, E. van den; ZHAN, Z. Myths about linear and monotonic associations: Pearson's r, spearman's , and kendall's . *The American Statistician*, Taylor & Francis, v. 76, n. 1, p. 44–52, 2022. Available at: <https://doi.org/10.1080/00031305.2021.2004922>. Citado na página 21.

HILBE, J. M. *Negative Binomial Regression*. 2nd. ed. Cambridge: Cambridge University Press, 2011. ISBN 978-0521198158. Citado na página 25.

KALSETH, J.; HALVORSEN, T. Health and care service utilisation and cost over the life-span: a descriptive analysis of population data. *BMC Health Services Research*, v. 20, n. 1, p. 435, 2020. Available at: <https://doi.org/10.1186/s12913-020-05295-2>. Citado na página 35.

KIM, H. et al. Evidence of asymmetric information in the automobile insurance market: Dichotomous versus multinomial measurement of insurance coverage. *The Journal of Risk and Insurance*, v. 76(2), p. 343–366, 2009. Citado 7 vezes nas páginas 5, 6, 18, 23, 24, 25, and 32.

MUGNATTO, S. *Aumentam reclamações de consumidores sobre cancelamentos unilaterais de planos de saúde*. 2024. Acessado em: 5 de junho de 2024. Available at: <camara.leg.br/noticias/1062863-aumentam-reclamacoes-de-consumidores/-sobre-cancelamentos-unilaterais-de-planos-de-saude/>. Citado na página 12.

NISHIJIMA, M.; POSTALI, F. A. S.; FAVA, V. L. Consumo de serviços médicos e marco regulatório no mercado de seguro de saúde brasileiro. *Revista Pesquisa e Planejamento Econômico*, v. 41(3), p. 509–531, 2011. Citado na página 17.

OCKé-REIS, C. O.; FIUZA, E. P. S.; COIMBRA, P. H. H. Inflação dos planos de saúde 2000-2018. *Nota Técnica do Instituto de Pesquisa Econômica Aplicada (IPEA)*, v. 54, 2019. Citado na página 11.

PIOVEZAN, S. Reclamações contra planos de saúde disparam e chegam a quase 900 por dia. *Folha de São Paulo*, v. 09 de Setembro de 2023, 2023. Citado na página 12.

RESENDE, M.; ZEIDAN, R. Adverse selection in the health insurance market: some empirical evidence. *The European journal of health economics*, v. 11(4), p. 413–418, 2010. Citado 2 vezes nas páginas 17 and 21.

ROBERTSON, C. T. et al. Distinguishing moral hazard from access for high-cost healthcare under insurance. *PLoS ONE*, v. 15, n. 4, p. e0231768, 2020. Citado na página 14.

ROTSCHILD, M.; STIGLITZ, J. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The Quarterly Journal of Economics*, v. 90(4), p. 629–649, 1976. Citado 2 vezes nas páginas 13 and 16.

SIMONS, K. et al. Age and gender patterns in health service utilisation: Age-period-cohort modelling of linked health service usage records. *BMC Health Services Research*, v. 23, p. 480, 2023. Available at: <https://doi.org/10.1186/s12913-023-09456-x>. Citado na página 35.

Sá, M. C. de. Risco moral e seleção adversa de beneficiários no mercado de saúde suplementar. *XXXII Encontro nacional de engenharia de produção*, 2012. Citado na página 17.

WINSSEN, K. P. M. van; KLEEF, R. C. van; VEN, W. P. M. M. van de. Can premium differentiation counteract adverse selection in the dutch supplementary health insurance? a simulation study. *The European journal of health economics : HEPAC : health economics in prevention and care*, v. 19(5), p. 757–768, 2018. Citado 2 vezes nas páginas 14 and 36.

WOOLDRIDGE, J. M. *Introductory Econometrics: A Modern Approach*. 1st. ed. Cincinnati, OH: South-Western College Publishing, 2003. Citado 2 vezes nas páginas 24 and 25.

WORLD HEALTH ORGANIZATION. *53rd Directing Council, 66th Session of the Regional Committee of WHO for the Americas*. Washington, D.C., USA, 2014. 53rd Directing Council, 66th Session of the Regional Committee of WHO for the Americas. Citado na página 12.

APPENDIX A – R and Python Code

Code for the implemented models and data processing available at: github.com/enricoruggieri/informationasymmetryPNS

The canonical models and extensions of preference heterogeneity and unused variables were implemented in Python, primarily using the *statsmodels* library.

The multinomial model was implemented in R, due to more suitable libraries for Ordered Probit Regression and Negative Binomial Regression required for this model.