UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE CIÊNCIAS ECONÔMICAS
DEPARTAMENTO DE CIÊNCIAS ECONÔMICAS

BRUNO GRILLO DE BERMÚDEZ

ANÁLISE DA ELASTICIDADE PREÇO DE BENEFICIÁRIOS DE PLANOS DE SAÚDE PRIVADOS EM RELAÇÃO ÀS CONSULTAS MÉDICAS

Porto Alegre
2015

BRUNO GRILLO DE BERMÚDEZ

ANÁLISE DA ELASTICIDADE PREÇO DE BENEFICIÁRIOS DE PLANOS DE SAÚDE PRIVADOS EM RELAÇÃO ÀS CONSULTAS MÉDICAS

Monografia submetida ao curso de graduação em Economia da Faculdade de Ciências Econômicas da Universidade Federal do Rio Grande do Sul, como quesito parcial para obtenção do título em Ciências Econômicas.

Orientador: Prof. Dr. Hudson da Silva Torrent

Porto Alegre

2015

à Edite, minha amada mãe.

RESUMO

O presente trabalho tem como objetivo central a identificação de padrões de comportamento de contratantes de planos de saúde frente a diferentes desenhos de contrato sob assimetria de informações, tais como risco moral e problema agente-principal. O foco principal é a mensuração da elasticidade-preço do bem consultas médicas, dada a existência de mecanismos – para uma parte da população – que alteram o custo marginal. A questão divide-se na verificação da existência do risco moral, ao observar que usuários cujo custo marginal é zero utilizam em excesso os recursos médicos – visto que estão abaixo do seu preço de mercado – e na magnitude deste efeito (elasticidade-preço). Para tanto, emprega-se modelos de contagem de dados – pois a variável de interesse é discreta – de uma e de duas etapas, cujo objetivo é tentar modelar a quantidade de consultas sob problema agente-principal. Na conclusão do estudo, verifica-se que o efeito da coparticipação (o mecanismo que torna o custo marginal diferente de zero) sob a quantidade de consultas é estatisticamente significativo.

Palavras-chave: economia da saúde, risco moral, modelo de contagem de dados, coparticipação, consultas médicas, econometria.

ABSTRACT

The present paper aims to identify health insurance contractors' behaviour patterns facing different sorts of contract designs under information asymmetry, such as moral hazard and principal-agent problem. The main focus lies in the price-elasticity of physician consult measurement given the existence of schemes - that apply only to a share of the population - whose purpose is to change the marginal costs through out-of-pocket payments. The problem is divided in two parts, being the first to assess whether moral hazard is present or not by observing that payees whose marginal cost is zero consume more resources - since their prices are under the market price - than their peers with higher copay, and then the magnitude of this phenomenon (price elasticity). For this purpose, one need to employ count data models - since the data of interest is discrete-, applying one-step and two-step models, whose intention is to model the physician consult under the principal-agent problem framework. At the conclusion, it is verified that the copayment effect (which turns the marginal cost greater than zero) on the amount of consults is statistically significant.

Keywords: health economics, moral hazard, count data models, copayment, physician consults, econometrics.

LISTA DE FIGURAS

LISTA DE TABELAS

1 Introdução

O principal objetivo desta monografia é verificar se diferentes desenhos de contrato levam os agentes a um uso mais parcimonioso dos recursos da área da saúde, especificamente em relação às consultas médicas. Investigar-se-á se o mecanismo de coparticipação – que consiste em um pagamento *ad valorem* sobre o custo total gerado por utilização ou um custo nominal – influencia a tomada de decisão dos agentes ao determinar a quantidade consumida do bem consultas médicas. Este mecanismo tem sua origem na literatura de risco moral, pois ao atribuir um custo marginal igual a zero para qualquer nível de consulta, os agentes não considerarão os preços para tomar decisões e espera-se que eles tomem decisões diferentemente do que o fariam caso necessitem desembolsar uma determinada quantia pelo uso adicional de um recurso. Conforme verificado na teoria econômica, espera-se que quanto maior for o preço do bem, menor será a quantidade demandada. O desiderato é, primeiramente, verificar se este fator moderador (a coparticipação) afeta o nível ótimo que cada agente consumirá e determinar, concomitantemente, a magnitude da variação na quantidade consumida em consequência de uma variação nos preços, ou seja, a elasticidade-preço.

Tentar-se-á estimar uma função de demanda por consultas médicas e de lá extrair-se-á a elasticidade preço. A literatura corrente sugere que há duas abordagens para verificar os efeitos de demanda dos agentes por bens de saúde, sendo uma a abordagem de uma etapa e outra de duas etapas. A idéia subjacente do método de duas etapas é a existência de dois processos governando a demanda, sendo o primeiro a decisão de consumir ou não, ao passo que o segundo determina o quanto. A explicação teórica para tal deriva-se do problema agente-principal, onde o paciente determina se haverá ou não contato, e subsequentemente ao contato, o médico determinará o consumo.

O estudo está estruturado em quatro capítulos de desenvolvimento do assunto e mais um de conclusão. No primeiro capítulo, faz-se uma revisão da teoria do consumidor com o intuito de elucidar os aspectos empregados na análise. Muitos fatores cruciais para entender o processo de tomada de decisões – como possuir ou não um plano, qual quantidade consumir, que depende do conjunto de possibilidades do consumidor – relacionam-se com a elasticidade-preço. Destarte, faz-se uma revisão completa, embora não detalhada, de aspectos importantes da teoria do consumidor, tais como a demanda, elasticidade-preço, utilidade do consumidor e tomada de decisões frente a incertezas. Considera-se o caso de seguro de contingência num contexto onde

ele é ótimo – quando não há assimetria de informação, tal como risco moral e problema agente-principal – e subsequentemente o caso onde este arranjo não é ótimo, uma vez que os agentes tem incentivos a consumir mais devido ao preço estar abaixo do preço de mercado. No segundo capítulo, sumarizar-se-á a literatura atinente à questão central, com o intuito de mostrar metodologias empregadas, problemas frequentes na análise e conclusões. No terceiro capítulo, procura-se mostrar um guia econométrico para lidar com o tipo de dados utilizados, onde começa-se com a regressão por mínimos quadrados ordinários, nesta seção faz-se uma abordagem completa – porém não exaustivamente detalhada – e mostra-se o método de estimação, hipóteses necessárias para inferir a consistência e eficiência dos estimadores, propriedades e testes para verificar perturbações das hipóteses, assim como maneiras de corrigi-las. Posteriormente abordam-se métodos de estimação que baseiam-se em máxima-verossimilhança, hipóteses e propriedades decorrentes destas, assim como testes para verifica-los e eventuais correções necessárias. Os métodos empregados serão calculados via R e serão estimados via algoritmos que estimam os estimadores de máxima verossimilhança por métodos iterativos – que não serão enfatizados neste trabalho. No terceiro capítulo haverá a aplicação dos métodos candidatos, estatísticas descritivas da variável de interesse, critérios para a escolha do método, assim como testes para os resultados obtidos pelo método escolhido.

2 Aspectos teóricos do comportamento

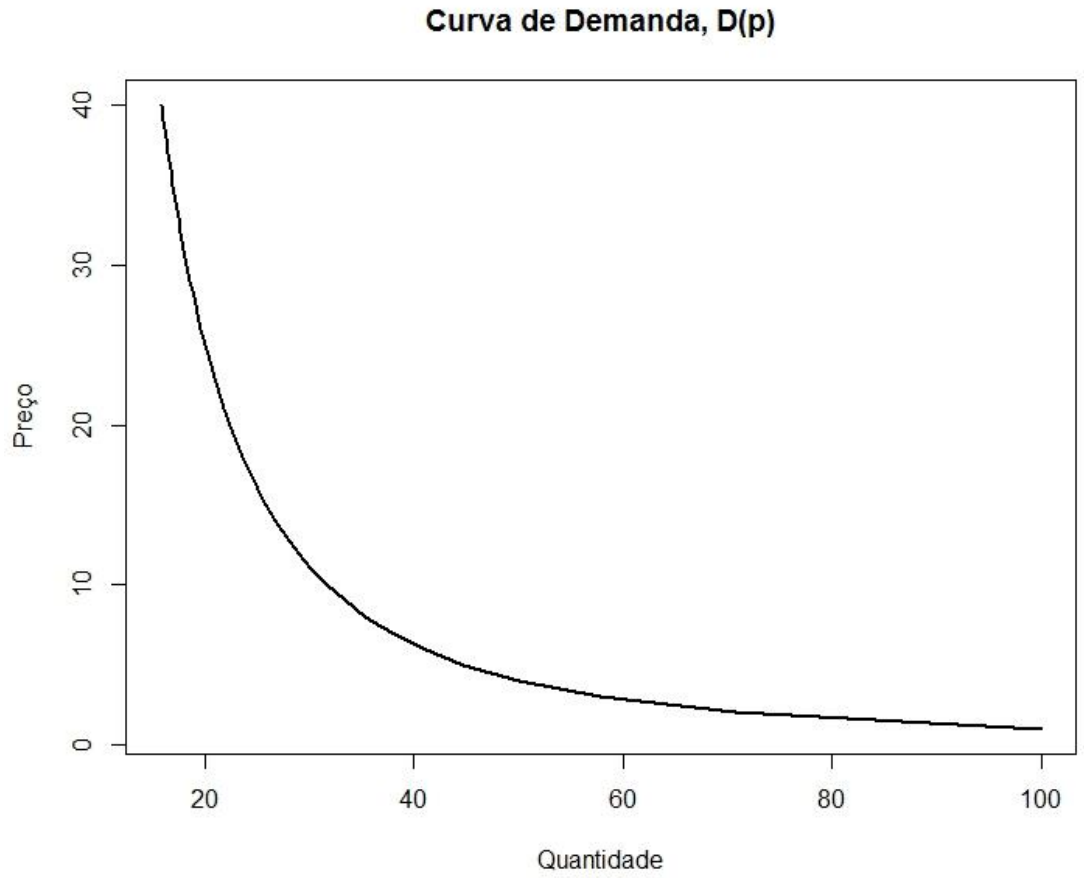
Neste capítulo, procura-se abordar pontos teóricos da demanda, assim como demonstrar argumentos aplicados à área da saúde. A teoria baseia-se em dos princípios, que são destacados por Varian (2006) como: (i) o princípio da otimização e (ii) o princípio de equilíbrio. O primeiro argumenta que os agentes tentam escolher o melhor padrão de consumo dada a restrição de recursos disponíveis, ao passo que a segunda assume que os preços ajustam-se até o ponto onde a demanda por um bem é igual à quantidade ofertada.

2.1 Demanda e elasticidade

A demanda é um princípio econômico que descreve o desejo/comportamento dos agentes econômicos de usufruir de determinado bem – assim como a quantidade dele – dado o seu preço, p. Visto que a demanda altera-se de acordo com o preço, podemos escrevê-la como uma função de p, assim, a demanda é escrita como $D(p)$. A curva de demanda é – convencionalmente[1] – negativamente relacionada ao preço, a inclinação da curva de demanda é dada por $\frac{\Delta q}{\Delta p}$. Em outras palavras, a quantidade demandada de um bem é menor se o preço deste bem aumentar, desta forma, diz-se que a primeira derivada da demanda em função do preço é negativa, $\frac{\partial D}{\partial p} < 0$. Pode-se visualizar um exemplo de curva de demanda na figura 1.

[1] Há casos em que pode ocorrer de maneira distinta, como, por exemplo, em bens de Giffen.

Figura 1 – Função de Demanda



Fonte: Elaboração Própria

Um conceito importante no contexto de demanda é o preço de reserva, p^*, que é o preço máximo que um agente está disposto a despender para consumir uma unidade de um determinado bem. Desta forma, se o preço de um bem for maior do que o preço de reserva do agente, a quantidade demandada será zero, isto é, se $p > p^*$, $D(p) = 0$. De acordo com a dotação monetária do agente, é possível que ele escolha consumir mais unidades de um determinado bem – já que o consumo do agente está restrito à renda e riqueza dele. No mundo de apenas um demandante, a demanda seria maior do que 0 se e somente se o preço fosse igual ou menor do que o preço de reserva do agente, desta forma, $D(p) > 0 \Leftrightarrow p \leq p^*$.

Para entender a magnitude da variação da quantidade demandada, faz-se necessário explicar o conceito de elasticidade-preço. Marshall – em seu livro *Principles of Economics* - sugere que há apenas uma lei universal, que é a redução do desejo de uma pessoa por um bem concomitantemente ao aumento da oferta deste bem *ceteris paribus*. Nos termos empregados por Marshall, essa diminuição pode ser lenta ou rápida. Caso

esta variação seja lenta, um aumento no preço causará uma redução relativamente alta na sua demanda por ele, entretanto, caso esta variação seja rápida, uma leve redução no preço causará uma variação pequena na quantidade do bem demandado por ele. No primeiro caso, a demanda pelo bem é elástica, ao passo que no último ela é inelástica. Com isto, Marshall afirma que a demanda é elástica – em diferentes magnitudes para bens distintos - tanto para uma redução quanto para um aumento no preço.

Formalmente, a elasticidade-preço é definida como a variação percentual na quantidade demandada dada a variação percentual do preço, desta forma, define-se ela como uma medida de sensibilidade. Uma propriedade importante da elasticidade é a independência de unidades, ou seja, não há importância se os preços são determinados em reais brasileiros ou euros, assim como se as quantidades são medidas em kg ou g. É importante notar que o grau de inclinação da curva de demanda também é uma medida de sensibilidade, todavia, ao utilizar a quantidade em kg em vez de gramas, a inclinação ficaria mil vezes menor, assim como alterar-se-ia caso unidades monetárias distintas – que não possuam equivalência unitária – fossem empregadas. Por isso, torna-se muito mais conveniente expressar variações no consumo dado variações no preço em termos de elasticidade, já que elimina a necessidade de especificar as unidades utilizadas. A elasticidade, ϵ, é expressa como

$$\epsilon = \frac{\Delta q/q}{\Delta p/p} = \frac{p}{q}\frac{\Delta q}{\Delta p}$$

Assim como é a demanda de um indivíduo, é também a demanda do mercado inteiro. Dito de maneira geral, a resposta da demanda a variações de preço é grande ou pequena de acordo com a variação na quantidade demandada. Em outras palavras, se um aumento no preço diminui muito ou pouco a quantidade demandada, e uma redução aumenta muito ou pouco esta. Marshall ainda salienta que as demandas de agentes podem ser distintas das agregadas dependendo do preço de reserva de cada agente. Outro ponto destacado por Marshall é que a demanda depende ainda do preço de bens rivais – ou seja, bens substitutos. Neste caso, definir-se-ia a elasticidade preço-cruzada. Além do preço de bens substitutos, há fatores elencados por Marshall como, por exemplo, renda, pois alguns bens relativamente caros são – ou são quase – proibitivos para as classes inferiores, ao passo que mal são significantes para os ricos.

2.2 Risco

A definição de risco no dicionário é: "Possibilidade de perigo – incerto mas previsível – que ameaça dano a pessoa ou a coisa". Diversas situações não apresentam resultados certos – dado o caráter não-determinístico de muitos fenômenos –, não obstante, podem ocorrer com uma determinada probabilidade. Em caso de ocorrência do evento, a utilidade do indivíduo é afetada, podendo causar tanto efeitos positivos – como, por exemplo, ao ganhar o prêmio de uma loteria – quanto negativos – ao ocorrer, por exemplo, um acidente. Muitos resultados acabam interferindo o bem-estar de maneiras não-mensuráveis, dado o caráter subjetivo das restrições impostas pela ocorrência do evento. Entretanto, para fins de esclarecimento e simplicidade de análise, faz-se necessário considerar variáveis que sejam objetivas e quantificáveis. Por isso, supor-se-á que os riscos acarretam apenas prejuízos pecuniários (afetando, com isso, a dotação monetária do agente e, por consequência, o seu consumo), assim como a existência de um conjunto de possibilidades restrito. Os diferentes resultados possíveis de um evento estocástico são definidos como estado de natureza.

Suponha um agente com dotação monetária y e com utilidade determinada por $u(c)$[2], onde c representa o seu consumo, que é uma função da cesta de bens consumida – denotada por x – sujeita à condição de que o gasto com consumo não exceda sua dotação monetária, isto é, $c(x)$ sujeito à restrição $p_x q_x \leq y$, onde p_x[3] representa o preço do bem e q_x a quantidade consumida. Ele está sujeito à ocorrência de um evento, cuja probabilidade de ocorrência é dada por ρ e a de não-ocorrência por $1 - \rho$, quando o evento materializa-se, o agente precisa arcar com um custo K. Se este evento não ocorrer, não será necessário arcar com custo algum, ou seja, $K = 0$. Supõem-se que o agente utilizará estes recursos para consumo, que é função da dotação monetária, denotado por $c(x)$. Deste modo, o agente enfrenta duas possibilidades

Cenário 1: $c(x)$ sujeito a $p_x q_x \leq y$ com probabilidade ρ

Cenário 2: $c(x)$ sujeito a $p_x q_x \leq y - K$ com probabilidade $1 - \rho$

É evidente que a utilidade do agente no cenário um é superior à no cenário dois, visto que ele terá menos recursos disponíveis para despender em consumo no segundo

[2] Supõe-se que a primeira derivada do consumo em relação ao consumo é positiva, $\frac{\partial u}{\partial c} > 0$, e que a segunda derivada é negativa, $\frac{\partial u^2}{\partial^2 c}$. Intuitivamente, quanto maior o consumo, maior será a utilidade do agente, contudo, o retorno de utilidade da unidade adicional de consumo é menor do que a anterior.

[3] Para não haver confusão com as notações, é crucial notar que o preço é denotado pela letra p, ao passo que a probabilidade é denotada pela letra grega rô ρ.

caso. Uma forma possível de controlar os eventuais danos sofridos é através de um seguro, de forma que o agente mantenha sua renda constante ao pagar um prêmio, que variará de acordo com (i) a potencial extensão do dano e (ii) a probabilidade de ocorrência. Quanto maior o risco, maior será o prêmio – denotado por π – cobrado *ceteris paribus*, esta relação também é verdadeira de acordo com a extensão do sinistro. Tanto a extensão quanto a probabilidade podem ser funções de outras variáveis.

Os agentes têm diferentes preferências sobre possíveis cestas de consumo e tomarão ações que reflitam suas preferências de consumo perante diferentes circunstâncias. Quanto os agentes estão dispostos a pagar para reduzir o risco e manter sua renda constante? Isso está muito relacionado ao grau de aversão de risco[4], agentes muito conservadores estarão mais inclinados a comprar um seguro e evitar riscos, ao passo que indivíduos mais arrojados estarão menos propensos à aquisição de um seguro.

A preferência de adquirir ou não proteção contra a incerteza depende muito da probabilidade que o agente crê que o evento ocorrerá e como dar-se-á sua dotação nos diferentes estados de natureza, ou seja, a crença do indivíduo atinente à probabilidade de ocorrência do evento influenciará o quão disposto ele estará a pagar para substituir consumo em um estado por consumo em outro. Por conseguinte, escreve-se a função de utilidade do agente relacionada ao consumo em cada estado e a probabilidade de ocorrência deles. Num caso dicotômico, temos que a função de utilidade do agente é dada por

$$U(c_1, c_2, \rho_1, \rho_2) = \rho_1 c_1 + \rho_2 c_2$$

Onde c_1 é o consumo no estado de natureza 1, c_2 é o consumo no estado dois (lembrando que os eventos são excludentes), ρ_1 é a probabilidade de ocorrência do primeiro estado de natureza e $\rho_2 = 1 - \rho_1$ é a probabilidade do segundo estado de natureza. De acordo com Varian (2006), essa função representa as preferências individuais de consumo em cada estado. Essa expressão é denominada – na literatura – de valor esperado, pois indica o nível de consumo médio, ponderado pelas probabilidades de ocorrência de cada evento.

É possível escrever a utilidade de forma que ela não seja diretamente relacionada ao consumo em cada estado, mas a funções do consumo em cada estado, $f(c)$, ou seja,

$$U(c_1, c_2, \rho_1, \rho_2) = \rho_1 f(c_1) + \rho_2 f(c_2)$$

[4] Que será discutido na sequência

Esta função em relação ao consumo também é ponderada pela probabilidade de ocorrência do evento. Quando um estado tem probabilidade igual a 1, a utilidade é dada por

$$U(c_1, c_2, \rho_1, \rho_2) = f(c_i)$$

Podendo, neste caso, o subscrito assumir dois valores. Essa expressão é conhecida na literatura como utilidade esperada (em razão de representar a utilidade média ponderada pelas probabilidades de ocorrência)[5]. Essa forma funcional satisfaz a propriedade de independência entre a taxa marginal de substituição (TMS) de dois bens e outros bens, ou seja,

$$TMS_{1,2} = -\frac{\Delta U(c_1, c_2, c_3)/\Delta c_1}{\Delta U(c_1, c_2, c_3)/\Delta c_2}$$

$$= -\frac{\rho_1 \Delta u(c_1)/\Delta c_1}{\rho_2 \Delta u(c_2)/\Delta c_2}$$

Desta forma, a taxa marginal de substituição depende apenas da dotação do bens 1 e 2. A taxa marginal de substituição é uma medida que mensura a taxa a que um agente está disposto a substituir o consumo de um bem pelo consumo de outro bem. Pode-se pensar nela como a quantidade de bem que um consumidor está disposto a trocar por outro bem, considerando que esse novo bem é igualmente prazeroso.

Suponha que um consumidor possua uma cesta (x_1, x_2) e que haja uma variação na cesta $(\Delta x_1, \Delta x_2)$, tal que Δ denota a magnitude e sentido da variação, movendo-o para uma nova cesta $(x_1 + \Delta x_1, x_2 + \Delta x_2)$. A taxa marginal de substituição determina a mudança que deixa o agente no mesmo nível de utilidade. Para ele manter-se no mesmo ponto, a mudança na utilidade resultante de um aumento de consumo de x_1 precisa ser exatamente igual à redução de utilidade associada à variação negativa de x_2. Desta forma, tem-se que

$$\frac{\partial u(x_1, x_2)}{\partial x_1}\Delta x_1 + \frac{\partial u(x_1, x_2)}{\partial x_2}\Delta x_2 = \Delta U = 0$$

A partir disto, deduz-se que a taxa marginal de substituição é

[5] As preferências podem ser descritas de outras formas, como, por exemplo, uma função Cobb-Douglas, cuja forma é $c_1^{\rho_1} c_2^{\rho_2}$. Apesar de representar as mesmas preferências que a sua transformação monotônica $\rho_1 \ln c_1 + \rho_2 \ln c_2$, ela não terá a propriedade de utilidade esperada, pois rompe com a hipótese de independência. Esta hipótese implica que a utilidade do consumo contingente deve ser aditiva. Intuitivamente, ao ocorrer algum evento, a utilidade do agente será determinada pelo consumo efetivo após a ocorrência e não pelo consumo que poderia ter ocorrido caso o evento fosse outro, pois os eventos são dicótomos.

$$\frac{\Delta x_2}{\Delta x_1} = -\frac{\dfrac{\partial u(x_1, x_2)}{\partial x_1}}{\dfrac{\partial u(x_1, x_2)}{\partial x_2}}.$$

A taxa marginal de substituição determina a inclinação da curva de indiferença. A curva de indiferença contém todas as combinações possíveis de (x_1, x_2) que mantém o consumidor no mesmo nível de utilidade, quanto mais bens o consumidor puder comprar, maior será a curva de indiferença em que ele estará. A escolha ótima da cesta será sempre o ponto onde a curva de indiferença é tangente à linha de restrição orçamentária – que determina as possíveis cestas dada a dotação monetária do agente. O caso de seguro pode ser visto como se o agente tivesse dois consumos possíveis, um quando ocorre o evento e outro quando não, este exemplo será tratado adiante.

Sob incertezas, a utilidade esperada dos agentes difere bastante de acordo com a maneira que o indivíduo lida com o risco, na literatura existe três denominações para os agentes de acordo com o modo que o agente encara o risco, que são: (i) avesso ao risco, (ii) propenso ao risco e (iii) neutro ao risco. Agentes propensos ao risco tem uma função de utilidade convexa, de tal modo que a inclinação fica mais íngreme ao aumentar a sua riqueza – denominada por w – (que converte-se em consumo), isto é, $\frac{d^2U(w)}{dw^2} > 0$. Por outro lado, a função de utilidade de agentes avessos ao risco é côncava, isto é $\frac{d^2U(w)}{dw^2} < 0$. O agente neutro ao risco é aquele que não importa-se com o risco, mas apenas com o valor esperado.

2.3 Princípios do seguro

O valor do seguro de saúde está calcado na imprevisibilidade dos gastos com saúde. Embora os agentes saibam algo sobre sua necessidade de serviços médicos, a quantia exata que eles gastarão é, em um nível significante, incerta. Os gastos com saúde são extremamente variáveis.

Agentes avessos ao risco quererão proteger-se contra o potencial risco de necessitar despender uma grande quantia. Uma maneira de fazer isso é tomar dinheiro emprestado quando está doente e devolvê-lo quando estiver bem, contudo, há alguns entraves para tal, pois há dificuldades em obter o empréstimo devido à incerteza concernente ao tempo de vida remanescente e/ou à recuperação do estado de saúde que permita o

pagamento deste. Uma alternativa mais razoável seria poupar recursos enquanto saudável a fim de dispor de recursos para eventuais gastos ao ficar doente. Contudo, algumas doenças exigem tratamentos com custos mais elevados que outras e os gastos com uma doença muito severa tornam a poupança para este fim quase impraticável. Far-se-ia também necessário reduzir muito o consumo para poupar para despesas que seriam encaradas por poucos. A solução natural é assegurar contra o risco de doença por compartilhamento de risco com os demais da população. O consumo anual seria reduzido para apenas o prêmio, que é o custo médio do provimento de saúde.

2.3.1 Seguro de contingência

O modelo mais comum de seguro é o que a doença impõe um custo fixo e a apólice está precificada no seu preço atuarial. Ao considerar o caso de uma doença, tem-se que as pessoas estão saudáveis com probabilidade $1-\rho$, o que requer gasto zero com saúde; e os agentes ficam doentes com probabilidade ρ. Denota-se a presença de doença por $d=1$ e a ausência por $d=0$. O gasto requerido para o tratamento de uma pessoa que está doente é denotado por m. Assume-se que a saúde de um doente após o dispêndio é $h=H[d,m]$ e que o gasto em saúde leva o agente ao estado de saúde perfeito, ou seja, $H[1,m]=H[0,0]$[6].

Os indivíduos tem utilidade u, que é uma função do consumo, definido por x, e a saúde, h. Deste modo, tem-se que $u=U(x,h)$. O consumo é definido pela renda remanescente dos gastos em saúde (tanto pelo prêmio do seguro quanto pelo eventual gasto necessário caso não esteja assegurado). A renda é descrita como y e se assume que os agentes não podem recorrer ao mercado financeiro para tomada de empréstimos, o prêmio do seguro π. Define-se o agente sem seguro com N e o que possui seguro com I. Deste modo, tem-se que o consumo é:

$$x=\begin{cases} y, & se\ d=0\ e\ \mathrm{N} \\ y-m, & se\ d=1\ e\ \mathrm{N} \\ y-\pi, & se\ d=\{0,1]e\ \mathrm{I} \end{cases}$$

[6] A situação é mais complexa quando o agente não consegue retornar ao estado de saúde perfeito após o gasto médico, e a utilidade marginal da renda é afetada pelo estado de saúde. Um custo fixo k pode ser imposto ao indivíduo relativo ao seu novo estado de saúde, tanto por gastos necessários – que reduziriam a renda – quanto por redução de *capabilities*. Para entender *capabilities*, recomenda-se ler Amartya Sen.

Na ausência de seguro, a utilidade esperada do agente é dada por:

$$V_N = (1-\rho)U(y, H[0,0]) + \rho U(y-m, H[1,m])$$

$$V_N = (1-\rho)U(y) + \rho U(y-m)$$

Assume-se que (i) a utilidade é crescente em função do consumo, conquanto que a uma taxa decrescente, ou seja, $\frac{dU}{\mathrm{dx}} > 0$ e que $\frac{dU^2}{\mathrm{d^2x}} < 0$; (ii) que o dispêndio em saúde é vantajoso mesmo que o agente não esteja assegurado.

Caso o agente possua seguro, o prêmio justo estipulado pela seguradora precisaria ser $\pi = \rho m$, que é o preço atuarial. A companhia de seguro coletaria o prêmio referente a um período de tempo e paga m quando o indivíduo fica doente. Desta forma, ao optar por um seguro, a utilidade do agente sempre seria dada por:

$$V_1 = U(y-\pi)$$

Ao considerar a utilidade esperada do agente na ausência de seguro $V_N = (1-\rho)U(y) + \rho U(y-m)$, é possível aproximá-la por série de Taylor à[7]

$$V_N \approx U(y-\pi) + U'\left(\frac{U''}{2U'}\right)\pi(m-\pi)$$

Portanto, o valor do seguro pode ser visto como

$$\frac{V_I - V_N}{U'} \approx \frac{1}{2}\frac{-U''}{U'}\pi(m-\pi)$$

O lado esquerdo da equação acima representa a diferença de utilidade entre estar assegurado e não-assegurado, ponderada pela utilidade marginal de uma unidade monetária para remover o risco. O lado direito é o benefício da remoção de risco. Na equação, o $(-\frac{U''}{U'})$ é o coeficiente de aversão ao risco absoluto, pode ser visto como o grau que a incerteza sobre a utilidade marginal afeta negativamente a pessoa. Por $U'' < 0$ e $U' > 0$, o termo é positivo. O termo $\pi(m-\pi)$ representa a extensão com que a renda após o gasto médico altera-se pelo fato de a pessoa não ter seguro. Este termo

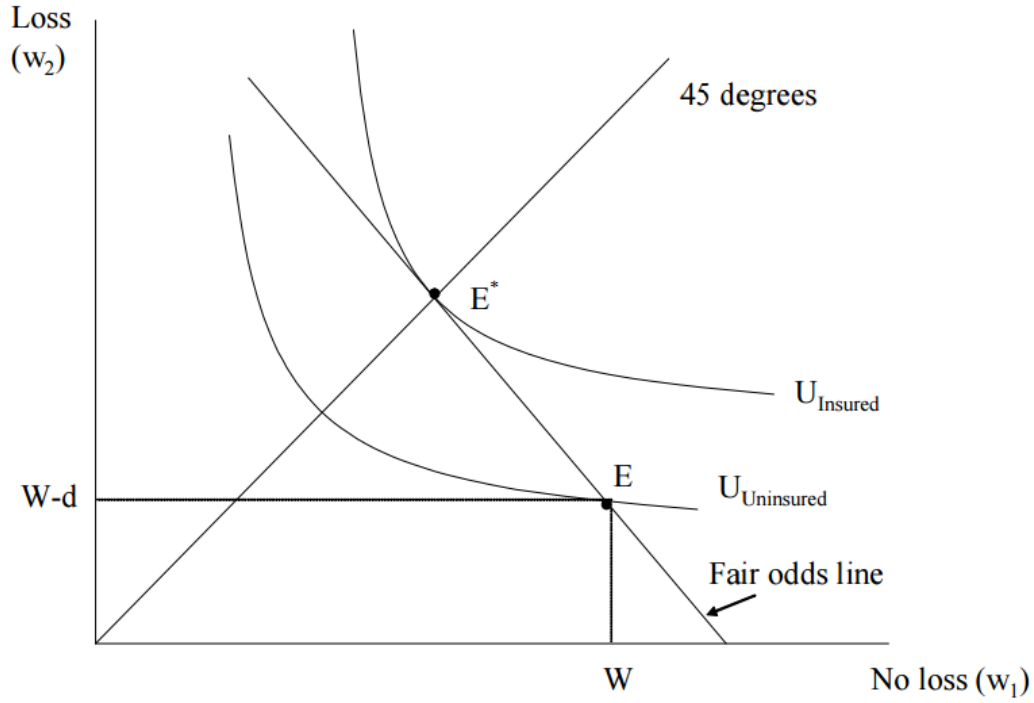
[7] Para demonstração, ver Cutler e Zeckhauser (2000).

também é positivo. O produto dos termo do lado direito da equação é, portanto, também positivo. A interpretação é que um seguro com prêmio justo é preferível a não estar assegurado. O valor atribuído ao valor unitário adicional despendido no compartilhamento de risco aumenta com a aversão de risco e com a variabilidade do gasto médico, isto é, se o agente for avesso ao risco e puder incorrer num custo médico muito alto, o valor atribuído ao gasto com a eliminação do risco é maior.

Intuitivamente, os agentes avessos ao risco gostariam de suavizar a utilidade marginal da renda, com outras palavras, os agentes preferem transferir renda de quando a utilidade marginal é baixa para quando a utilidade marginal é alta. Conforme visto acima, a utilidade marginal da renda é decrescente, portanto, quando o agente não tem gastos com tratamento médico $U(y)$, a utilidade marginal de sua renda é mais baixa do que quando está doente $U'(y-m)$ e sua renda é igual a $y-m$. Transferir renda de períodos em que está saudável para períodos onde está doente até o ponto em que a utilidade marginal é equalizada maximiza a utilidade total, assumindo que o prêmio seja justo. O seguro conduz essas transferências ao cobrar um prêmio em adiantamento e reembolsar os gastos médicos mais tarde.

A maneira gráfica de expor isto é mostrada na figura 2. Pensa-se em dois contextos – saudável (*no-loss*) e não saudável (*loss*) – como se eles fossem dois bens. Indivíduos gostariam de ter mais consumo em cada estado. Na ausência da possibilidade de ficar doente, ou seja, $\rho = 0$, os agentes seriam capazes de consumir y em cada estado. Todavia, devido ao gasto médico, os agentes poderiam consumir $y-m$ (representado por $W-d$ na figura) quando doentes. Isto é demonstrado no ponto E da figura 2.

Figura 2 – Utilidade sob Incerteza



Fonte: Autor, David (2010) - Lecture Note, MIT.

A linha de *fair odds insurance* é a restrição orçamentária na forma implícita do indivíduo. A inclinação da linha é dada por $-\frac{1}{\rho}$, a curva de indiferença de consumo também tem sua inclinação afetada pela probabilidade de ficar doente, uma vez que quanto menor for a probabilidade, menor será a propensão a abdicar de consumo quando saudável para usufruir quando doente. Os agentes podem trocar o consumo de quando estão doentes pelo consumo de quando estão saudáveis a uma taxa dada pelo prêmio do seguro.

Uma pessoa pode trocar consumo quando está doente por consumo quando está saudável, a taxa de troca é dada pelo prêmio cobrado pelo seguro. Os indivíduos decidirão comprar algum nível de seguro. Se o preço do seguro tem um preço atuarialmente justo, os indivíduos decidirão ficar completamente assegurados – pois terão o mesmo consumo independentemente da situação. Arrow (1963) sugere que sob o prisma do indivíduo – visto que a pessoa tem preferência estrita pelo seguro cujo preço é atuarialmente justo a assumir os riscos sozinha –, ele ainda preferirá uma apólice com preço atuarial injusto, dado que não seja muito injusto. O ponto ótimo é mostrado em E^*, os agentes estão numa curva de indiferença mais elevada. Numa versão de mundo simplificado – sem complicações oriundas de assimetria de informação – este tipo de apólice é eficiente, visto que a quantia paga é igual ao custo

do tratamento apropriado para a doença do indivíduo. Como supôs-se que cada doença tem um custo fixo, não há possibilidade da pessoa consumir nem mais nem menos e não há, portanto, recursos desperdiçados.

2.4 Assimetrias de informação na saúde

Na área da saúde, assim como diversas outras áreas da economia, há problemas de assimetria de informação, que leva os agentes a não tomarem a melhor decisão devido à falta de informação completa. Essa assimetria manifesta-se no comportamento que um tomador de seguros adotará ao estar segurado – e a operadora de seguros não consegue distingui-los para cobrar mais deste beneficiário. Outra maneira em que se observa esse tipo de assimetria é na ida ao médico, que possui mais informações, e acaba determinando o que o paciente – que em muitos casos não possui informação alguma sobre o diagnóstico – fará. Estes são alguns exemplos de assimetria de informação, sendo o primeiro caso um problema de risco moral, ao passo que o segundo é um problema de agente-principal. Estes conceitos serão explicados neste capítulo, assim como suas consequências.

2.4.1 Risco moral e consequências

Risco moral refere-se à possível má-conduta de um indivíduo agindo de certa maneira com os recursos de outros, que é distinta do modo que agiria caso esses recursos fossem próprios. É esperado que este agente utilize os recursos de maneira não parcimoniosa e que esteja propenso a tomar riscos que não tomaria caso contrário. Na saúde, por exemplo, ele usará muito mais recursos médicos do que usaria se tivesse que pagar por si mesmo. Tendo em vista que o seguro é um arranjo onde todos os participantes arcam com as despesas/perdas de alguns, isso cria um risco moral de usar mais recursos e tornar, por consequência, o prêmio do seguro mais elevado. Apesar da conotação do termo sugerir que isso é uma falha moral, isso não é o verdadeiro significado. Arrow (1985) emprega um termo mais próximo ao significado, que seria "ação oculta" (*hidden action*). Todavia, este pensamento – apesar de não ter sido estudado especificamente – está presente na literatura desde os tempos de Adam Smith:

> "Os diretores de tais companhias administram mais do dinheiro de outros do que o próprio, não é de esperar que dele cuidem com a mesma irrequieta vigilância com a qual os sócios de uma associação privada freqüentemente cuidam do seu". (Smith, Adam. A Riqueza das Nações, p.214)

Risco moral é uma preocupação porque conflita com os objetivos de divisão de risco. O seguro tem valor porque permite que os agentes transfiram a renda de quando menos se necessita para quando mais se necessita, entretanto, esta transferência não é perfeita em razão de os indivíduos aumentarem o consumo quando este é subsidiado. Este fenômeno cria um problema no desenho de contratos de seguro, uma vez que as seguradoras enfrentam o *trade off* entre o benefício de mais indivíduos compartilhando o risco contra o custo de um risco moral. O aumento da amplitude do seguro compartilha os riscos mais amplamente, entretanto, também leva a um aumento das perdas porque os indivíduos escolhem mais cobertura (risco moral).

Uma forma em que o risco moral se manifesta é na disposição a tomar um risco. As pessoas podem tomar menos cuidado de si quando estão asseguradas do que tomariam caso não estivessem. Contudo, essas informações são difíceis de observar / obter. A extensão do risco moral na saúde não é tão grande em algumas instâncias, pois algumas consequências podem afetar permanentemente a utilidade dos agentes devido a eventuais limitações que podem ser ocasionadas. Não haveria risco moral se as ações fossem as mesmas independentemente do agente estar ou não assegurado. Deste modo, é difícil de imaginar que as pessoas começariam a fumar por ter cobertura para câncer de pulmão, contudo, é provável que alguns se submetam a exames e/ou cirurgias que não o fariam caso tivessem que arcar com os custos ou parte deles. A ação de buscar cuidado médico não é oculta, apenas a motivação é. Outra forma de risco moral é o indivíduo esforçar-se menos para procurar provedores de baixo custo.

De acordo com De Meza (1983), ao tratarmos o risco moral na terminologia adotada na teoria do consumidor, este é o efeito substituição das pessoas gastando mais em saúde quando o preço deste bem é baixo, e não o efeito renda de consumidores gastando mais em cuidado médico devido ao seguro, que transfere os recursos do estado em que está saudável para quando está doente, fazendo-os ficar mais ricos quando doentes, que é similar ao apontado por Arrow (1963) em seu artigo seminal. Em síntese, ele sugere que agentes com plano de saúde consomem mais recursos médicos do que o fariam se tivessem que pagar o seu preço inteiro, destarte, os agentes avaliam o valor de

um serviço adicional abaixo do seu valor se mercado, o que reflete esses custos no restante da sociedade.

Na presença de risco moral, a política de pagamentos fixos não é mais adequada, pois os beneficiários terão custo marginal igual a zero dado o consumo de quantidades adicionais. Neste caso, é necessário coibir o uso indiscriminado de recursos. Para este fim, as companhias podem colocar uma taxa que deve ser paga conforme o gasto médico – que é denominada coparticipação –, deste modo, o custo marginal dos beneficiários será diferente de zero – como ocorre quando paga-se apenas o prêmio – e passa a ser a coparticipação. Por aumentar o preço do bem, a quantidade ótima consumida também altera-se, posto que a utilização de unidades adicionais não tem preço zero e, por consequência, o dispêndio nelas reduzirá a renda disponível para consumo de outros bens.

Para entender o caso de política ótima, considera-se um caso onde a política de pagamento fixo não é. Suponha que em vez de imaginarmos o beneficiário como doente ou saudável, o indivíduo tem um intervalo de potencial de severidade de doenças, determinado por s, cuja distribuição é dada pela função de densidade $f(s)$. O estado de saúde do beneficiário mantém-se o mesmo, $h = H[s, m]$. O s do paciente determinará o tratamento ótimo. A seguradora não consegue observar s. Desta forma, fazer uma apólice com prêmio fixo para todos não é ótima. A função de utilidade *ex ante* do agente é

$$V_I = \int U(y - \pi - co(m(s)), H[s, m(s)])f(s)\, ds$$

Onde $m(s)$ informa quanto de cuidado médico o indivíduo com condição s escolhe receber. É importante observar que $y - \pi - co(m(s))$ é a dotação do agente que está livre para consumo, desta forma, quanto maior for a $co(m(s))$, menor será a quantidade de recursos remanescentes para despender em consumo. Conforme citado anteriormente, a utilidade do agente tem relação positiva com consumo, assim como com seu estado de saúde. Deste modo, o agente precisa encontrar o arranjo ótimo para estas variáveis.

Considera-se, primeiramente, a política ótima – o montante de cuidado médico que o agente gostaria de contratar caso pudesse desenhar um contrato perfeito e eliminar o risco moral. Quando s é observável, a taxa de coparticipação depende somente de s, que

pode ser escrita, desta forma, como $co(s)$. O agente escolherá $m^*(s)$ que maximiza a utilidade

$$Max_{m(s)} \int U(y - \Pi - co(s), H[s,m]) f(s)\, ds$$

Onde $\Pi = \int \big(m(s) - co(s)\big) f(s) ds)$. A solução para o problema é dada por

$$H_m U_H = E[U_x]$$

Onde os subscritos denotam as derivadas parciais e $x = y - \Pi - co(s)$. O lado esquerdo representa o ganho de utilidade por despender uma unidade monetária a mais em cuidado médico – ou seja, é o produto do efeito do cuidado médico na saúde e o efeito da saúde na utilidade. O lado direito da equação é a esperança ponderada da utilidade marginal do consumo em diferentes estados de saúde, dado por

$$E[U_x] = \int U_x(y - \Pi - co(s), H[s,m])\, f(s)\, ds$$

A expressão $H_m U_H = E[U_x]$ sugere que com a melhor política, a utilidade marginal esperada oriunda do gasto adicional de uma unidade monetária iguala-se ao custo de utilidade causado pela perda de um dólar. Supondo que o ótimo local é o ótimo global.

Ao saber a função de utilidade dos agentes e os parâmetros que determinam as elasticidades de gasto médico, pode-se combiná-las para desenhar o contrato ótimo – a apólice com preço atuarial justo que maximiza a utilidade esperada dada a restrição que os agentes agirão de maneira maximizadora (e que o risco moral vai ocorrer). Essa política é inerentemente a segunda-melhor, pois ao calibrar o nível de generosidade, equilibra-se com os benefícios de maior divisão de risco entre os agentes contra os custos incorridos devido ao risco moral. O objetivo é encontrar o compartilhamento de risco $co(m)$ que maximiza a utilidade esperada.

A seguradora buscará encontrar a função $co^*(m^\#)$ que produz a maior utilidade esperada

$$E[U^*] = Max_{c(m^\#)} \int U(Y - \pi - co^*(m^\#),\ H[s,m^\#])\ f(s) ds$$

Onde $m^\#$ é a solução ótima da seguinte equação $Max_{m(s)} U(y - \Pi - co(m), H[s,m]) \quad \forall\, s$, ou seja, o gasto médico que maximiza a sua utilidade quando estiver doente. Visto que as seguradoras não podem determinar o estado de saúde de cada indivíduo, não é possível diferenciar pagamentos com base na severidade da doença (assim como não é permitido, visto regulação do setor).

Outra restrição para a seguradora é o fato de que a receita deve superar os gastos esperados. Desta forma,

$$\pi = \int \left[m^{\#}(s) - co\left(m^{\#}(s)\right)\right] f(s)\, ds$$

Haverá dois fatores se balanceando, o primeiro é a redução do consumo excessivo devido ao maior pagamento *out-of-pocket*[8] por cuidados médicos. Se a taxa de coparticipação é elevada em algum nível, os indivíduos naquela faixa pagarão mais por cuidados médicos, assim como as pessoas com níveis de gastos mais altos (pois a coparticipação foi elevada). Esse aumento na taxa de coparticipação aumenta a eficiência do provimento. O efeito adverso é uma perda nos benefícios de divisão de riscos, uma vez que ao fazer com que os indivíduos paguem mais *out-of-pocket*, aumenta-se o risco a qual eles estão expostos e reduz, por consequência, o seu bem-estar. A taxa de coparticipação ótima deve balancear esses dois incentivos.

Uma importante diferença entre a política ótima e o mundo real é que no último há uma taxa de coparticipação constante, ao passo que a primeira pode exigir estruturas não-lineares. Blomqvist (1997) encontrou – via dados coletados do RAND[9] – que a política de coparticipação ótima seria de 27% para gastos até US$ 1.000,00 e 5% para gastos acima de aproximadamente US$ 30.000,00.

Zeckhauser (1970) argumenta que há um *trade-off* entre optimalidade e simplicidade de coparticipação. Caso serviços ou doenças difiram em grau de risco moral, a taxa de coparticipação ótima diferirá entre elas também, isto é, há uma quantidade de doenças possíveis e o risco moral para um agente pode diferir de acordo com a doença (assim como o local, visto o problema agente-principal).

2.4.2 Problema agente-principal

Jensen e Meckling (1976) definem a relação de agência como um contrato no qual um indivíduo (o principal) solicita que um terceiro (o agente) realize algum serviço em seu nome/benefício no qual envolve a delegação da tomada de decisões para o agente. Se ambas as partes forem agentes maximizadores de utilidade, haverá boas razões para acreditar que o agente não agirá sempre no melhor interesse do principal, conforme

[8] O termo *out-of-pocket* significa gastos que o beneficiário tem maiores gastos ao utilizar o seguro, isto é, coparticipação.

[9] Ver capítulo 3.

citação de Adam Smith. No caso geral, o principal pode limitar as divergências de interesses ao estabelecer incentivos para o agente – como, por exemplo, um ônus pecuniário ao agente caso ele não tome as ações necessárias e/ou cause dano ao principal –, contudo, ao estabelecer tais incentivos, o principal deverá incorrer em custos de monitoramento. Além disso, haverá um custo oriundo da divergência entre a decisão que o principal e o agente consideram ótimas, este custo é chamado de perda residual. Por definição, tem-se que os custos de agência são definidos pela soma de (i) os custos de monitoramento do principal; (ii) a perda residual; e (iii) os gastos com ônus do agente.

No caso específico da relação paciente (principal) e médico (agente), há um problema de assimetria de informações. Arrow (1963) aponta que o mercado da saúde é caracterizado por um alto nível de incerteza, visto que talvez nem o médico, tampouco o paciente, está seguro sobre a doença existente e qual é o tratamento ótimo a ser dado. Todavia, é muito mais provável que o médico tenha mais conhecimento sobre a situação do paciente do que o paciente tem.

Devido à complexidade da saúde e medicina, a assimetria de informação entre provedor e consumidor é muito maior que na maioria dos mercados, como, por exemplo, o mercado de bens de consumo. Quanto mais informação o paciente adquirir sobre o seu estado de saúde e possibilidades de tratamento, mais improváveis são os desvios do papel dos provedores como agentes perfeitos[10].

O problema agente-principal mais citado na literatura de economia da saúde é o de demanda induzida pelo médico. McGuire (2000) sugere que essa dinâmica ocorre quando o médico influencia que a demanda do paciente por cuidados médicos seja diferente do que sua interpretação do que seria de melhor interesse para o paciente. Isto é, tem-se que o problema agente-principal ocorre na relação entre paciente (principal) e médico (agente), onde o principal espera que o agente tome as melhores decisões por ele quando estiver doente, pois é o agente quem tem poder de decisão sobre a quantidade a ser dispendida em recursos médicos, com o principal encarando os custos dessa decisão.

Demanda induzida por médico implica na atividade persuasiva de alterar a curva de demanda do paciente de acordo com o seu (agente) interesse próprio. Assim como em qualquer caso de problema de agência, o grau da indução depende da assimetria de

[10] Denomina-se agente perfeito quando o principal não detém informação alguma sobre o "bem".

informação entre o agente e principal – sendo a relação positiva, ou seja, quanto maior a assimetria, maior pode ser o grau de indução –. A propensão de ocorrência deste problema é muito maior quando a remuneração do médico é dada por *fee-for-service*[11], cujo pagamento dá-se pela quantidade produzida, visto que o incentivo a aumentar o volume de serviços para aumentar seu lucro é mais evidente. Neste tipo de pagamento, o médico não recebe incentivo algum a tratar o paciente de maneira eficiente, ou seja, consumindo o menor número de recursos possível.

O problema de agência foi tratado por Platão – ainda que não de maneira específica – em seu livro, A República, com seu trabalho em forma de diálogos socráticos[12],

> "Sócrates — Portanto, o médico, na medida em que é médico, não objetiva nem prescreve a sua própria vantagem, mas a do doente? Com efeito, reconhecemos que o médico, no sentido exato da palavra, governa o corpo e não é homem de negócios." (Platão, p. 22)

Platão não define o médico como um agente auto-interessado e maximizador de utilidade[13], mas como um agente altruísta, cujo único desiderato é prover melhores condições de saúde ao paciente.

[11] Fee-for-service é um método de remuneração em que o médico recebe um salário variável em função da quantidade de procedimentos que realiza.

[12] As obras de Platão têm o formato de diálogo, onde geralmente Sócrates dialoga com algum(uns) interlocutor(es) e representa a ideia do autor.

[13] Neste ponto, faz-se necessário argumentar que a utilidade de um agente pode não ser determinada apenas por questões objetivas (ex: consumo de bens materiais), mas também por questões subjetivas, como, por exemplo, ética.

3 Evidências na literatura

3.1 Aspectos gerais

O impacto da coparticipação nas decisões dos agentes foi amplamente estudado nos anos 70. Uma substancial parte da literatura deste período obtém a elasticidade da demanda por cuidado médico – dispêndio total – via dados de corte transversal ou dados em painel. Feldstein (1971) foi um dos primeiros a estimar elasticidade-preço usando dados em painel em nível de hospital, entretanto, o escopo ficou restrito a hospitais não orientados pelo lucro[14]. Ele identificou que o efeito de taxas de coparticipação ficou na ordem de -0,5. Os artigos subsequentes empregaram dados em nível de paciente e desenhos de experimento mais sofisticados. As elasticidades que surgiram desses estudos variavam entre -0,14, Phelps e Newhouse (1972, e -1,5, Rosett e Huang (1973). A implicação dessas elasticidades é de que o risco moral mostra-se uma força significativa.

Os autores da época comentam duas grandes dificuldades que foram enfrentadas. Primeiramente, a generosidade do seguro – tanto em nível regional ou individual – pode ser endógena. Uma cobertura mais generosa pode aumentar a utilização de serviços médicos – conforme previamente demonstrado – ou, alternativamente, regiões onde pessoas necessitam ou desejam mais cuidados médicos podem ser áreas onde os agentes demandam mais seguro. Não é possível separar estes efeitos estatisticamente sem a utilização de instrumentos (variáveis instrumentais) para a taxa de cobertura do seguro na região, todavia, estes instrumentos não estavam disponíveis para muitos autores. Em segundo lugar, muitos artigos – geralmente devido às limitações de dados – não conseguiam distinguir a taxa média e marginal de coparticipação. A maioria dos estudos relaciona o gasto com recursos médicos à coparticipação média da área. Porém, a teoria prevê que este gasto relaciona-se à coparticipação marginal. Considerando-se que a política de preço entre os contratos são não lineares, o preço médio e o preço marginal podem diferir substancialmente. Muitos críticos da época acreditavam que o uso de recursos médicos era apenas guiado pelas necessidades dos pacientes e que não era

[14] Exemplos de hospitais não maximizadores de lucro: hospitais filantrópicos e hospitais públicos.

afetado por nenhum outro fator econômico, ou seja, a demanda era completamente inelástica.

Manning (1987) utiliza dados do experimento que ficou famoso nos Estados Unidos como RAND Health Insurance Experiment[15] (HIE). O autor é um dos autores mais importantes na área de economia da saúde, seus artigos são referência tanto de método quanto de comprovação de questionamentos antigos. Neste estudo, Manning utiliza três formas para estimar as elasticidades – devido à comparabilidade com estudos anteriores-, que são:

i. Estimação da elasticidade da coparticipação pura via análise da variação da demanda considerada em eventos de cuidados médicos em vez de utilizar o dispêndio anual por paciente. Manning sugere que os beneficiários que não excederam o limite superior da sua disponibilidade de gastos (nos Estados Unidos, um dos planos tinha a característica de determinar um limite de gastos em que o beneficiário não haveria custos de coparticipação, somente após exceder este limite, haveria custos com co-pagamentos) antes de haver "out-of-pocket" (fora do bolso, ou seja, pagos via coparticipação) irão descontar o preço nominal pela probabilidade de exceder o limite (pois com essa probabilidade, o preço realmente pago em coparticipação é zero) ao tomar uma decisão de consumo marginal de consumo de serviços médicos. Nesta abordagem, Manning utiliza dados apenas de beneficiários que estão a US$ 400 abaixo do limite. Isto dá uma aproximação do efeito puro do preço caso as pessoas tratam a probabilidade de exceder os seus limites o mais próximo de zero possível.

ii. A segunda estimativa é feita através da estimação da função de utilidade indireta.

iii. A terceira estimativa origina-se de um cálculo similar aos utilizados na literatura, aqui utilizou-se taxas médias de coparticipação. A demonstração usual de um viés positivo na estimação da elasticidade ao empregar-se a taxa média de coparticipação – Newhouse et al (1980ª) – não aplica-se ao HIE devido ao equilíbrio entre planos. O tamanho do viés, se houver, depende de dois efeitos que tem sentidos distintos. Para pequenos volumes

[15] O governo dos Estados Unidos financiou um experimento desenhado para estimar a elasticidade-preço de atendimento médico. Este estudo, que foi nomeado como "Rand Health Insurance Experiment", coletou dados de aproximadamente 6.000 pessoas em seis regiões diferentes que receberam diferentes planos por um período de três a cinco ano (a alocação deu-se aleatoriamente).

de gasto, planos exibirão menor gasto do que haveria um plano de coparticipação puro com taxas de coparticipação entre 16 e 31% (porque a coparticipação efetiva é provavelmente maior); para grandes volumes despendidos que excedam significativamente o limite, o oposto será verdadeiro (devido ao fato de que a taxa marginal de coparticipação será zero, não positiva). O autor sugere que a verificação do efeito que predomina é uma questão empírica que os dados experimentais não podem resolver, contudo, ao comparar os resultados dessas três abordagens, o terceiro método gera valores que são sensivelmente menores – porém ainda próximos – aos outros dois métodos (o que sugere que o primeiro viés é predominante.)

Os resultados sugerem que não restam dúvidas que as elasticidades por bens médicos é diferente de zero e que, de fato, a resposta ao compartilhamento de custos é não-trivial. Outra constatação – que origina da primeira – é que mesmo com a redução de utilização de serviços médicos, não há, em média, nenhum efeito adverso no nível de saúde.

Um dos pontos de interesse neste artigo é a parte estatística, posto que o autor utiliza duas etapas para estimação, sendo que na primeira ele emprega a estimação de resposta binária Tobit e mínimos quadrados ordinários[16].

As variáveis de controle utilizadas são, também, importantes na análise. O autor analisa, inclusive, as respostas de diferentes grupos (que podem determinar as variáveis a serem usadas como variáveis de controle):

a. Idade: o autor observou diferenças estatisticamente significativas na resposta do nível de coparticipação de acordo com a idade, sendo as fases iniciais (crianças) as menos sensíveis às mudanças.
b. Renda: neste caso, o autor não encontrou diferenças significativas entre a resposta dos níveis de coparticipação no nível de utilização dos grupos de renda inferior ao dos grupos de renda superior. Houve somente diferença significativa para os agentes de baixa renda e que são doentes crônicos.
c. Tempo no plano: o autor observa que planos sem coparticipação levaram a elevados níveis de utilização (embora transitórios) nos primeiros períodos, ao passo que planos com coparticipações muito elevadas levaram a uma defasagem

[16] Estes modelos serão detalhados no capítulo 4.

da demanda (ou seja, os beneficiários ficaram com demanda reprimida) por serviços médicos. Isto sugere que agentes que não tiveram planos de saúde terão excesso de demanda nos primeiros meses de contrato caso este não regule via coparticipação.

Manning (1985) busca – em mais um artigo da série RAND – entender as diferenças entre os resultados obtidos pelos incentivos gerados pelas diferentes modalidades de planos de saúde, sendo uma delas a (i) modalidade pré-paga, onde o beneficiário paga uma mensalidade e não arca com nenhum custo extra, todavia, ele está restrito a um atendimento num núcleo específico de atendimento (ou seja, a seguradora possui um núcleo de saúde e os beneficiários só podem consultar-se lá e a internação/ida ao hospital é condicionada ao parecer do clínico geral deste núcleo da seguradora), e (ii) a outra a *fee-for-service*. Para isolar a relação entre pré-pagamento e o uso de serviços, conduziu-se um experimento com 3095 pessoas que buscava responder duas perguntas:

i. Quando agentes que recebiam assistência via planos com taxa por serviço são aleatoriamente designados a receber tratamento médico via planos pré-pagos. Como o uso deles diferencia-se dos similares que permaneceram no plano anterior?

ii. Quando agentes que recebiam assistência via planos com taxa por serviço são aleatoriamente designados a receber tratamento médico via planos pré-pagos. Como o uso deles diferencia-se dos agentes que já estavam em planos pré-pagos?

O autor aponta que a diferença entre a taxa de admissão hospitalar – cujos gastos representam em torno de 50% do montante total gasto em saúde nos Estados Unidos – entre os dois tipos de plano foi da ordem de 40%, embora o nível de visitas ambulatoriais apresentou-se similar. Visitas ambulatoriais são idas a hospitais que não requerem internação hospitalar, ou seja, é diagnosticado e tem um tratamento recomendado que não necessita hospitalização. Pode-se atribuir a menor taxa de admissão hospitalar do plano de saúde com atendimento via núcleo ao fato de os médicos dos planos de saúde darem o parecer sobre a necessidade de internar ou não. Com isso, evitam-se algumas internações desnecessárias e/ou que só ocorreriam devido aos incentivos financeiros de hospitais em internar algum cliente, aliado à eventual meta de internações estabelecida.

O autor aponta também que observou-se que o nível de consultas preventivas (consideradas as idas ao médico) é maior nos planos direcionados (ainda que não possa atribuir-se isto ao menor índice de hospitalização) do que nos planos *fee-for-service*.

Newhouse *et al* (1976) é o artigo do experimento RAND que traz a elasticidade preço, o método estatístico utilizado foi o de mínimos quadrados ordinários em dados de corte transversal. Os planos tiveram distintas taxas de coparticipações e as estimativas de elasticidade foram feitas a partir da comparação dos níveis de utilização nos diferentes planos. O experimento encontrou que a elasticidade-preço para os diversos tipos de custo é de -0,2 e que é estatisticamente diferente de zero, embora menor do que proposta pela literatura anterior. Os resultados obtidos são considerados padrão na literatura e pode-se dizer que a literatura econômica posterior a este estudo aceita que o modelo padrão de seguro de saúde leva a risco moral na demanda. Neste mesmo estudo, os autores estimam a elasticidade-preço somente por consultas médicas, também utilizando dados de corte transversal e aplicando mínimos quadrados ordinários. O resultado encontrado foi de -0,42.

Chandra (2010) afirma que a teoria econômica sugere que uma ferramenta para controlar custos médicos é o aumento do compartilhamento dos custos de cuidados médicos. Pari passu ao aumento do compartilhamento dos custos assistenciais, há uma redução no preço pago mensalmente. O que o autor sugere é que um maior nível de coparticipação pode, em teoria, diminuir os incentivos tanto do beneficiário quanto da operadora de saúde suplementar de comprometer-se com o risco moral.

O contexto em que a pesquisa de Chandra está inserido é o de expansão do número de pessoas cobertas por planos de saúde, principalmente considerando pessoas de baixa renda – que pode supor-se que elas possuam demanda reprimida por serviços de saúde. O impacto desses aumentos de coparticipação – a fim de reduzir o risco moral - são particularmente importantes para incluir populações de baixa renda em planos de saúde. O autor aponta que por um lado, um desenho de plano melhor elaborado pode reduzir pressões fiscais associadas à expansão do seguro, todavia, por outro lado há a possibilidade de que recipientes não sejam capazes de reduzir judiciosamente a utilização e experimentem baixos níveis de uso hospitalar devido ao alto custo de coparticipação. Para entender a magnitude dos fenômenos, o autor buscará verificar se as elasticidades-preço continuam similares às encontradas nos estudos anteriores (que datam trinta anos), pois ele supõe que houveram grandes mudanças estruturais na assistência médica que podem ter alterado a magnitude da resposta dos agentes ao aumento nos preços. Algumas mudanças citadas pelo autor são: (i) mudanças na prática médica, (ii) aumento no uso de medicamentos prescritos, (iii) o crescimento do

diagnóstico por imagem, (iv) desenvolvimento de cirurgias minimamente invasivas que requerem mais equipamentos.

Os resultados obtidos por Chandra apontam que a elasticidade da camada de pessoas pertencentes ao quartil inferior de renda são bem similares às obtidas no Health Insurance Experiment, que foi um experimento pioneiro conduzido há cerca de trinta anos. Ele aponta para o fato marcante de as elasticidades continuarem tão parecidas, apesar de que tenham havido mudanças estruturais importantes na composição dos custos médicos. Este resultado é de grande valia, visto que o autor aponta que a elasticidade-preço aparenta ser estável no longo prazo, mesmo que tenham mudanças estruturais pesadas, ou seja, pode-se supor com base na literatura que a elasticidade é estável no longo prazo. Como não há dados de coparticipação disponível para os períodos anteriores (somente o de uso), faz-se necessário partir dessa premissa.

3.2 Modelagem com problema agente-principal

Na literatura econômica, a estimação de modelos de duas partes é amplamente utilizado devido à alta incidência de uso zero, o que pode violar a hipótese de que uma distribuição de probabilidade com um único parâmetro (primeiro e segundo momento descritos por apenas um parâmetro) descreve adequadamente o processo gerador de dados, assim como os dados serem governados por mais do que um processo gerador de dados.

Um argumento teórico empregado por Zweifel (1981) é de que a especificação do modelo de duas partes descreve melhor o problema do agente-principal, onde o médico (agente) determina a utilização em nome do paciente (principal) assim que o primeiro contato é feito. Manning (1981) argumenta que a decisão de receber algum cuidado é praticamente do consumidor, embora o médico influencie a decisão sobre a quantidade de cuidado. Pohlmeier e Ulrich (1995) separam a análise em contato e frequência, tal que o paciente determina se consultar-se-á ou não com o médico (análise de contato) e que fica, essencialmente, a critério do médio a intensidade do tratamento (análise de frequência).

Por outro lado, Grossman (1972) utiliza a abordagem que supõe que a demanda por serviços médicos é determinada pelo paciente. Grossman faz uma ressalva ao tipo de dados que encontram-se disponíveis, pois argumenta que para dados de “cross-section”,

onde os eventos são computados durante um determinado período de tempo e não sobre um episódio de doença. Desta forma, a estrutura de agente-principal subjacente à distinção entre usuário e não usuário poderá não ser apropriada para lidar com os dados, dado que até pessoas saudáveis serão usuárias, visto que as utilizações podem ter outras origens, como, por exemplo, uso preventivo.

O autor defende que o modelo de uma parte – que não distingue usuários e não-usuários – produz uma melhor estrutura, visto que consegue distinguir grupos com demanda média alta e com demanda média baixa. Assim como argumenta que o modelo de duas partes pode fazer uma distinção entre pessoas saudáveis e não-saudáveis (apesar das proxies de status de saúde percebido), cujas demandas são caracterizadas por baixas e altas médias, respectivamente.

Deb e Trivedi (2002) utilizam os dados do RAND – que como dito anteriormente, foi um experimento para identificar como os agentes consomem serviços de saúde de acordo com os diferentes tipos de plano de saúde – para estimar o efeito de coparticipação no nível de consultas. A principal vantagem deste estudo é que os planos foram distribuídos aleatoriamente – e não escolhidos livremente pelos agentes – e os autores não precisam, portanto, lidar com os efeitos de endogeneidade do tipo de plano escolhido. Neste artigo, os autores comparam os dois tipos de modelo – duas partes e uma parte – e tentam determinar qual deles adequa-se melhor aos dados. Os autores argumentam que não é claro qual deles tem melhor desempenho a priori, dados os argumentos supracitados.

Os autores testam as diferentes abordagens e sugerem que há forte evidência em favor os modelos de uma parte, que é amparada pelos ajustes tanto de testes dentro da amostra quando por validação-cruzada (*cross validation*)[17]. Os autores argumentam que os resultados sugerem um enquadramento econométrico[18] superior do modelo de uma parte para prever demanda por saúde, medida em idas ao médico, pois é mais provável que produza melhores estimativas. Outra indicação é de que nenhuma das abordagens aceita ou descarta a estrutura de agente-principal teoricamente suposta.

[17] A validação cruzada consiste em estimar valores que não estão presentes na amostra, dado o conjunto de variáveis explicativas. Este tipo de validação é uma medida para lidar com um possível overfitting.

[18] Os autores empregam o termo "*econometric framework"*, ou seja, que a estrutura do modelo leva a resultados, em termos de previsão, superiores.

4 Métodos estatísticos

Neste capítulo, abordar-se-ão os métodos candidatos para determinar o modelo[19] a ser empregado. Far-se-á, ainda, uma rotina para encontrar a especificação do modelo. Apesar de observar que a variável resposta apresenta obliquidade positiva elevada e leptocurtose, empregar-se-á, a fins de exemplificação, uma rotina para identificação do melhor modelo (que pode acabar abrangendo outros tipos de dados relacionados à área), modelo estimado via mínimos quadrados ordinários. Empregar-se-á modelos de contagem, cuja distribuição mais comum é a Poisson. Os outros modelos surgem devido a violações de hipóteses da distribuição Poisson, tal como não satisfazer a propriedade de equidispersão – ou seja, média e variância são iguais –. Com isto, faz-se necessário empregar modelos que lidem com este problema, tanto assumindo apenas uma etapa de estimação (negativo binomial) e em duas etapas (*hurdle* e zero inflado), cuja idéia central é a de dois processos geradores de dados, sendo que o primeiro atribui um processo gerador para os zeros e outro para os inteiros positivos não-nulos, ao passo que o segundo atribui um processo gerador de zeros – que na maior parte da literatura são zeros estruturais – e outro processo gerador de dados para os inteiros positivos considerando o zero.

4.1 Mínimos quadrados ordinários

Considere o seguinte modelo:

$$y = X\beta + u$$

Tal que contenha um componente sistemático ($X\beta$) e um componente estocástico (u).

Visto matricialmente como

[19] Até agora, a palavra modelo foi tratada na sua verdadeira acepção, que é a simplificação de um fenômeno – que por muitas vezes é complexo e as relações são caóticas, o que torna praticamente impossível uma representação perfeita deste – todavia, a partir deste momento, considera-se modelo o método empregado para estimar as relações estipuladas pelo modelo. Essa nomenclatura está em consonância com a empregada em livros e artigos referência da área. Destarte, ao dizer que compara-se o modelo MQO com o modelo Poisson, deve-se ler que compara-se o modelo estimado por MQO com o modelo estimado por máxima verossimilhança que usa a distribuição Poisson como função de ligação, ambos empregando a mesma relação suposta.

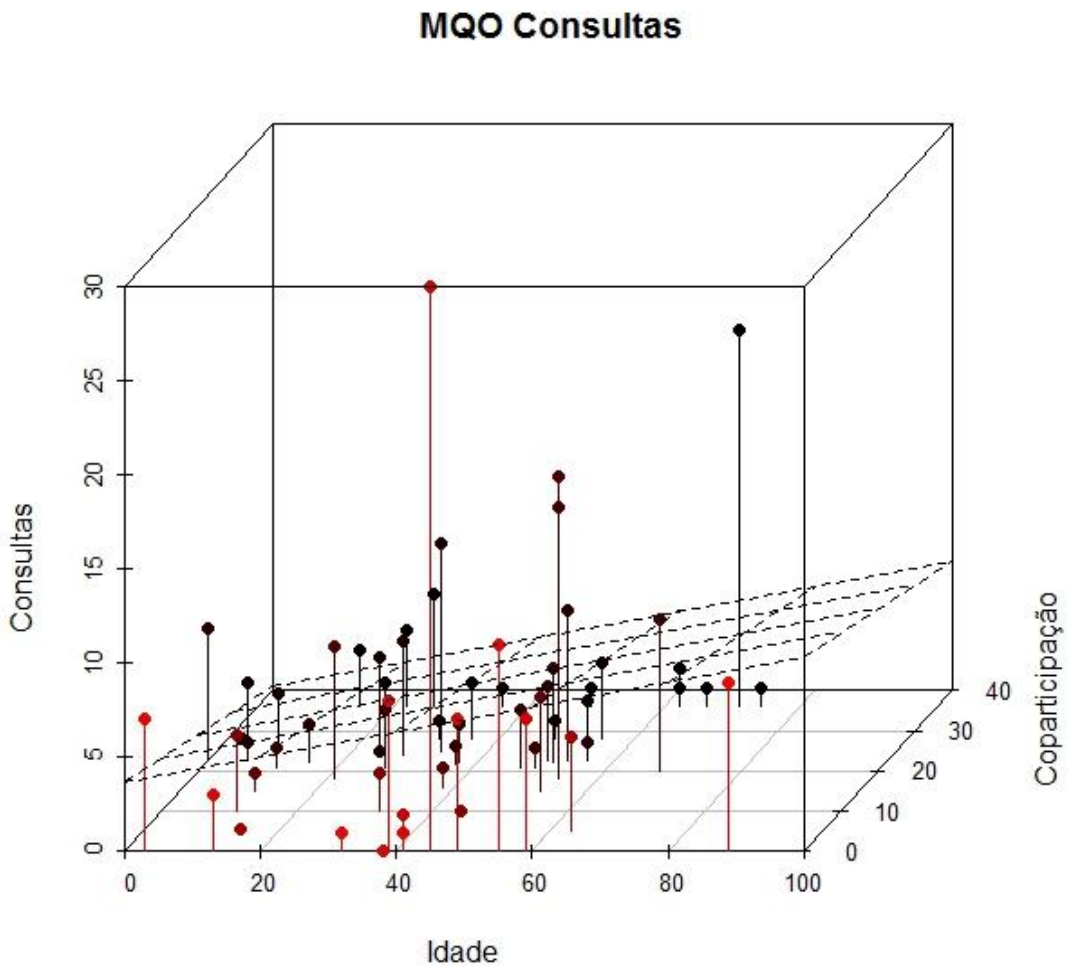
$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}_{n\times(k+1)} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{(n-1)} \end{bmatrix}_{(k+1)\times 1} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix}_{nx1}$$

Onde X é uma matriz $n \times (k+1)$, tal que n é a quantidade de observações e k é o número de variáveis independentes para as n observações. Neste trabalho, empregar-se-ão apenas modelos com um termo constante, que é adicionado na primeira coluna da matriz X e será representado pelo parâmetro β_0. O y é um vetor $n \times 1$ que contém as observações da variável dependente. O u é um vetor $n \times 1$ que contém os erros. O β é um vetor $(k+1) \times 1$ que contém os parâmetros populacionais que deseja-se estimar.

Para determinar os parâmetros, atribui-se uma função de perda (na língua inglesa, *loss function*), que sugere o quão preocupado se está com os resíduos. O método de mínimos quadrados ordinários utiliza a forma quadrática dos resíduos para encontrar os estimadores[20] e busca a solução dos estimadores em que a soma do quadrado dos resíduos seja a menor possível (de maneira simples, diz-se que estima-se os parâmetros da forma em que minimize-se o quadrado da distância vertical entre as estimativas da regressão e as observações). Na figura abaixo é possível verificar isto graficamente.

[20] Os estimadores serão denominados com chapéu, como, por exemplo, β sendo o parâmetro populacional e $\hat{\beta}$ sendo o estimador do parâmetro populacional.

Figura 3 – Plano do MQO: $Consultas = \beta_0 + \beta_1 Idade + \beta_2 Coparticipação$



Os pontos são as observações de uma amostra de tamanho 60, a altura deles determina o número de consultas realizados no ano. Empregou-se uma linha para ficar visível a posição da observação na coparticipação. O plano presente na figura representa os pontos estimados pela regressão presente no título. A distância vertical entre o ponto e o plano representa o resíduo daquela observação. Elaboração Própria.

Os resíduos são definidos por

$$u = y - X\hat{\beta}$$

A soma do quadrado dos resíduos (SQR) é dada por $u'u$

$$[u_1 \quad u_2 \quad \cdots \quad \cdots \quad u_n]_{1\times n} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_n \end{bmatrix}_{n\times 1} = [u_1 \times u_1 + u_2 \times u_2 + \cdots + u_n \times u_n]_{1x1}$$

Uma tautologia para descrever os resíduos é

$$u'u = \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)$$

$$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$[21]

Para achar os estimadores de mínimos quadrados ordinários é necessário tirar a primeira derivada da equação acima com relação a $\hat{\beta}$, o que fornece a seguinte equação

$$\frac{du'u}{d\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

Para verificar que a solução encontrada é um mínimo, faz-se necessário tomar a segunda derivada com relação à $\hat{\beta}$ – que resulta em $2X'X$. Enquanto X for posto completa, essa matriz é definida e positiva e é, portanto, mínimo[22].

Da condição de primeira ordem, obtêm-se

$$(X'X)\hat{\beta} = X'y.$$

Duas constatações podem ser feitas a partir da matriz $X'X$, uma delas é que o resultado sempre será uma matriz quadrada, pois é $k \times k$. A outra é que ela é sempre simétrica.

Se o inverso de $X'X$ existe[23], multiplicar ambos lados das equações acimas fornece-nos

$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y$$

Visto que, por definição, $(X'X)^{-1}(X'X) = I$, onde I é a matriz identidade $k \times k$, tem-se que

$$I\hat{\beta} = (X'X)^{-1}X'y$$
$$\hat{\beta} = (X'X)^{-1}X'y$$

4.1.1 Hipóteses

O modelo de regressão linear precisa satisfazer algumas hipóteses para a determinação de não viés dos estimadores, tais que:

4.1.1.1 Linearidade nos parâmetros

[21] A solução baseia-se no fato de que a transposta de um escalar é o escalar, como, por exemplo, $y'X\hat{\beta} = \left(y'X\hat{\beta}\right)' = \hat{\beta}'X'y$

[22] Para maiores detalhes e demonstrações, verificar Greene (2003).

[23] Essa condição será descrita adiante, na parte das hipóteses

Isso sugere que a relação entre a variável dependente e as variáveis explicativas é linear, isto é, os estimadores são linearmente relacionados à variável explicada.

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u_i$$

Ou, em forma matricial,

$$y_i = X_i \beta + u_i$$

Um modelo é linear nos parâmetros se as condições de primeira ordem associadas ao problema de minimização do quadrado dos resíduos gerarem um sistema linear nos parâmetros.

Exemplos de modelos lineares nos parâmetros são (i) $y = \beta_0 + \beta_1 \cos(x_1) + u$; (ii) $y = x^{\beta} e^{u}$, que após tomar o log em ambos os lados tem-se $\ln y = \beta \ln x + u$. Nos exemplos, evidencia-se que a linearidade refere-se à maneira na qual os parâmetros e o termo de erro aleatório entram na equação.

O modelo linear não conseguiria, por exemplo, estimar a seguinte equação $y = \beta_0 + \frac{1}{\beta_1 + \beta_2} + u$, visto que não há transformação possível para torná-la linear nos parâmetros.

4.1.1.2 Colinearidade não perfeita – Condição de Posto

Essa hipótese assume que não há dependência linear entre as variáveis explicativas, diz-se que um conjunto de vetores é linearmente dependente se algum desses vetores do conjunto pode ser escrito como combinação linear dos outros. Ela também é conhecida como a condição de identificação

Na amostra (população) nenhuma das variáveis independentes é constante e não há relações lineares exatas entre as variáveis independentes. A primeira parte supõe que as variáveis independentes tem variância diferente de zero, ao passo que a segunda parte relaciona-se à condição de posto, pois se as variáveis forem combinações lineares perfeitas, a matriz não será invertível e não poder-se-á calcular os estimadores.

Tem-se a matriz X das variáveis independentes, a hipótese de inexistência de multicolinearidade perfeita implica que $posto(X) = k + 1$, pois $n \geq k + 1$,

intuitivamente, pode-se dizer que para estimar $k+1$ parâmetros, é necessário que tenha-se no mínimo $k+1$[24] observações.

4.1.1.3 Média condicional zero

Assume-se que o termo de erro tenha média condicional igual a zero em toda observação – ou seja, quaisquer valores assumidos pelas variáveis independentes–, que pode ser escrito como

$$E[u_i|x_{1i}, x_{2i}, \dots, x_{ki}] = e[u_i|X_i] = 0$$

Desta forma, pode-se dizer que nenhuma observação de $\boldsymbol{x}$ transmite informações sobre o valor esperado do erro, ou seja, assume-se que não há nenhuma informação sobre $E[u_i|]$ contida em nenhuma observação $\boldsymbol{x}_i$ e o erro é, portanto, composto apenas valores aleatórios retirados de uma população.

A hipótese de média condicional zero implica que a média não-condicional também seja zero, pois

$$E[u_i] = E_{\boldsymbol{x}}\big[E[u_i|X_i]\big] = E_x[0] = 0$$

Já que para cada u_i, a $Cov[E[u_i|X],\ X] = Cov[u_i, X]$, a hipótese de média condicional zero implica que $Cov[u_i,\ X] = 0$ para todas as observações.

Greene (2003) considera que a hipótese de média condicional zero não é restritiva na maioria dos casos – os que usam modelo com intercepto –, o autor exemplifica que num modelo de duas variáveis – sendo uma delas a constante – e que a média de u_i é $\psi \neq 0$. Então, $\beta_0 + \beta_1 x + u$ é o mesmo que $(\beta_0 + \psi) + \beta_1 x + (u - \psi)$. Considerando $\beta_0' = \beta_0 + \psi$ e $u' = u - \psi$ produz o modelo original. Caso o modelo original não contenha um termo constante e $E(u_i)$ pode ser expresso como uma função linear de X_i, a transformação do modelo produzirá resíduos com média zero. Caso contrário, a média diferente de zero dos resíduos será uma parte substantiva da estrutura do modelo, o que sugere um potencial problema em modelos sem intercepto. O autor sugere que deve-se – como regra geral – utilizar o intercepto, exceto em casos em que a teoria subjacente determine que não se deva.

Wooldridge (2005) demonstra que sob essas três hipóteses supracitadas os estimadores não serão viesados. Um estimador $\widehat{\beta_i}$ de β_i será não viesado se

[24] A partir deste ponto, por fins de facilidade de notação, denotar-se-á $k = k + 1$. Isto é, sempre inclui-se o intercepto.

$$E\left[\widehat{\beta}_i\right] = \beta_i, \qquad \forall\, i = 1, 2, \dots, k$$

Para todos os possíveis valores de β_i. Se o estimador for não-viesado, sua distribuição de probabilidade terá um valor esperado igual ao parâmetro que ele estará a estimar. A inexistência de viés não significa que a estimativa obtida será igual ao parâmetro, ou mesmo muito aproximada, mas que caso fosse possível extrair amostras indefinidas da população, calcular os estimadores para cada uma dessas amostras e calcular a média das estimativas de todas as amostras aleatórias, então obter-se-ia β_i (o parâmetro).

4.1.1.4 Homocedaticidade

O erro u tem a mesma variância dada a quaisquer valores das variáveis explicativas, ou seja,

$$Var[u_i|x_{1i}, \dots, x_{ki}] = Var[u_i|X_i] = \sigma^2, \qquad \forall\, i = 1, \dots, n$$

Isso sugere que independentemente dos valores assumidos pelas variáveis explicativas, a dispersão do erro será idêntica. Ao atender a hipótese de variância constante – e ausência de autocorrelação[25][26], embora esta não seja importante para problemas de corte transversal – a matriz de variância-covariância dos resíduos será uma matriz de bloco diagonal[27] (do termo inglês "block diagonal matrix") cujos valores serão a variância do resíduo e as covariâncias (fora da diagonal) serão iguais a zero. Desta forma,

$$\begin{bmatrix} E[u_1u_1|X] & \cdots & E[u_1u_n|X] \\ \vdots & \ddots & \vdots \\ E[u_nu_1 & \cdots & E[u_nu_n|X] \end{bmatrix}$$

Dado que a covariância entre os resíduos é igual a zero – devido à hipótese de autocorrelação – tem-se que:

$$E[uu'|X] = \sigma^2 I$$

É importante notar que a heterocedasticidade não altera as propriedades de consistência dos parâmetros. Quando a hipótese de homocedasticidade não é satisfeita, a perda de eficiência do modelo de mínimos quadrados ordinários pode ser substancial –

[25] A hipótese de ausência de autocorrelação é mais evidente para séries temporais, pois assume que os resíduos $E\left[u_iu_j = 0\right], \forall\, i \neq j$, ou seja, os resíduos do período anterior não pode influenciar o resíduo do período subsequente (o que é constantemente violado na modelagem de ativos financeiros). Entretanto, não é muito provável para dados de corte cruzado que o resíduo de uma pessoa vá influenciar o resíduo de outra.

[26] Quando as duas hipóteses são atendidas – homocedasticidade e inexistência de autocorrelação –, os resíduos são chamados de esféricos.

[27] A matriz de bloco diagonal é uma matriz simétrica cujos elementos na diagonal são matrizes simétricas de qualquer tamanho – inclusive 1x1 – e cujos elementos externos à diagonal são nulos.

conforme apontado por Breusch e Pagan (1979) – e que o viés nos erros-padrão dos estimadores pode levar a inferências inválidas (pois a estatística t-student do estimador é calculada com base nos erros-padrão).

Tome-se, por exemplo, a quantidade de consultas de beneficiários de planos de saúde como função da idade. Mesmo que considerando a idade, a quantidade de consultas dos beneficiários de idade mais elevada apresentará maior variância do que a quantidade de consultas dos beneficiários mais jovens.

As quatro hipóteses supracitadas são conhecidas como as Hipóteses de Gauss-Markov para corte transversal. Sob elas o estimador de MQO é o melhor estimador linear não viesado (BLUE, da expressão inglesa *best linear unbiased estimator*). O teorema Gauss-Markov sugere que ao satisfazer as hipóteses supracitadas, os estimadores de mínimos quadrados serão os com menor variância dentro da classe dos estimadores lineares e não-viesados.

4.1.1.5 Normalidade dos resíduos

O erro é normalmente distribuído com média zero e variância σ^2: $u \sim Normal(0, \sigma^2)$ e é independente de todas as variável explicativas $x_1, \dots, x_k$. Uma implicação desta hipótese é de que os erros observados são estatisticamente independentes e não correlacionados.

De acordo com Greene (2003), a hipótese de normalidade é geralmente vista como desnecessária para o modelo de regressão, exceto nos casos em que uma distribuição alternativa é assumida[28].

As cinco hipóteses supracitadas são conhecidas como hipóteses do modelo linear clássico. Wooldridge (2005) afirma que esse conjunto de hipóteses pode ser visto como o conjunto de hipóteses Gauss-Markov mais a hipótese de erros normalmente distribuídos. Sob essas hipóteses, os estimadores de mínimos quadrados ordinários são os estimadores não-viesados de menor variância, sugerindo que estes tem a menor variância dentre os estimadores não-viesados.

[28] O modelo linear generalizado assume uma distribuição a priori e os resíduos informam se a escolhe é adequada. Abordar-se-á este ponto ao decorrer do trabalho.

4.1.2 Testes de inferência

4.1.2.1 Heterocedasticidade Breusch-Pagan

Empregar-se-á o teste Breusch-Pagan em detrimento do teste White devido a algumas restrições deste último. De acordo com Greene (2003), o teste White é extremamente geral – apesar de ser uma virtude, traz limitações potencialmente sérias – e pode revelar erros de especificação da forma funcional[29]. Outros pontos levantados pelo autor quanto ao teste são que o teste é não construtivo – ou seja, ao rejeitar a hipótese nula, o resultado não aponta o que deve ser feito posteriormente – e que não se pode dizer muito sobre o poder[30] do teste – exceto em contexto de problemas específicos e que pode ser muito baixo contra algumas alternativas.

Breusch e Pagan (1979) projetam um teste multiplicador de Lagrange da hipótese que $\sigma_i^2 = \sigma^2 f(\alpha_0 + \alpha' \boldsymbol{z}_i)$, onde $\boldsymbol{z_i}$ é um vetor de variáveis independentes cujo primeiro elemento é unitário (dados que geram o intercepto) e que o modelo é homocedástico se $\alpha = 0$. O teste é feito via uma regressão

$$g_i = z_i' \delta + e_i$$

Onde $g_i = \frac{u_i^2}{(u'u)/n}$, desta regressão utiliza-se a soma dos quadrados explicados (SQE) desta regressão para chegar-se à estatística LM do teste, tal que

$$LM = \frac{1}{2} SQE$$

De maneira análoga, define-se $\boldsymbol{Z}$ como a matriz com todas observações $\boldsymbol{z_i}$e $\boldsymbol{g}$ como todas as observações de $g_i = \frac{u_i^2}{(u'u)/n} - 1$ [31], então

$$LM = \frac{1}{2} [\boldsymbol{g'Z(Z'Z)^{-1}Z'g}]$$

Sob a hipótese nula de homocedasticidade ($H_0: \alpha_1 = \ ... = \alpha_k = 0$), a estatística LM segue uma distribuição chi-quadrada com graus de liberdade iguais ao número de

[29] Como, por exemplo, a omissão do quadrado de uma variável explicativa. Para mais detalhes e demonstrações, ver Thursby (1982).

[30] De acordo com Greene, o poder de um teste estatístico é a probabilidade com que ele levará corretamente à rejeição da hipótese nula falsa.

[31] O fato de reduzir-se um origina-se da primeira diferença da função de log-verossimilhança com respeito ao parâmetro α. Para detalhes, ver Breusch e Pagan (1979) página 1289.

variáveis em $\boldsymbol{Z}$. Este teste tem como hipótese a normalidade dos resíduos, caso ela não seja atingida, faz-se necessário outra abordagem que seja robusta à não-normalidade.

4.1.2.2 Multicolinearidade

É importante notar que a multicolinearidade não é uma questão de existência, posto que ela sempre existe, mas uma questão de grau. Quanto mais elevado for o grau de multicolinearidade, mais próximo de zero fica o determinante de $(X'X)$, de tal forma que $(X'X)^{-1}$ aumenta consideravelmente. Como a variância do estimador é positivamente influenciada por $(X'X)^{-1}$, ela aumenta consideravelmente quando o determinante de $(X'X)$ aproxima-se de zero. Com isso, o intervalo de confiança gerado para o estimador fica muito grande e a estatística do teste t – que é dado por $t = \frac{\widehat{\beta}}{Var[\widehat{\beta}]}$ para testar se $\beta = 0$ – reduz-se, embora tenha um R^2 individual alto. Outra consequência da multicolinearidade é a eventual divergência entre o teste de significância individual e o teste de significância global (teste F), ou seja, o teste F rejeita a hipótese de que os estimadores são globalmente iguais a zero, enquanto o teste t aceita a hipótese de que o estimador é igual a zero. Um sintoma da multicolinearidade severa é o fato de a estimativa do parâmetro oscilar muito dadas pequenas alterações nos dados. Conforme Greene (2003), caso haja relação perfeita entre variáveis, os estimadores não poderão ser obtidos devido à impossibilidade de inverter a matriz $(X'X)$ e a variância é infinita.

Um teste comumente utilizado para verificar a multicolinearidade é o fator de inflação da variância (FIV), cuja forma de cálculo é

$$FIV(\widehat{\beta}_\iota) = \frac{1}{1 - R_i^2}$$

O cálculo do fator de inflação da variância origina-se da variância do estimador. A lógica subjacente ao FIV é que há um aumento da variância do estimador atribuído à não-ortogonalidade das variáveis (caso de ausência de multicolinearidade)

$$Var[\widehat{\beta}_\iota] = \frac{\sigma^2}{SQT_i}\frac{1}{(1 - R_i^2)}$$

Onde SQT é a soma dos quadrados totais – representando a variação amostral em x_i – e R_i^2 é o coeficiente de determinação da regressão de x_i contra todas as demais variáveis explicativas (incluindo um intercepto).

Belsley et al (1980) sugere que valores acima de 20 são indicativos de problema. Neste trabalho empregar-se-á apenas a abordagem de fator de inflação de variância, para um discussão mais completa, verificar Hill et al (2001).

4.1.2.3 Má-Especificação Reset

O teste RESET (da sigla da expressão inglesa "*regression specification error test*") de má-especificação da forma funcional fornece-nos um indicador de evidência de não-linearidade. O teste consiste em verificar se a inclusão da potência dos valores previstos da variável explicada pelo modelo original é significativa para explicar variações da variável independente.

A ideia subjacente ao teste é que os valores ajustados – que são, por definição, uma combinação linear das variáveis explicativas – elevados ao quadrado podem ser vistos como uma combinação linear do quadrado das variáveis explicativas e a suas interações (produtos cruzados).

Considera-se o modelo original[32]

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u_i$$

Após rodar o modelo, faz-se necessário utilizar os valores previstos, $\hat{y}$, elevá-los à segunda potência – é possível utilizar potências mais elevadas, todavia, Wooldridge (2005) aponta que não há uma resposta objetiva para definir quantas utilizar e que os trabalhos aplicados empregam a segunda a terceira potência – e inseri-los no modelo original, de tal forma que a regressão expandida assuma a seguinte forma

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \cdots + \delta_l \hat{y}^l + erro$$

As hipóteses do modelo são

[32] Se o modelo satisfizer a hipótese de média condicional zero, as funções não-lineares das variáveis independentes não devem ser significativas quando adicionadas ao modelo. Conforme Wooldridge, a hipótese pode não ser atendida quando especifica-se a forma funcional da maneira incorreta, como, por exemplo, omitindo o quadrado de uma variável na equação.

$$\begin{cases} H_0: \delta_1 = \cdots = \delta_l = 0 \\ H_1: \exists m. \delta_m \neq 0 \end{cases}$$

Coeficientes significativos podem indicar má-especificação, todavia, o modelo não aponta qual correção deve ser feita.

4.1.2.4 Estimativa robusta à heterocedasticidade dos erros-padrão

Recapitulando o ponto abordado em 4.1.1.4., a hipótese de homocedasticidade sugere que a variância residual dada por

$$Var(u) = E[uu'|X] = \sigma^2\Omega = \Sigma$$

Tenha $\Omega = I$, ou seja, a matriz identidade. A variância do estimador de mínimos quadrados ordinários é dada por

$$Var(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

Na formulação original, White (1980) sugere que $\widehat{X'\Sigma X}$ é um estimador consistente de $X'\Sigma X$ usando a matriz da amostra

$$\Sigma = diag\{\hat{u}_i^2\}$$

Onde $\hat{u}_i^2$ são os resíduos da estimativa de mínimos quadrados ordinários.

Outra abordagem padrão na literatura é a sugerida por MacKinnon e White (1985), onde ajusta-se a estimativa com graus de liberdade e tamanho da amostra, ou seja,

$$\Sigma = \frac{n}{n-k} diag\{\hat{u}_i^2\}$$

Onde n é o tamanho da amostra e k é o número de dimensões de β (número de variáveis no modelo).

4.2 Método de máxima verossimilhança

Verossimilhança foi definida por Fisher (1922), que descreve-a como

> "The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed."[33], (FISHER, p.310).

[33] "a verossimilhança com que algum parâmetro (ou um conjunto de parâmetros) deveria assumir algum valor (ou conjunto de valores) é proporcional à probabilidade que caso ele o fosse, a totalidade de observações seria essa observada", ou seja, o princípio sugere que ao escolher um estimador (ou conjunto de estimadores) que maximize a verossimilhança, será mais provável que os dados gerados sejam iguais aos observados.

Os modelos baseados em verossimilhança são aqueles cuja distribuição de probabilidade conjunta da variável dependente é especificada. Assume-se que a variável dependente y_i, dado o vetor de regressores $\boldsymbol{x}_i$ e o vetor de parâmetros θ, tem a distribuição com distribuição de probabilidade $f(y_i|\boldsymbol{x}_i, \theta)$. O princípio da verossimilhança utiliza o estimador de θ que maximiza a probabilidade conjunta de observar os valores da amostra $y_1, \dots, y_n$. Essa probabilidade – vista como uma função dos parâmetros condicionados aos dados – é denominada de função de verossimilhança, cuja denotação é

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i, \theta)$$

Uma hipótese importante é a de independência ao longo dos distintos subscritos i. Por questões computacionais e de simplicidade, a verossimilhança é geralmente tratado pelo logaritmo natural da verossimilhança. A maximização da verossimilhança é equivalente à maximização da função de log-verossimilhança e não há prejuízos em empregá-la. Ela assume a seguinte forma

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{x}_i, \theta)$$

4.2.1 Condições de regularidade

As condições de regularidade são equivalentes – em papel desempenhado – às hipóteses elencadas para a determinação das propriedades dos estimadores de mínimos quadrados ordinários. Quando as condições de regularidade são satisfeitas, os estimadores de máxima verossimilhança possuirão as propriedades de consistência e normalidade assintótica. Ademais, os estimadores de máxima verossimilhança possuem a propriedade desejável de alcançar o limite inferior de Cramer-Rao e são, desta maneira, eficientes. Crowder (1976) elenca as condições de regularidade da seguinte maneira

i. A função de distribuição de probabilidade $f(y, \boldsymbol{x}, \theta)$ é globalmente identificada e $f\left(y, \boldsymbol{x}, \theta^{(1)}\right) \neq f\left(y, \boldsymbol{x}, \theta^{(2)}\right),\ \forall\, \theta^{(1)} \neq \theta^{(2)}$
ii. $\theta \in \Theta$, onde Θ possui dimensões finitas, é fechado e compacto.

iii. Derivadas contínuas e limitadas de $\ln L(\theta)$ no mínimo até a terceira ordem.

iv. A ordem de diferenciação e integração da verossimilhança pode ser revertidas[34]

v. O vetor de regressões $\boldsymbol{x}_i$ satisfaz

   a. $\boldsymbol{x}_i'\boldsymbol{x}_i < \infty$

   b. $\frac{E[w_i^2]}{\sum E[w_i^2]} = 0, \ \forall\, i$, onde $w_i \equiv \boldsymbol{x}_i' \frac{\partial \ln f(y_i|\boldsymbol{x}_i,\theta)}{\partial \theta}$

   c. $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E[w_i^2|\Omega_{i-1}]}{\sum_{i=1}^{n} E[w_i^2]} = 1$, onde $\Omega_{i-1} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{i-1})$

Newey e McFadden (1994) comentam mais detalhadamente sobre essas condições de regularidade. A primeira condição é a condição de identificação, que assegura que o limite de $\frac{1}{n}\ln L(\theta)$ tenha apenas um único ponto de máximo. A segunda condição elimina possíveis problemas no contorno de Θ e pode ser relaxada se $\ln L(\theta)$ for globalmente côncava. A terceira condição geralmente é relaxada até a condição de segunda ordem e serve para identificar o ponto de máximo. A quarta condição é uma condição-chave que elimina distribuições de probabilidade cuja extensão de y_i depende de θ. A última condição elimina qualquer observação que faça uma contribuição exagerada para a verossimilhança.

Devido à terceira condição, considerar-se-á que o limite de $\frac{1}{n}\ln L(\theta)$ é maximizado num ponto dentro do espaço de Θ. Os estimadores de máxima verossimilhança são, desta forma, a solução das condições de primeira ordem de

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f_i}{\partial \theta} = 0$$

Onde $f_i = f(y_i|\boldsymbol{x}_i,\theta)$ e $\frac{\partial \ln L(\theta)}{\partial \theta}$ é um vetor com dimensões $q \times 1$, tal que $q = k + 1$.

A distribuição assintótica dos estimadores de máxima verossimilhança é geralmente obtida sob a hipótese de que a distribuição de probabilidade está corretamente especificada. Isto é, o processo gerador de dados de y_i tem distribuição de probabilidade $f(y_i|\boldsymbol{x}_i,\theta_0)$, onde θ_0 é o valor do parâmetro.

Sob as condições de regularidade, $\hat{\theta} \xrightarrow{p} \theta_0$[35], isto é, o estimador de máxima verossimilhança converge em probabilidade para o valor do parâmetro, destarte, o estimador é consistente. Assim como,

[34] Isto é, pode-se expressar a derivada de uma integral como a integral de uma derivada.

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N[0, \boldsymbol{A}^{-1}]$$

Isto é, o termo da esquerda converge em distribuição para o termo da direita, onde o inverso da matriz $\boldsymbol{A}$ é a variância dos estimadores. A matriz $\boldsymbol{A}$ possui dimensões $q \times q$ e é definida como

$$\boldsymbol{A} = \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{i=1}^{n} \frac{\partial^2 \ln f_i}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\right]$$ [36]

Conforme apontado em Greene (2003), uma consequência das condições de regularidade – especificamente a terceira e a quarta – é a igualdade da matriz de informação, ou seja,

$$E\left[\frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta'}\right] = -E\left[\frac{\partial \ln L(\theta)}{\partial\theta}\,\frac{\partial \ln L(\theta)}{\partial\theta'}\right]$$

para todos os valores de $\theta \in \Theta$.[37]

Valendo-se da hipótese de independência ao longo dos subscritos i e definindo

$$\boldsymbol{B} = \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{i=1}^{n} \frac{\partial \ln f_i}{\partial\theta}\frac{\partial \ln f_i}{\partial\theta'}\bigg|_{\theta_0}\right]$$

A igualdade da matriz de informação implica que $\boldsymbol{A} = \boldsymbol{B}$. Essa igualdade é verdadeira somente no ponto do valor do parâmetro (verdadeiro valor). Cameron e Trivedi (1992) sugerem que a condição de igualdade pode ser vista como uma condição para a não-existência de má-especificação[38], isto é, que a média condicional e a variância condicional são corretamente especificadas.

Quando há problema de má-especificação, a estimativa da matriz de variância-covariância dos estimadores não é consistente. Uma forma generalizada de obter uma matriz variância-covariância consistente dos estimadores de máxima verossimilhança – mesmo sob a especificação incorreta da distribuição de probabilidade da variável dependente – é conhecida como sanduíche.

---

[35] A prova de convergência em probabilidade de um estimador é demonstrada em Newey e McFadden (1994).

[36] Esta notação de derivada segue o padrão de Leibniz e sugere que é o valor da segunda derivada de $\ln f_i$ em relação à θ no ponto θ_0. Uma forma alternativa de escrever é $\frac{\partial^2 \ln f_i}{\partial\theta\partial\theta'}(\theta_0)$.

[37] Para verificar a demonstração, olhar Greene (2003), capítulo 17.

[38] Por exemplo, na regressão por mínimos quadrados ordinários, a presença de heterocedasticidade pode ser oriunda de uma má-especificação do modelo.

4.2.2 Estimador consistente - *Sandwich*

Uma das hipóteses dos estimadores de máxima verossimilhança é a especificação correta da distribuição de probabilidade da variável dependente – que quando não atendida, leva a inconsistências –, contudo, White (1982) sugere um estimador consistente à má-especificação da distribuição de probabilidade de y.

Supõe-se que assume-se a distribuição de probabilidade $f(y|\boldsymbol{x},\theta)$, que sob as condições de regularidade é consistente. Todavia, a distribuição de probabilidade escolhida não é a mesma que a distribuição de probabilidade do processo gerador de dados – dada por $f(y|\boldsymbol{z_i},\theta)$ –, neste caso, a má-especificação levará à inconsistência dos estimadores. White (1982) sugere que $\hat{\theta} \xrightarrow{p} \theta_*$, onde o valor do pseudoparâmetro θ_* é o valor que maximiza a convergência em probabilidade $plim\frac{1}{n}\sum_{i=1}^{n}\ln f(y_i|\boldsymbol{x_i},\theta)$ e o a convergência é obtida sob o processo gerador de dados $f^*(y_i|\boldsymbol{z_i},\gamma)$. Desta forma,

$$\sqrt{n}\left(\hat{\theta}-\theta_*\right) \xrightarrow{d} N[0,\boldsymbol{A}_*^{-1}\boldsymbol{B}_*\boldsymbol{A}_*^{-1}]$$

Onde

$$\boldsymbol{A} = \lim_{n\to\infty}\frac{1}{n}E_*\left[\sum_{i=1}^{n}\frac{\partial^2 \ln f_i}{\partial\theta\partial\theta'}\bigg|_{\theta_*}\right]$$

E

$$\boldsymbol{B} = \lim_{n\to\infty}\frac{1}{n}E_*\left[\sum_{i=1}^{n}\frac{\partial \ln f_i}{\partial\theta}\frac{\partial \ln f_i}{\partial\theta'}\bigg|_{\theta_*}\right]$$

Onde $f_i = f(y_i|\boldsymbol{x_i},\theta)$ e as esperanças E_* são obtidas em relação ao processo gerador de dados $f^*(y_i|\boldsymbol{z_i},\gamma)$.

O estimador *sandwich* é denominado desta maneira devido ao formato que estima-se consistentemente, ou seja, $\widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{A}}^{-1}$, cujo formato remete a um sanduíche de $\widehat{\boldsymbol{B}}$. O estimador foi elaborado no contexto de estimadores de máxima verossimilhança – cuja média condicional é inconsistente –, o estimador de máxima verossimilhança do modelo linear generalizado é visto como um caso especial dos estimadores de máxima verossimilhança. No caso do modelo linear generalizado, os estimadores podem ser consistentes mesmo sob má-especificação da distribuição de probabilidade, já que a consistência neste caso é essencialmente a exigência de média condicional seja corretamente especificada, isto é, $E[\eta_i|\boldsymbol{x_i}] = \boldsymbol{x_i'}\hat{\beta}$ e as consequências dessa má-

especificação são problemas na inferência dos erros-padrão dos estimadores e nas estatísticas t-student destes.

4.2.3 Inferência

Há diversos testes para avaliar o quão bem os modelos se ajustam aos dados. As estatísticas deles indicam o nível de verossimilhança do modelo, isto é, o quão bem maximiza-se a função de verossimilhança. Essas medidas são:

4.2.3.1 Critério de informação de Akaike

O critério de informação de Akaike (AIC, da sigla inglesa) é uma medida de ajuste, ela emprega a log-verossimilhança, todavia, ela adiciona termos que penalizam a estatística de acordo com o número de variáveis no modelo. Desta forma, ele tenta balancear a qualidade do ajuste e a introdução de variáveis. É computado da seguinte maneira

$$AIC = -2\ln L(\theta) + 2l$$

Onde l é o número de parâmetros adicionados ao modelo – incluindo o parâmetro de dispersão, quando houver – e $\ln L(\theta)$ é a log-verossimilhança do modelo. Dentro de um conjunto de modelos a escolher, o preferido é o com menor valor de AIC, já que o modelo recompensa a qualidade de ajuste (dada pela função de verossimilhança).

4.2.3.2 Critério de informação Bayesiano

O critério de informação Bayesiano (BIC, da sigla inglesa) é similar ao AIC, todavia, inclui-se o tamanho da amostra no cálculo[39]. O cálculo da estatística dá-se por

$$BIC = -2\ln L(\theta) + l\ln n$$

Onde n é o tamanho da amostra. A interpretação dada é similar à do AIC.

[39] A motivação da adição do tamanho da amostra é oriunda de hipóteses feitas por Schwarz (1978).

4.2.3.3 Pearson chi-square

Uma medida padrão de qualidade de ajuste é a Pearson Statistic, cujo cálculo é uma soma ponderada dos resíduos, calculado da seguinte maneira

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\omega}_i}$$

Onde $\hat{\mu}_i$ é o estimador da média condicional e $\hat{\omega}_i$ é o estimador da variância condicional. A intuição do teste é que se a média e variância estiverem corretamente especificada, então $E\left[\sum_{i=1}^{n} {(y_i - \mu_i)^2}/{\omega_i}\right] = n$, pois $E\left[{(y_i - \mu_i)^2}/{\omega_i}\right] = 1$. Os valores mais próximos ao tamanho da amostra são os modelos que apresentam melhor ajuste.

4.2.4 Testes

Conforme Engle (1984), os três testes clássicos para testes de hipóteses no contexto de máxima verossimilhança, que são: (i) razão da verossimilhança (TLR), (ii) teste de Wald (TW) e (iii) Multiplicador Lagrange (TLM)[40]. Os testes utilizam as mesmas hipóteses, cuja nula é

$$H_0: \boldsymbol{r}(\theta) = 0^{41}$$

Onde $\boldsymbol{r}$ é um vetor com dimensões $h \times 1$ de possíveis restrições não-lineares em θ, tal que $h \leq q$, onde $q = k + 1$. A hipótese alternativa é

$$H_1: \boldsymbol{r}(\theta) \neq 0.$$

Engle (1984) sugere que a diferença essencial entre os testes é a abordagem. De maneira geral, a abordagem do teste LM começa no modelo nulo e responde se um movimento em direção ao alternativo seria um ganho, ao passo que o teste de Wald começa no alternativo e considera o movimento em direção ao nulo. O TLR compara as duas hipóteses diretamente.

[40] O teste LM não será abordado neste trabalho, pois a implementação do TLR e TW são computacionalmente mais simples – visto que não necessitam usar o vetor sore, isto é, as primeiras derivadas da log-verossimilhança sobre θ – e produzem resultados similares.

[41] As restrições podem assumir diferentes formas, isto é, elas podem ser tanto lineares, como, por exemplo, $\theta_1 + \theta_2 = 1$, quanto não-lineares, como, por exemplo, $\theta_3\theta_4 = 1$.

Sob as condições de regularidade, os testes são distribuídos assintoticamente como uma chi-quadrado χ^2 com h – que é o número de restrições – graus de liberdade quando a hipótese nula é verdadeira, denotado de maneira simplificada como $\chi^2(h)$. A hipótese nula é rejeitada quando a estatística do teste exceder $\chi^2(h;\alpha)$, onde α é o nível de significância adotado.

4.2.4.1 Razão da verossimilhança

No contexto de regressões, o TLR é calculado ao estimar dois modelos e comparar o ajuste dos modelos distintos, tal que um seja a versão restrita (com menos variáveis) e o outro a versão não-restrita (com mais variáveis). A remoção de variáveis independentes geralmente leva a um pior ajuste do modelo aos dados (isto é, a verossimilhança será menor), todavia, é necessário verificar se a diferença observada é estatisticamente significativa ou não. O TLR faz isto ao comparar a verossimilhança dos dois modelos, caso essa diferença seja significativa, a versão não-restrita ajusta-se melhor aos dados do que a versão restrita. A estatística do teste é calculada

$$T_{LR} = -2\left[\ln L\left(\hat{\theta}_R\right) - \ln L\left(\hat{\theta}_{\tilde{\mathrm{N}}}\right)\right]$$

Onde $L(\theta_R)$ é a função de verossimilhança do modelo restrito e a $L(\theta_{\tilde{\mathrm{N}}})$ é a do modelo não-restrito.

A intuição do TLR é que se a hipótese nula for verdadeira, o máximo da função de verossimilhança do modelo restrito e não-restrito deveriam ser o mesmo e a estatística do teste seria aproximadamente zero.

4.2.4.2 Teste de Wald

A estatística do teste de Wald é calculada por

$$T_W = \boldsymbol{r}\left(\hat{\theta}_{\tilde{\mathrm{N}}}\right)' \left\{ \left.\frac{\partial \boldsymbol{r}(\theta)'}{\partial \theta}\right|_{\hat{\theta}_{\tilde{\mathrm{N}}}} \left[\frac{1}{n}\widehat{\boldsymbol{A}}\left(\hat{\theta}_{\tilde{\mathrm{N}}}\right)^{-1}\right] \left.\frac{\partial \boldsymbol{r}(\theta)}{\partial \theta'}\right|_{\hat{\theta}_{\tilde{\mathrm{N}}}} \right\}^{-1} \boldsymbol{r}\left(\hat{\theta}_{\tilde{\mathrm{N}}}\right),$$

Onde $\widehat{\boldsymbol{A}}(\hat{\theta}_{\tilde{\mathrm{N}}})$ é um estimador consistente da matriz variância-covariância com o valor estimado no modelo não-restrito.

A vantagem do teste de Wald é que ele necessita a estimação apenas de um modelo, que é o não-restrito. Se a hipótese nula for aceita, isso sugere que a remoção de

variáveis do modelo não prejudicará substancialmente o ajuste do modelo, já que um preditor com um coeficiente que é muito pequeno em relação ao seu erro-padrão não estará, no geral, ajudando muito na previsão da variável dependente.

4.3 Modelo linear generalizado

No seu artigo, Nelder e Wedderburn (1972) notaram que muitos modelos de regressão linear que eram tidos como padrão na estatística eram membros de uma família e que poderiam ser tratado da mesma maneira. McCullagh e Nelder (1989) traçam uma breve cronologia, expor-se-á ela de maneira sucinta alguns exeplos: (i) uma distribuição binomial com ligação probit, empregada por Bliss (1935); (ii) uma distribuição binomial com ligação logit, empregada por Berkson (1944); (iii) modelo log-linear de contagem de dados (distribuição Poisson com ligação log), empregado por Birch (1963).

É sabido desde o tempo de Fisher (1934) que muitas distribuições comumente utilizadas eram pertencentes a uma mesma família, cuja denominação dada por Fisher foi de família exponencial. Nelder e Wedderburn (1972) unificaram estes modelos de regressão no seu artigo "Generalized Linear Models", demonstraram a possibilidade de tratar estes modelos da mesma maneira e que as estimativas de máxima verossimilhança desses modelos poderiam ser obtidas utilizando o método de mínimos quadrados ponderados iterados (*iterative weighted least squares, do termo inglês*).[42]

4.3.1 Estrutura

O modelo linear generalizado consiste em três componentes, que são:

i. Componente Aleatório:

O componente aleatório especifica a distribuição condicional da variável resposta (Y_i) (sendo i o número da observação da amostra de tamanho n) dado os valores das

[42] O método de estimação precisa ser iterativo devido ao caráter circular dessa estimação, uma vez que os valores de μ_i e de η_1 são desconhecidos e dependem dos parâmetros que deseja-se estimar. Há o emprego de outros métodos iterativos, contudo, a abordagem destes foge do escopo deste trabalho.

variáveis explicativas no modelo. Na formulação original proposta por Nelder e Wedderburn (1972), a distribuição de Y_i é pertencente à classe das distribuições exponenciais – como, por exemplo, gaussiana, binomial, gamma ou gaussiana-inversa. Todavia, trabalhos posteriores estenderam para classe das distribuições exponenciais multivariadas (como, por exemplo, a distribuição multinomial) e de algumas distribuições que não são da classe das exponenciais (tal como a binomial negativa de dois parâmetros).

ii. Preditor Linear

Uma função linear dos regressores, tal como:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} = \boldsymbol{x}'\beta$$

Que assim como no modelo linear, os regressores X_{ij} são funções previamente especificadas das variáveis explicativas e podem, portanto, incluir (i) variáveis explicativas quantitativas, (ii) transformações de variáveis explicativas quantitativas, (iii) variáveis qualitativas (dummies), (iv) iterações e assim por diante.

iii. Função de Ligação $g(.)$

A função de ligação precisa ser infinitamente diferenciável (contínua) e invertível. A função de ligação tem por objetivo transformar a esperança da variável resposta $\mu_i \equiv E(Y_i)$ para o preditor linear:

$$g(\mu_i) = \eta_i$$

Como a função de ligação é invertível, pode-se escrever analogamente:

$$\mu_i = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

Portanto, o modelo linear generalizado pode ser pensado como (i) um modelo linear para transformação da saída (response) esperada ou (ii) um modelo de regressão não-linear para a saída.

Uma propriedade conveniente da classe das exponenciais é de que a variância condicional de Y_i é uma função de sua média e, possivelmente, um parâmetro de dispersão.

4.3.2 Distribuições - ligação

O modelo linear generalizado utiliza distribuições da família das distribuições exponenciais lineares (*linear exponential family*, da expressão anglicana) – formuladas por Fisher (1934). Firth (1991) descreve a forma generalizada delas como[43]

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

Onde $f(y|\theta,\phi)$ é a função de probabilidade da variável discreta aleatória y ou a função distribuição de probabilidade para variável contínua aleatória y; $a(.), b(.), c(.)$ são funções que variam de acordo com a distribuição. A função $b(.)$ é tal que $E[y] = \mu = b'(\theta)$, onde $b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}$. A função $a(.)$ é tal que $V[y] = a(\phi)b''(\theta)$, onde $b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}$. A função $c(.)$ é uma constante de normalização. Tome-se, por exemplo, a distribuição normal, descrita como

$$f(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Ou

$$f(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\phi^2}}\, e^{-\frac{(y-\theta)^2}{2\phi^2}}$$

Através de uma manipulação algébrica, tem-se que

$$f(y|\mu,\sigma) = exp\left\{\frac{y\theta - \frac{\theta^2}{2}}{\phi} - \frac{1}{2}\left[\frac{y^2}{\phi} + \ln 2\pi\phi\right]\right\}$$

Onde $\theta = g(\mu) = \mu$; $\phi = \sigma^2$; $a(\phi) = \phi$; $b(\theta) = \frac{\theta^2}{2}$ e $c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\phi} + \ln 2\pi\phi\right]$. Firth (1991) argumenta que é útil expressar as diversas distribuições pertencentes à família das exponenciais de maneira generalizada, pois propriedades gerais da família podem ser aplicadas a casos individuais,

[43] A notação empregada por Firth (1991) é levemente distinta, entretanto, para torna-las homogêneas neste trabalho, alterou-se algumas notações para corresponderem com a utilizada por demais autores.

$$b'(\theta) = \frac{db(\theta)}{d\theta} = \mu$$

e que

$$Var[y] = a(\phi)b''(\theta) = a(\phi)\frac{d^2b(\theta)}{d\theta^2} = \alpha(\phi)v(\mu)$$

A variância condicional, $v(.)$, é vista como uma função da média condicional $v(\mu)$ e de um parâmetro de dispersão (que no caso da distribuição Poisson é zero, uma vez que ela tem a característica de equidispersão).

4.4 Regressão Poisson

Um modelo tradicional para contagem de um evento de interesse é o modelo Poisson, que sugere o uso da distribuição de probabilidade Poisson, cuja função de probabilidade, é

$$P[Y = y] = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0, 1, 2, \ldots$$

Onde μ é a intensidade. Referir-se-á à distribuição como $P[\mu]$. Os dois primeiros momentos da distribuição são

$$E[Y] = \mu$$
$$Var[Y] = \mu$$

A esperança e variância mostradas acima demonstram a propriedade de equidispersão da distribuição Poisson, ou seja, a igualdade entre variância e média.

A regressão Poisson é derivada da função de probabilidade Poisson ao parametrizar a relação entre o parâmetro μ e os regressores x. A hipótese padrão é utilizar a parametrização na forma exponencial

$$\mu_i = e^{x_i'\beta}, \qquad i = 1, \ldots, N$$

Onde N é o tamanho da amostra e existem Kregressores independentes, incluindo a constante. Dado que a distribuição é equidispersa, tem-se que $Var[Y_i|x_i] = e^{x_i'\beta}$, portanto, a regressão Poisson é intrinsecamente heterocedástica, pois a sua variância não é constante.

Partindo-se da hipótese de independência das observações $(y_i|x_i)$, a estimativa dos parâmetros dar-se-á por máxima verossimilhança. Conforme visto anteriormente, ao

maximizar a log-verossimilhança, maximiza-se, necessariamente, a máxima verossimilhança. A função log-máxima verossimilhança dá-se por

$$\ln L(\beta) = \sum_{i=1}^{N} (y_i x_i' \beta - \mathrm{e}^{\mathrm{x}_i' \beta} - \ln y!)$$

O estimador Poisson de máxima verossimilhança, denominado por $\hat{\beta}$é a solução para Kequações não-lineares correspondentes à condição de primeira ordem para máxima verossimilhança,

$$\sum_{i=1}^{N} \left(y_i - e^{x_i' \beta}\right) x_i = 0$$

A função log-verossimilhança é globalmente côncava, consequentemente, a resolução das equações leva a estimativas únicas dos parâmetros.

A matriz variância-covariância robusta à má-especificação é dada na forma *sandwich*

$$V[\hat{\beta}] = \left(\sum_{i=1}^{n} e^{x_i' \boldsymbol{\beta}} \boldsymbol{x_i x_i'}\right)^{-1} \left(\sum_{i=1}^{n} \omega_i \boldsymbol{x_i x_i'}\right) \left(\sum_{i=1}^{n} e^{x_i' \boldsymbol{\beta}} \boldsymbol{x_i x_i'}\right)^{-1}$$

Onde $\omega_i = V[y_i|\boldsymbol{x_i}]$ é a variância condicional de y_i. Caso a hipótese de especificação correta seja atendida, tem-se que $\omega_i = \mu_i$ devido à propriedade de equidispersão.

4.4.1 Interpretação

Cameron e Trivedi (1998) consideram um modelo de contagem de dados para uma variável Y que assume valores inteiros não negativos, cuja distribuição é Poisson. O modelo é escrito como

$$E[y_i|x_i] = \lambda_i = e^{x_i' \beta} = e^{\beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}}$$

Conforme Cameron e Trivedi (1998), a interpretação de modelos de contagem de dados com a distribuição Poisson difere da dada aos coeficientes de mínimos quadrados ordinários devido à exponenciação. Os autores demostram que:

$$\frac{\partial E[y_i|x_i]}{\partial x_{ji}} = e^{\beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}} \beta_j = E[y_i|x_i] \beta_j$$

Portanto, conclui-se que uma variação unitária no regressor $j_{ésimo}$ leva a uma mudança na média condicional no tamanho de $e[y_i|x_i]\,\beta_j$, que no modelo linear seria apenas β_j. Os autores sugerem que o impacto de uma variação unitária no regressor $j_{ésimo}$ pode ser visto como uma mudança proporcional de β_j em $E[y_i|x_i]$, pois $\frac{\partial E[y_i|x_i]/E[y_i|x_i]}{\partial x_{ji}} = \beta_j$.

Nos casos em que a variável explicativa também é transformada no logaritmo natural dela, conclui-se que o β_j é a elasticidade.

4.4.2 Restrições

Devido à sua característica de parâmetro único para localização e escala – que na distribuição normal são μ e σ^2, respectivamente –, todos os momentos de y são funções de μ. Este problema aparece quando os dados são sobredispersos, neste caso, distribuições com dois parâmetros levam vantagem em relação à Poisson.

Uma forma em que esta restrição manifesta-se é que em muitas aplicações a função de probabilidade Poisson prevê a probabilidade de assumir zero é consideravelmente menor do que o efetivamente observado. Na literatura, denomina-se este problema como problema de excesso de zeros, visto que há mais zeros nos dados do que na previsão Poisson.

Outra forma em que a restritividade aparece é quando a variância excede a média, denominado sobredispersão. Isto é um problema devido à propriedade de equidispersão da distribuição Poisson. A sobredispersão tem consequências similares à não verificação da hipótese de homocedasticidade na regressão linear. Supondo que a média condicional foi corretamente especificada, o estimador Poisson de máxima verossimilhança é ainda consistente. A sobredispersão gera subestimação dos erros-padrão dos estimadores e superestima a estatística t-student, por consequência, é importante empregar um estimador robusto da variância dos estimadores. O estimador robusto é conhecido como White-Huber, ou *sandwich*.

4.4.3 Teste para Sobredispersão

Para testar a sobredispersão e subdispersão, é necessário, primeiramente, especificar a forma dela, a maioria dos modelos de contagem de dados a especifica da seguinte forma

$$Var[y_i|x_i] = \mu_i + \alpha\, g(\mu_i),$$

Onde α é um parâmetro desconhecido e $g(.)$ é uma função conhecida, comumente $g(\mu) = \mu^2$ ou $g(\mu) = \mu$. Assume-se que tanto sob a hipótese nula quanto sob a hipótese alternativa a média é corretamente especificada (por exemplo, $e^{x_1'\beta}$), considerando que sob a hipótese nula o $\alpha = 0$, de modo que $Var[y_i|x_i] = \mu_i$.

O teste assume $H_0: \alpha = 0$ versus $H_1: \alpha \neq 0$, calcula-se após a estimativa do modelo ao construir valores ajustados $\widehat{\mu_i} = e^{x_i'\widehat{\beta}}$e computar a regressão auxiliar de MQO sem constante

$$\frac{(y_i - \widehat{\mu_i})^2 - y_i}{\widehat{\mu_i}} = \alpha \frac{g(\widehat{\mu_i})}{\mu_i} + u_i,$$

Onde u_ié o termo de erro aleatório. As estatísticas t-student dos estimadores são assintóticamente normais sob a hipótese nula de equidispersão, ou seja, $\alpha = 0$.

4.5 Binomial Negativo

O modelo binomial negativo – um exemplo específico de "modelo mistura" – pode ser obtido de diversas maneiras.

Uma variável discreta não-negativa y segue a distribuição Poisson e é condicionada ao parâmetro λ, de modo que $f(y|\lambda) = \frac{e^{-\lambda}y^{\lambda}}{y!}$. O parâmetro λ é estocástico, ou seja, não é uma função completamente determinística dos regressores, e é descrito como $\lambda = \mu v$, onde μ é uma função determinística dos regressores e $v > 0$ é independente e identicamente distribuído com distribuição de probabilidade $g(v|\alpha)$. Isto pode ser considerado heterogeneidade, pois dado que diferentes observações podem ter diferentes λ (o que demonstra heterogeneidade), essas diferenças são oriundas do componente não-observado aleatório v. Quando $E[v] = 1$, tem-se que $E[\lambda|y] = \mu$, que é idêntico à Poisson. Portanto, pode-se dizer que a Poisson é um caso específico da

binomial negativa e, por consequência, as interpretações mantem-se idênticas às dadas no modelo Poisson.

A distribuição marginal de y – não condicionada ao parâmetro aleatório v, mas condicionada aos parâmetros determinísticos μ e α (conforme mostrado na seção do Modelo Poisson) – é obtida pela integração de v. Que produz

$$h(y|\mu,\alpha) = \int f(y|\mu,v)\, g(v|\alpha)\, dv,$$

onde $g(v|\alpha)$ é chamada de distribuição mista e α denota o parâmetro desconhecido da distribuição mista. A integração defini uma distribuição média.

A distribuição negativa binomial foi originalmente[44] derivada como um caso limite das mistura entre a função de probabilidade Poisson e da função de probabilidade Gama. Portanto, trabalhar-se-á ela a mistura das duas, tal que y siga a distribuição Poisson , $f(y|\lambda) = \frac{e^{-\lambda}y^{\lambda}}{y!}$, e que v siga a distribuição gama – $g(v) = \frac{v^{\delta-1}e^{-v\delta}\delta^{\delta}}{\Gamma(\delta)}$ tal que $v, \delta > 0$, com $E[v] = 1$ e $Var[v] = \frac{1}{\delta}$.

Obtêm-se, então, a negativa binomial

$$h[y|\mu,\delta] = \int_0^{\infty} \frac{e^{-\mu v}\,(\mu v)^y}{y!}\,\frac{v^{\delta-1}e^{-v\delta}\delta^{\delta}}{\Gamma(\delta)}\,dv$$

$$= \frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}$$ [45]

Onde $\alpha = \frac{1}{\delta}$ e $\Gamma(.)$ denota a integral gama. A Poisson pode ser vista como um caso especial da negativa binomial quando $\alpha = 0$, assim como a geométrica pode ser vista como um caso especial quando $\alpha = 1$.

Os dois primeiros momentos da distribuição negativa binomial são

$$E[y|\mu,\alpha] = \mu$$

$$Var[y|\mu,\alpha] = \mu(1+\alpha\mu) = \mu + \alpha\mu^2$$

A variância excede a média se $\alpha > 0$ e $\mu > 0$. Deduz-se também que a hipótese de equidispersão não é satisfeita quando $y|\lambda$ é Poisson e a heterogeneidade não-observada tem forma multiplicativa $\lambda = \mu v$, onde $E[v] = 1$.

[44] Greenwood M, Yule GU (1920).
[45] Para ver a demonstração completa, Cameron e Trivedi (1998).

Há duas abordagens padrão para o modelo negativo binomial – sendo denominadas de NB1 e NB2 –, ambas especificam o μ da mesma forma, ou seja, $\mu = e^{x_i'\beta}$, contudo, a especificação do parâmetro α é distinta

$$\begin{cases} NB1: Var[y|\mu,\alpha] = (1+\gamma)\mu, & \alpha = \dfrac{\gamma}{\mu} \\ NB2: Var[y|\mu,\alpha] = \mu + \mu^2 \end{cases}$$

Como observado, a variante NB1 tem uma função de variância linear, ao passo que a variante NB2 possui função de variância quadrática. De forma generalizada, os modelos da classe binomial negativo tem a função de variância determinada por $\mu_i + \alpha\mu_i^p$, onde a NB2 tem $p = 2$ e a NB1 tem $p = 1$. As distribuições de probabilidade são iguais à exposta acima (representada por $h[y|\mu,\delta]$), todavia, o α^{-1} é substituído por $\alpha^{-1}\mu^{2-p}$. Ambas variantes tem parâmetros estimados por máxima verossimilhança

4.5.1 Negativo Binomial 2

Conforme supracitado, a NB2 tem a função variância determinada por $\mu + \mu^2$e a distribuição de probabilidade

$$f(y|\mu,\alpha) = \frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}$$

Que reduz-se à Poisson quando $\alpha = 0$. A função $\Gamma(.)$ é a função gamma[46], e tem como uma de suas propriedades que $\frac{\Gamma(y+a)}{\Gamma(a)} = \prod_{j=0}^{y-1} j + a$ quando y é inteiro. Deste modo,

$$\ln\left(\frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})}\right) = \sum_{j-0}^{y-1} \ln(j+\alpha^{-1})$$

Ao substituir a igualdade acima na função de distribuição de probabilidade, tem-se que a função de máxima verossimilhança para a média exponencial $\mu_i = e^{x_i'\beta}$ é

[46] $\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt,\ a > 0$

$$\ln L(\alpha,\beta) = \sum_{i=1}^{n}\left\{\sum_{j-0}^{y-1}\ln(j+\alpha^{-1}) - \ln y_i! - (y_i+\alpha^{-1})\ln\left(1+\alpha e^{x_i'\beta}\right) + y_i\ln\alpha + y_i x_i'\beta\right\}$$

As condições de primeira ordem são

$$\sum_{i=1}^{n}\frac{y_i-\mu_i}{1+\alpha\mu_i}x_i = 0\,, \qquad para\ o\ \hat{\beta}$$

e

$$\sum_{i=1}^{n}\left\{\frac{1}{\alpha^2}(\ln(1+\alpha\mu_i) - \sum_{j=0}^{y_{i-1}}\frac{1}{j+\alpha^{-1}} + \frac{y_i-\mu_i}{\alpha(1+\alpha\mu_i)}\right\} = 0, \qquad para\ o\ \hat{\alpha}$$

4.5.2 Negativo Binomial 1

Diferentemente da NB2, a NB1 tem a função variância determinada por $\mu+\alpha\mu$e a distribuição de probabilidade

$$f(y|\mu,\alpha) = \frac{\Gamma(\alpha^{-1}\mu+y)}{\Gamma(\alpha^{-1}\mu)\Gamma(y+1)}\left(\frac{\alpha^{-1}\mu}{\alpha^{-1}\mu+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}\mu}\right)^{y}$$

Ao aplicar a propriedade $\frac{\Gamma(y+a)}{\Gamma(a)} = \prod_{j=0}^{y-1} j+a$ à distribuição de probabilidade da NB1, tem-se que

$$\ln\left(\frac{\Gamma(y+\alpha^{-1}\mu)}{\Gamma(\alpha^{-1}\mu)}\right) = \sum_{j-0}^{y-1}\ln(j+\alpha^{-1}\mu)$$

Ao substituir a igualdade acima na função de distribuição de probabilidade, tem-se que a função de máxima verossimilhança para a média exponencial $\mu_i = e^{x_i'\beta}$ é

$$\ln L(\alpha,\beta) = \sum_{i=1}^{n}\left\{\sum_{j-0}^{y-1}\ln(j+\alpha^{-1}e^{x_i'\beta}) - \ln y_i! - (y_i+\alpha^{-1}e^{x_i'\beta})\ln(1+\alpha) + y_i\ln\alpha\right\}$$

As condições de primeira ordem são

$$\sum_{i=1}^{n}\left\{\left(\sum_{j=0}^{n-1}\frac{\alpha^{-1}\mu_i}{j+\alpha^{-1}\mu_i}\right)x_i + \alpha^{-1}\mu_i x_i\right\} = 0\,, \qquad para\ o\ \hat{\beta}$$

e

$$\sum_{i=1}^{n} \frac{1}{\alpha^2}\left\{-\left(\sum_{j=0}^{y_{i-1}} \frac{\mu_i}{j+\alpha^{-1}}\right) - \alpha^{-2}\mu_i \ln(1+\alpha) - \frac{\alpha}{1+\alpha} + y_i\alpha\right\} = 0, \qquad para\ o\ \hat{\alpha}.$$

4.5.3 Discussão

Conforme Cameron e Trivedi (1998), o modelo NB2 é o mais popular devido à propriedades não compartilhadas pelos demais modelos NB, que são (a) matriz de informação é bloco diagonal, ou seja, os termos externos à diagonal da matriz de informação são iguais a zero, portanto, tem-se que a $Cov\left[\widehat{\beta_{NB2}}, \widehat{\alpha_{NB2}}\right] = 0$, (b) robustez à má-especificação distribucional e (c) é um caso geral da distribuição geométrica quando $\alpha = 1$. Os estimadores do modelo NB2 são robustos à má-especificação da distribuição devido ao fato de pertencer à família exponencial linear quando o parâmetro de dispersão α é conhecido. Deste modo, os estimadores NB2 serão consistentes para β se a média condicional estiver corretamente especificada (essa condição é verificada pela condição de primeira ordem, cujo valor esperado é zero se a média condicional for corretamente especificada. Isso mantém-se pois $E[y_i - \mu_i|\,\boldsymbol{x_i}] = 0$).

Os erros padrão dos estimadores NB2 de máxima verossimilhança serão geralmente inconsistentes se houver qualquer má-especificação distribucional, ou seja, eles não serão consistentes se a condição de que a variância é especificada por $Var[y_i|\boldsymbol{x_i}] = \mu_i + \alpha\mu_i^2$ não estiver correta. Mesmo que a variância esteja bem especificada, a variância estimada pode ser inconsistente. As condições para que a estimativa de α seja consistente, é necessário que (i) $E[y_i - \mu_i|\boldsymbol{x_i}] = 0$ e que (ii) $\sum_{i=1}^{n}\left\{\ln(1+\alpha\mu_i) - \sum_{j=0}^{y_i-1}\frac{1}{(j+\alpha^{-1})}\right\} = 0$[47], essa condição é atendida somente se a variável dependente segue, de fato, uma distribuição negativa binomial.

O único modelo consistente à má-especificação funcional da família negativa binomial é o NB2. Então, a consistência dos estimadores de máxima verossimilhança para β requer que os dados sigam a distribuição negativa binomial.

[47] Lembrando que esta é parte da condição de primeira ordem em função do α.

4.6 *Hurdle* ou duas partes

Segundo Cameron e Trivedi (1998), em muitas ocasiões os modelos Poisson e Binomial Negativo não conseguem prever muito bem a ocorrência de zeros excessivos e faz-se necessário, portanto, que empregue-se modelos distintos que consigam lidar com isto. Neste contexto surge o modelo Hurdle (ou modelo de duas partes), este modelo relaxa a hipótese de que os zeros e os valores positivos originam-se do mesmo processo gerador de dados. A ideia central é que existe uma distribuição de probabilidade que governa o resultado binário de um evento que pode não-ocorrer ou ter realização positiva. Caso realização seja positiva, há uma distribuição de probabilidade que rege a magnitude desse evento. Procura-se, portanto, modelar esses dois processos distintos, ou seja, os zeros e os valores positivos. Atribui-se uma função de probabilidade $f_1(.)$ para o processo gerador dos zeros, tal que $P[y = 0] = f_1(0)$. Já os valores positivos vem da função de probabilidade truncada $f_2(y|y > 0) = \frac{f_2(y)}{(1-f_2(0))}$[48], que é multiplicado por $P[y > 0] = 1 - f_1(0)$ para assegurar que soma das probabilidades seja igual a 1. Deste modo,

$$g(y) = \begin{cases} f_1(0) & se\ y = 0 \\ \dfrac{1 - f_1(0)}{1 - f_2(0)}\ f_2(y) & se\ y \geq 1 \end{cases}$$

Com isso, o modelo reduz-se ao modelo padrão se $f_1(.) = f_2(.)$, ou seja, quando a função de probabilidade da geração de zeros é igual à função de probabilidade da contagem dos valores positivos, o modelo padrão é um caso específico do hurdle. Desta maneira, os processos geradores de dados – tanto dos zeros quanto dos positivos – não são restritos a serem iguais. Este modelo consegue lidar tanto com excesso de zeros quanto com pouquíssimos zeros, pois estima-se os dois processos.

A estimação de máxima verossimilhança do *hurdle* envolve a solução da condição de primeira ordem para os dois termos, ou seja, maximiza-se separadamente a verossimilhança para o processo correspondente aos zeros e para os positivos. A

[48] A função de distribuição de probabilidade é dada por $f(y|\theta)$ e a função de distribuição de probabilidade acumulada é dada por $F(y|\theta) = \Pr(Y < y)$, onde θ é o vetor de parâmetros. Ao omitir, por exemplo, o zero, temos que a distribuição de probabilidade truncada é dada por $f(y|\theta, y \geq 1) = \frac{f(y|\theta)}{1-F(0|\theta)}$, $y = 1, 2, 3, ...$ Emprega-se essa solução para não gerar inconsistência nos estimadores.

interpretação dada ao *hurdle* é que ele reflete um procedo de tomada de decisão de duas etapas, tal que a primeira e a segunda são regidas por diferentes mecanismos.

O *hurdle* combina um modelo de contagem de dados para os dados truncados à esquerda em $y = 1$ e um modelo de resposta binária para os dados censurados à direita em $y = 1$.[49]

A interpretação dos estimadores do modelo de contagem do *hurdle* mantém-se idêntica às dos demais modelos de uma etapa – já que a sua estrutura é a mesma – assim como os estimadores do modelo de resposta binária (que prevê se o evento irá ou não ocorrer) tem a mesma interpretação.

Conforme visto em Mullahy (1986), os dois primeiros momentos do *hurdle* são determinados pela probabilidade de ultrapassar o limiar e pelos momentos da distribuição truncada nos zeros, isto é

$$E[y|x] = \Pr[y > 0|\,\boldsymbol{x}]\; E_{y>0}[y|y > 0, \boldsymbol{x}]$$

E a variância pode ser vista como

$$V[y|\boldsymbol{x}] = \Pr[y > 0|\,\boldsymbol{x}]\; V_{y>0}[y|y > 0, \boldsymbol{x}] + \Pr[y = 0|\boldsymbol{x}]\; E_{y>0}[y|y > 0, \boldsymbol{x}]$$

Para maiores detalhes e demonstrações, verificar Mullahy (1986).

4.7 Zero-inflated models

Outra abordagem para lidar com o excesso de zeros é conhecida como modelos de zero inflados. A lógica é similar à empregada no modelo *hurdle*, todavia, considera-se que alguns zeros são estruturais – ou seja, a observação não poderá assumir um valor diferente de zero –. Assim como no hurdle, atribui-se uma função de probabilidade de contagem $f_2(.)$ com um processo binário, cuja função de probabilidade é $f_1(.)$. Caso o processo binário assuma o valor 0, com probabilidade $f_1(0)$, então, $y = 0$, entretanto, caso o processo binários assuma o valor 1, com probabilidade $f_1(1)$, então y assume valores de contagem $0, 1, 2, \ldots$ oriundos de $f_2(.)$. Deste modo, observa-se que a

[49] É importante distinguir a censura e o truncamento. O primeiro refere-se à agregação de dados maiores do que um determinado valor, embora estes sejam observados na amostragem, ao passo que o último não observa os dados em diferentes valores (não são capturados na amostragem). Ou seja, a diferença é que na censura, tanto os resultados quanto as variáveis são observadas, enquanto que no truncamento não observa-se o resultado, tampouco as variáveis. Um exemplo de truncamento é coletar amostra sobre frequência de viagens aéreas num determinado período com pessoas que estão viajando num avião (ou seja, observar-se-ão apenas indivíduos cujo número de viagens é maior ou igual a um), já a censura ocorre em uma pesquisa que elimina os valores extremos com pouquíssimos casos.

contagem de zeros ocorre em duas maneiras distintas: (i) como realização do processo binário e (ii) como realização do processo de contagem quando a variável binária assume o valor 1. A função de probabilidade é dada por

$$g(y) = \begin{cases} f_1(0) + \big(1 - f_1(0)\big) f_2(0) & se\ y = 0 \\ \big(1 - f_1(0)\big) f_2(y) & se\ y \geq 0 \end{cases}$$

A intuição por trás desse modelo pode ser exemplificada pelo caso de viagens recreativas. Considere que seja necessário estimar a quantidade de viagens que uma pessoa realiza num determinado período de tempo, esta variável de contagem pode assumir diversos valores inteiros não-negativos. O processo que rege os zeros pode ocorrer de duas maneiras, (i) o agente não tem tempo livre ou dotação monetária neste período de tempo, ou (ii) o agente tem tempo livre e dotação monetária, entretanto, ele prefere outras formas de lazer. O primeiro processo é definido como um zero estrutural, ou seja, o valor da contagem não diferirá de zero devido a algum vetor de restrições, ao passo que o segundo é definido como zero amostral (tradução livre do termo inglês "sampling zero"), onde o agente pode assumir diversos valores, mas preferiu realizar nenhum.

Visto que não há existência teórica de zeros estruturais para o problema de consultas, este modelo não será empregado neste estudo.

5 Aplicação e escolha

5.1 Dados

Os dados utilizados foram extraídos da base de clientes de uma operadora de planos de saúde situada em Porto Alegre, Rio Grande do Sul. A amostra conta com mais de 600.000 observações, embora não tenha sido integralmente utilizada devido a questões computacionais, desta forma, a amostra selecionada foi de aproximadamente 340 mil. A base de dados disponível sofre de algumas limitações de (i) indisponibilidade de algumas variáveis – como, por exemplo, pacientes crônicos, renda e educação – amplamente utilizadas na literatura[50], e de (ii) impossibilidade de extração de séries temporais – visto que algumas variáveis são constantemente alteradas e o histórico, em consequência, não fica armazenado, como, por exemplo, a coparticipação –, todavia, consoante apontado por Chandra (2010), o comportamento dos agentes aparenta ser estável ao longo do tempo e isso não deve ser um problema.

Os dados estão em formato de corte transversal e compreendem o período de um ano. Utilizou-se apenas beneficiários que estivessem ativos durante todo o período, isto é, que poderiam ter se consultado ao longo de todo o período. Caso contrário, seria necessário controlar o tempo em que ele tinha possibilidade de realizar consultas. Isto pode ocorrer em consequência de novos contratos, rescisões de contratos ou período de carência. A adoção de um período de carência é uma maneira de coibir um comportamento oportunista, uma vez que agentes poderiam entrar no plano, consumir os recursos que desejavam – como, por exemplo, uma cirurgia – e cancelar o seu prêmio subsequentemente à realização.

A variável dependente é o número de consultas realizadas no período de um ano. As consultas utilizadas são as realizadas em consultório, isto é, elimina-se a quantidade de consultas realizadas em emergência, pois supõe-se que neste caso os agentes sejam pouco sensíveis a preços altos de coparticipação. As variáveis independentes disponíveis são: (i) produto, que é o tipo de plano e determina o tamanho da rede disponível para atendimento, isto é, o número de prestadores a que o beneficiário tem acesso. O produto A é o que possui maior disponibilidade de rede, ao passo que o C é o

[50] Deb e Trivedi (1997).

que possui a menor. Outro ponto importante a ser destacado sobre os beneficiários do produto C é o modelo de plano, pois o produto C segue uma estrutura com controle de consultas, uma vez que os beneficiários devem consultar-se num núcleo da operadora – que contém clínicos gerais que podem encaminhar os pacientes para médicos especialistas de acordo com a necessidade; (ii) cobertura, que determina a quais tipos de serviços o beneficiário terá acesso, ou seja, a variedade de procedimentos que estão cobertos pelo plano de saúde; (iii) gênero, isto é, feminino ou masculino; (iv) cidade do beneficiário[51], agregadas por grandes blocos, que são: (a) Porto Alegre, (b) algumas cidades da Grande Porto Alegre[52] e (c) demais cidades; (v) idade do beneficiário; (vi) tempo no plano tratada como categórica; (vii) modalidade de plano, ou seja, se o contratante recebe o plano como auxílio da empresa ou se decidiu comprá-lo; (viii) coparticipação.

Nota-se que a base contém boa parte das variáveis elencadas por Manning (1987) como relevantes, contudo, não há disponibilidade da variável renda. A omissão da variável renda compromete a interpretação das variáveis produto, cobertura e segmento, posto que estas possuem potencialmente alta correlação com a renda[53], fazendo com que os estimadores destas variáveis sejam viesados. Outro problema que surge com estas variáveis é a potencial endogeneidade da escolha do plano, que pode refletir outras características não-observáveis do agente. Dito isto, atribui-se a estas variáveis somente o papel de controle. O tempo no plano é uma variável importante, já que espera-se que novos beneficiários que não tinham acesso a planos de saúde consumam bastante nos primeiros meses, ao passo que reduzam o consumo nos anos subsequentes. Esta variável foi tratada como categórica – dividida em "0-12 meses", "13-24 meses", "25 meses ou mais" – a fim de evitar eventuais distorções, uma vez que beneficiários há bastante tempo no plano tem no mínimo "bastante tempo" de idade. O produto é classificado em três categorias: A, B e C, onde A representa o produto mais amplo. Essa mesma categorização aplica-se às coberturas.

[51] Adiciona-se a variável categórica cidade devido à disponibilidade de prestadores, uma vez que cidades grandes possuem um número maior de prestadores e médicos com consultório. Em cidades pequenas, os beneficiários tem poucas opções de médicos, o que os induz a utilizar menos consultas eletivas e deslocar-se até um hospital/pronto-atendimento.

[52] Cidades utilizadas para o bloco: Canoas, Gravataí, Cachoeirinha, Esteio, Viamão e São Leopoldo.

[53] A potencial alta correlação com a renda explica-se pelo preço dos planos de saúde, uma vez que planos com mais abrangência e cobertura custam muito mais caro, o que indica (i) maior renda e/ou (ii) maior preço de reserva por seguro. Um fator que pode, aparentemente, dificultar as conclusões de que esta é uma boa proxy para a renda é o segmento do plano, ou seja, se ele é pago pela empresa ou pelo próprio beneficiário. È de praxe que empresas paguem planos mais completos para seus funcionários de alto escalão e planos mais baratos para os funcionários de baixo escalão,

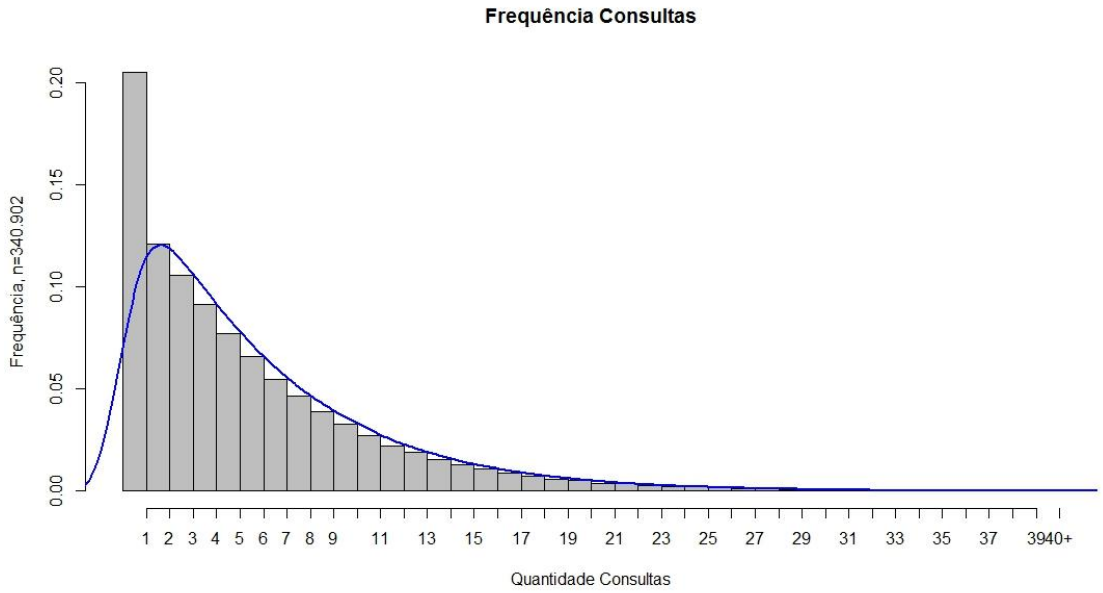
Tabela 1 – Definição variáveis e estatísticas descritivas

Variável	**Descrição**	**Média**	**DP**
Consultas Eletivas	Número consultas eletivas	5.81	5.540
Consultas Emergência	Número consultas emergência	1.07	1.810
Coparticipação	Preço da Coparticipação	17.74	12.380
Idade	Idade do beneficiário no final do período	37.53	21.130
<= 12 meses plano	= 1 se tiver há menos do que um ano no plano	0.12	0.328
> 24 meses plano	=1 se tiver há mais de dois anos no plano	0.65	0.477
Produto A	=1 se possuir produto A	0.59	0.491
Produto C	=1 se possuir produto C	0.15	0.358
Cobertura A	=1 se possuir cobertura A	0.87	0.340
Cobertura B	=1 se possuir cobertura B	0.10	0.186
Plano Empresarial	= 1 se o plano for empresarial	0.81	0.390
Gênero Feminino	= 1 se beneficiário for mulher	0.57	0.495
Cidade Porto Alegre	= 1 se cidade contratante Porto Alegre	0.77	0.421
Grande Porto Alegre	= 1 se região da GPA, exceto Porto Alegre	0.14	0.347

Fonte: Elaboração própria.

Ao analisar a quantidade de consultas, vemos que a média é 5.81 e a variância é 30,64 (quadrado do desvio-padrão), o que sugere que o processo gerador de dados não deve ser regido por Poisson, pois observa-se sobredispersão. A distribuição dela é assimétrica, com obliquidade (*skewness*, do termo inglês) positiva e elevada e leptocúrtica. Na figura 4 é possível verificar como ocorre a distribuição dela

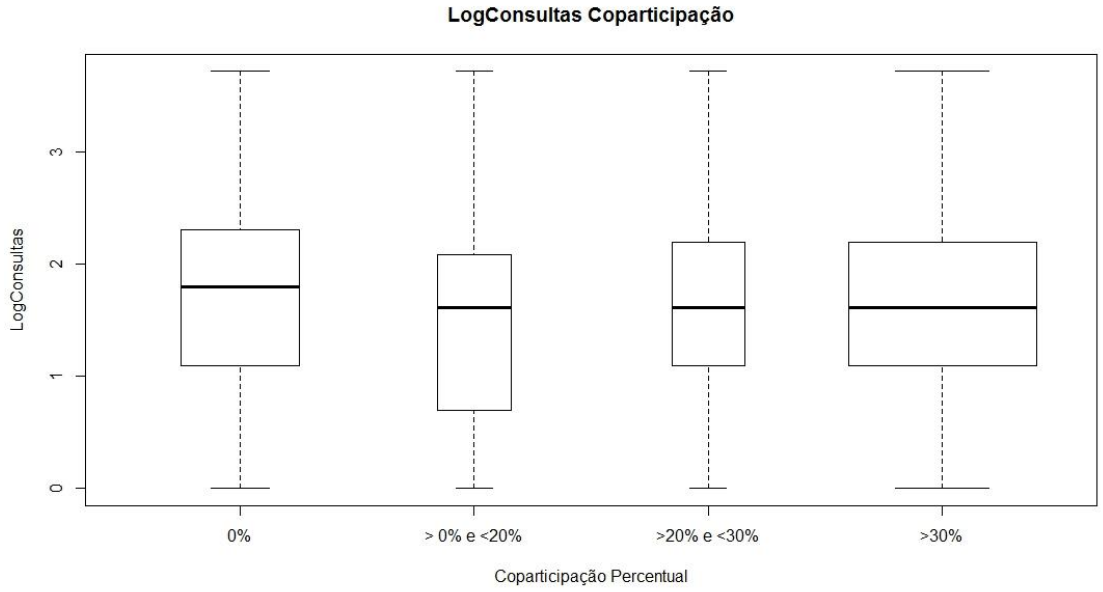
Figura 4 – Distribuição das consultas



Fonte: Elaboração própria

Ao analisar a dispersão de consultas cruzada – graficamente representada na figura abaixo – pelo nível de coparticipação, tem-se

Figura 5 – Consultas vs Coparticipação



Fonte: Elaboração própria

A leitura adequada da imagem acima é que a linha representa a mediana e a altura da caixa a distância interquartílica da variável consulta para cada um dos grupos

elencados no eixo x. O gráfico contém cinco elementos importantes: (i) a linha espessa dentro da caixa representa a mediana, (ii) o limite superior da caixa representa o quartil superior; (iii) o limite inferior da caixa representa o quartil inferior; (iv) a linha externa horizontal superior representa $Q3 + 1{,}5 \times (Q3 - Q1)$, onde $Q1$ é o primeiro quartil e $Q3$ é o terceiro quartil; e (v) a linha externa horizontal inferior representa $Q3 - 1{,}5 \times (Q3 - Q1)$.

A tabela abaixo contém as frequências de ocorrência de consultas, censurada no ponto 9.

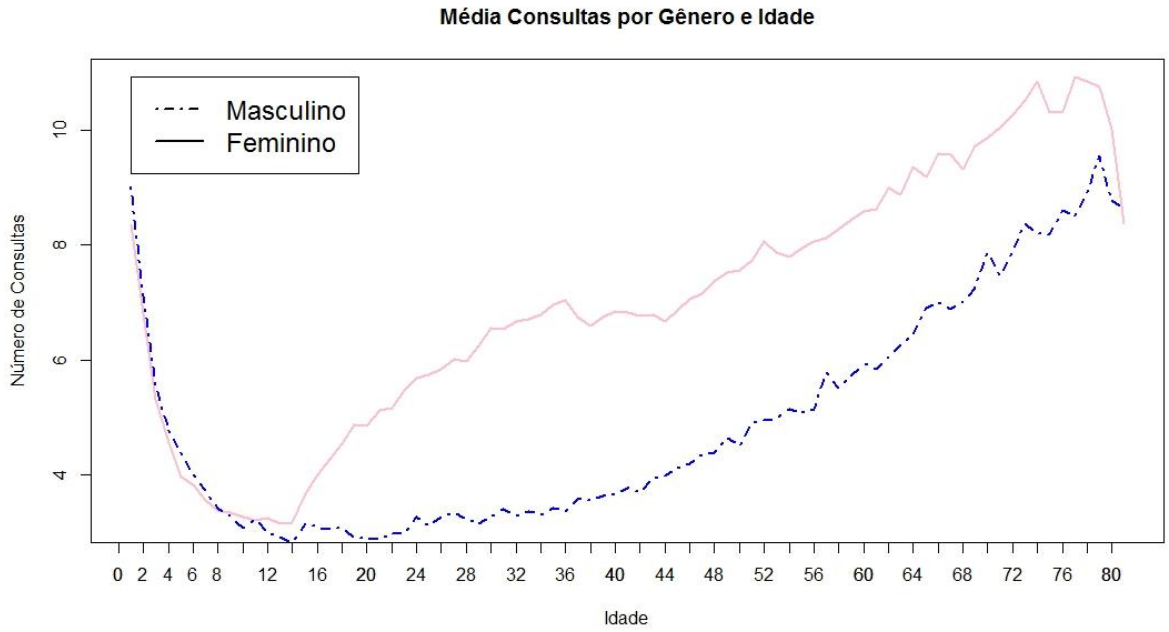
Tabela 2 – Frequência Consultas

Consultas	Freqüência
0	7.75%
1	12.73%
2	12.08%
3	10.55%
4	9.14%
5	7.73%
6	6.58%
7	5.50%
8	4.66%
≥ 9	23.27%

Fonte: Elaboração Própria

Para demonstrar o quão importante é a inserção das variáveis idade e gênero no modelo, elabora-se um gráfico que compara a média de consultas de acordo com a idade,

Figura 6 – Comparativo média consultas por gênero e idade



Fonte: Elaboração Própria

Observa-se que a quantidade de consultas é aparentemente idêntica entre os gêneros ao longo da infância. Uma das possíveis explicações para esse fenômeno é que a determinação da quantidade de consultas dá-se pelos pais da criança. A partir dos 14 anos, observa-se que a diferença do número de consultas entre os gênero começa a tomar forma, sendo o ápice da diferença entre 28 e 36 anos. Wang *et al* (2013) sugere que o maior hiato ocorre entre os 16 e 60 anos e que a maior parte deste é devido às consultas atinentes à reprodução. O autor ainda conclui que ao considerar pacientes em mesmo grupo de doença crônica, a média de consultas para pacientes do sexo feminino é apenas 8% maior do que a média para o masculino. Um ponto importante a ser destacado é o não-excesso de consultas por parte de mulheres, que é destacado por Wang *et al.*

5.2 Teste modelos candidatos

Os modelos serão testados em duas categorias diferentes: (i) modelos que não são baseados em máxima verossimilhança, isto é, MQO, e (ii) modelos que são, isto é, demais modelos. Divide-se em dois grupos devido à não comparabilidade destes, visto

que o ajuste de um é determinado pelo coeficiente de determinação ao passo que o outro é pela função de verossimilhança.

5.2.1 Verificação MQO

No modelo MQO, utilizou-se a forma log-log do modelo, em razão do resultado do estimador da coparticipação ser a própria elasticidade. Em consequência disto, foi requerida a transformação da variável dependente e da coparticipação para log, contudo, ambas apresentam observações cujo valor é zero. Para contornar este problema, empregou-se a solução aplicada em Deb e Trivedi (2002). A solução empregada foi tomar o log da variável selecionada mais um, ou seja, $\ln(\text{consultas} + 1)$.

O resultado do modelo está sumarizado na tabela abaixo.

Tabela 3 – Modelo MQO estimativa

Estimador	Coeficiente	Erros-Padrão	Significância
Intercepto	1.3816	0.00697	***
Coparticipação	-0.0456	0.00103	***
Idade	0.0036	0.00021	***
Idade²	0.0001	0.00000	***
Plano 13-24 meses	-0.0139	0.00457	**
Plano 24 meses +	-0.0644	0.00408	***
Produto A	0.1737	0.00482	***
Produto B	0.0956	0.00506	***
Cobertura A	0.1796	0.00531	***
Cobertura C	-0.2105	0.00895	***
Plano Familiar	-0.0090	0.00348	***
Masculino	-0.3610	0.00261	***
Grande Porto Alegre	0.0394	0.00407	***
Outras cidades	0.0062	0.00465	

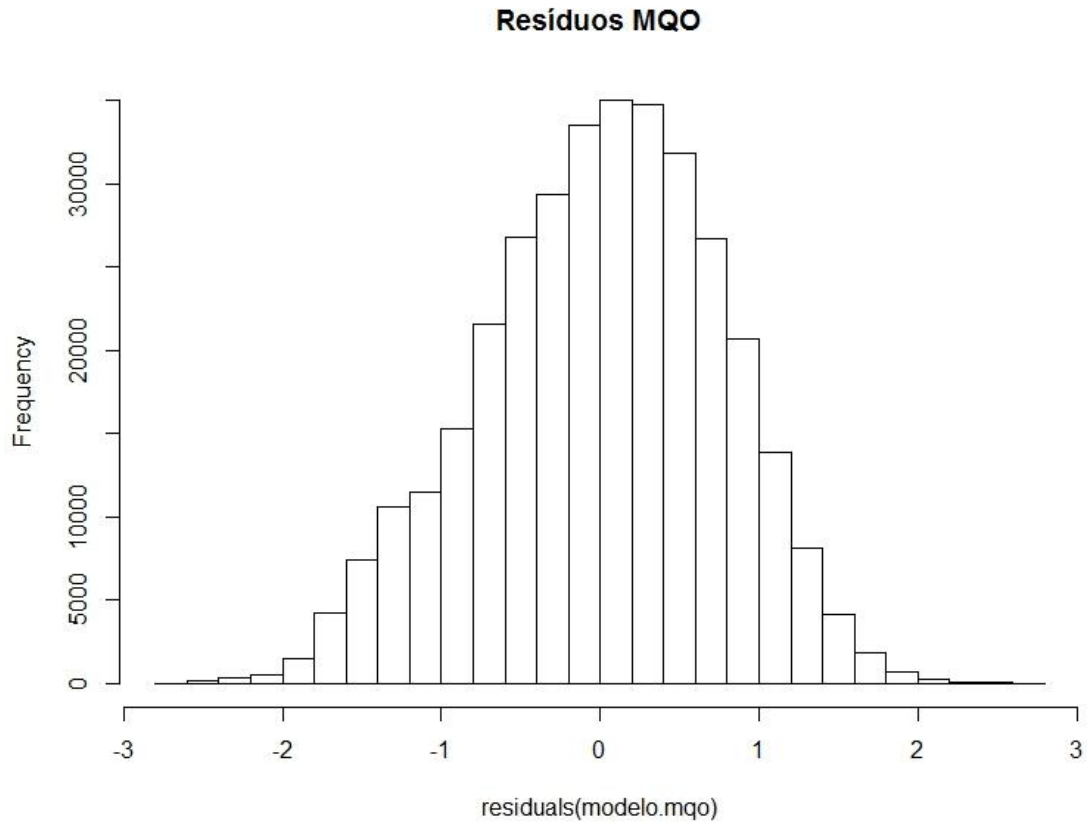
Códigos de Significância: (i) '***' a 0,001, (ii) '**' a 0,01, (iii) '*' a 0,05, (iv) '.' a 0,1, (v) ' ' a 1.
Fonte: Elaboração própria

Os erros-padrão foram estimados usando a correção proposta por White (1980), já que o teste de Breusch-Pagan indicou a presença de heterocedasticidade, com estatística BP = 1865,12, com 14 graus de liberdade e p-valor menor do que $2{,}2 \times 10^{16}$. A correção não alterou a significância de nenhum dos estimadores.

Quanto à multicolinearidade, o maior FIV encontrado (11,58) foi para as variáveis idade e quadrado da idade, o que é algo esperado e que não acarreta problemas. Para as demais, nenhuma apresentou FIV superior a 2.

O teste RESET não apresentou indícios de má-especificação da forma funcional. Os resíduos são normalmente distribuídos, de acordo com teste Jarque-Bera O histograma dos resíduos gerados pelo modelo está apresentado na figura abaixo.

Figura 7 – Resíduos MQO



Fonte: Elaboração própria

A elasticidade-preço estimada pelo método de mínimos quadrados ordinários foi de 4,56% e é significativa a um nível de 0,1%. Os testes calculados para verificação das hipóteses clássicas sugere que os estimadores são não-viesados, entretanto, não pertencentes à classe dos modelos de menor variância. O coeficiente de determinação, R^2, foi baixo (0,1358) e mostrou que as variações das variáveis independentes explicam apenas 13,58% das variações da variável independente.

5.2.2 Modelos estimados por máxima verossimilhança

Nesta fase, compara-se os modelos candidatos. Cinco modelos distintos foram testados, que são: (i) modelo Poisson, (ii) modelo negativo binomial 1, (iii) modelo negativo binomial 2, (iv) *hurdle* poisson e (v) *hurdle* NB2. Para calcular o modelo (i), utilizou-se o cálculo padrão da função nativa 'glm()' do software R, para os modelos (ii) e (iii), empregou-se também a função padrão 'glm()', todavia, utilizou-se uma alteração para a distribuição empregada, sendo necessário utilizar a ligação presente no pacote MASS[54]. Já para os modelos (iv) e (v), empregou-se a função 'hurdle()' do pacote pscl[55].

Os modelos e o número de parâmetros de cada estão expostos na tabela 4. É natural que os modelos NB tenham mais parâmetros do que o Poisson, posto que o modelo negativo binomial também estima a dispersão, ao passo que o Poisson toma ela como equidispersa.

Tabela 4 – Número de parâmetros de cada modelo

Modelo	Número de Parâmetros
Poisson	14
Negativo Binomial 1	15
Negativo Binomial 2	15
Hurdle Poisson	30
Hurdle NB2	31

Fonte: Elaboração própria

Na tabela 5, aplicou-se o teste razão verossimilhança para verificar se a inclusão de variáveis no modelo tem um incremento significativo na verossimilhança do modelo. O teste é implementado pela função 'lrtest()' do pacote 'lmtest'.

Tabela 5 – Teste Razão Verossimilhança

Modelo	χ^2	Significância
Poisson	219.610	***
Negativo Binomial 1	32.980	***
Negativo Binomial 2	57.176	***
Hurdle Poisson	178.287	***
Hurdle NB2	49.561	***

[54] Solução formulada por Brian Ripley, professor titular de Oxford e um dos principais colaboradores do R.

[55] Pacote disponibilizado por Simon Jackman, professor de Stanford, com auxílio de Achim Zeileis, da Universitaet Innsbruck.

Códigos de Significância: (i) '***' a 0,001, (ii) '**' a 0,01, (iii) '*' a 0,05, (iv) '.' a 0,1, (v) ' ' a 1.
Fonte: Elaboração própria

É possível observar que a adição das variáveis teve impacto significativo na verossimilhança dos modelos, porquanto rejeitou-se a hipótese nula de que o modelo restrito e não-restrito produzem aproximadamente a mesma verossimilhança.

A tabela 6 contém bastante informação, ela traz critérios de informação – que são baseados na função de verossimilhança –, a verossimilhança e a dispersão (quando for testada). Os erros-padrão (EP) dos estimadores já estão corrigidos pelo estimador *sandwich*, cujo cálculo é feito através da função '*sandwich*()' presente no pacote homônimo da função.

Tabela 6 – Comparação dos Modelos

	Modelo Poisson			Modelo NB1			Modelo NB2			Hurdle Poisson			Hurdle NB2		
	Coef	EP	Sig	Coef	EP	Sig	Coef	EP	Sig	Coef	EP	Sig	Coef	EP	Sig
Intercepto	1.40370	0.00420	***	1.48380	0.00846	***	1.47399	0.00850	***	1.55899	0.00371	***	1.49922	0.00674	***
Coparticipação	-0.05630	0.00054	***	-0.05690	0.00122	***	-0.05687	0.00121	***	-0.05274	0.00054	***	-0.05945	0.00128	***
Idade	0.00780	0.00012	***	0.00240	0.00025	***	0.00310	0.00025	***	0.00610	0.00001	***	0.00258	0.00025	***
Idade²	0.00003	0.00000	***	0.00010	0.00000	***	0.00009	0.00000	***	0.00004	0.00000	***	0.00010	0.00000	***
Plano 13-24 meses	-0.01891	0.00256	***	-0.02835	0.00549	***	-0.02721	0.00547	***	-0.02172	0.00255	***	-0.03488	0.00587	***
Plano 24 meses +	-0.08060	0.00228	***	-0.09282	0.00489	***	-0.09151	0.00488	***	-0.08104	0.00223	***	-0.10592	0.00522	***
Produto A	0.17471	0.00281	***	0.17714	0.00583	***	0.17680	0.00583	***	0.12114	0.00283	***	0.13288	0.00627	***
Produto B	0.08151	0.00296	***	0.08514	0.00612	***	0.08483	0.00613	***	0.03970	0.00299	***	0.04539	0.00664	***
Cobertura A	0.21062	0.00325	***	0.21726	0.00647	***	0.21616	0.00652	***	0.17757	0.00331	***	0.20874	0.00711	***
Cobertura C	-0.21463	0.00570	***	-0.20847	0.01098	***	-0.20921	0.01111	***	-0.14115	0.00582	***	-0.15403	0.01223	***
Plano Familiar	-0.00179	0.00187		0.00323	0.00415		0.00260	0.00412		0.00239	0.00187		0.00645	0.00429	
Masculino	-0.39307	0.00151	***	-0.40461	0.00315	***	-0.40305	0.00315	***	-0.34008	0.00153	***	-0.39201	0.00338	***
Grande Porto Alegre	0.06186	0.00229	***	0.06003	0.00490	***	0.06033	0.00489	***	0.06698	0.00231	***	0.07593	0.00523	***
Outras cidades	-0.00696	0.00266	**	-0.00077	0.00561		-0.00172	0.00561		-0.01589	0.00262	***	-0.01555	0.00603	***
AIC	2428231			1904983			1869184			2334747			1873430		
BIC	2428392			1905144			1869345			2334908			1873591		
lnL	-1214100			-952476.5			-934560.8			-1167344			-936700.1		
Dispersão				0.6844209			1.175035						0.4693312		

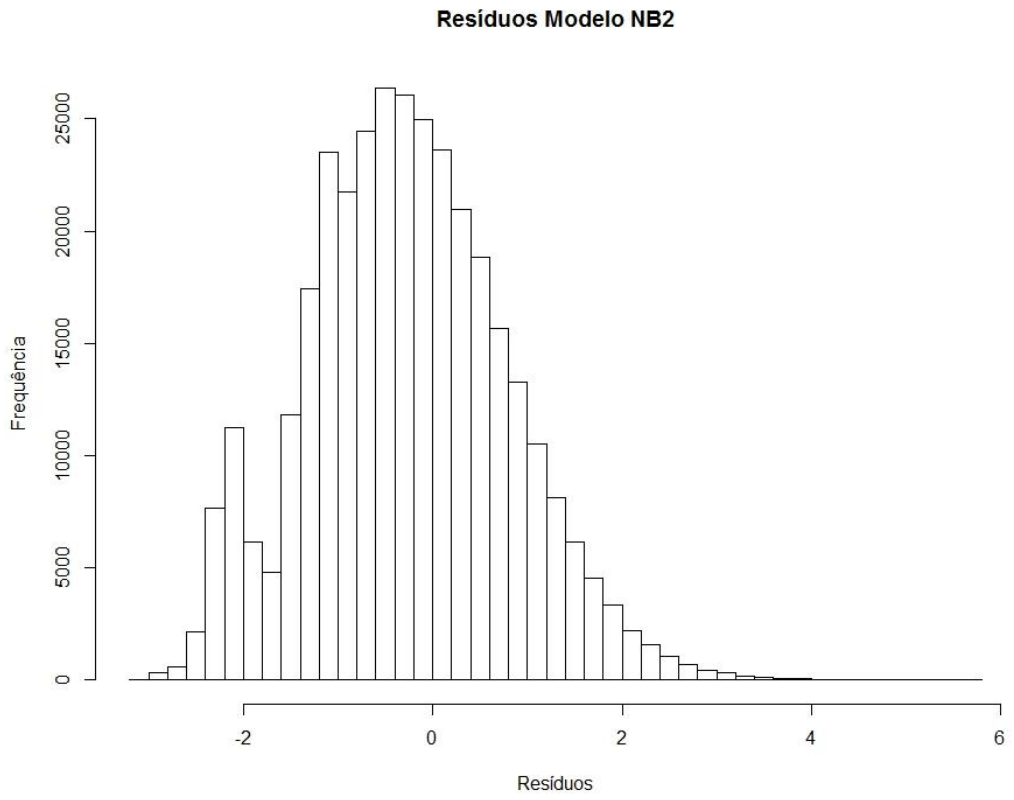
Códigos de Significância: (i) '***' a 0,001, (ii) '**' a 0,01, (iii) '*' a 0,05, (iv) '.' a 0,1, (v) ' ' a 1.
lnL é o log da verossimilhança.
Dispersão é o parâmetro calculado (se diferente de 0, não é equidispersão)
Fonte: Elaboração própria

A escolha do modelo foi simples, uma vez que tanto a verossimilhança quanto os critérios de informação apontam para o modelo NB2. Outro fator que corrobora com a escolha deste modelo é o não atendimento da hipótese de equidispersão – que justificaria a utilização da distribuição Poisson. Este resultado está em sintonia com o que fora afirmado por Deb e Travedi (2002), onde afirmam que o modelo de duas partes não fornece um enquadramento econométrico superior e que não é preciso modelar duas etapas distintas para capturar o problema agente-principal.

Na figura 8 é possível observar a distribuição dos resíduos gerados pelo modelo NB2, que aproxima-se de uma normal, o que é importante em modelos lineares

generalizados, uma vez que eles sinalizam se a escolha da distribuição é adequada. Para testar a normalidade, empregara-se o teste Jarque-Bera – que a verifica através da curtose e obliquidade – e aceitou-se a hipótese de normalidade.

Figura 8 – Histograma Resíduos Modelo NB2



Fonte: Elaboração própria

As frequências ajustadas estarão na próxima seção, comparadas aos valores observados.

5.3 Comparação MQO e NB2

A melhor – e talvez única – maneira de comparar estes modelos é verificando as frequências esperadas confrontadas com as frequências realizadas.

Na tabela abaixo há uma comparação entre as frequências observadas, frequência estimada por MQO e frequência estimada pelo modelo NB2, considerando ela em percentual.

Tabela 7 – Dados vs Previsão modelos

Consultas	Dados	MQO	NB2
0	7.73%	0.00%	0.00%
1	12.75%	0.00%	2.38%
2	12.09%	3.09%	16.96%
3	10.56%	21.58%	20.18%
4	9.15%	24.56%	21.14%
5	7.74%	23.80%	16.41%
6	6.58%	13.71%	9.50%
7	5.49%	6.72%	5.44%
8	4.66%	3.33%	3.10%
9 +	23.25%	3.22%	4.88%

Fonte: Elaboração própria

É possível notar que ambos os modelos não ajustam-se muito bem aos dados, sendo mais evidente nos dois primeiros intervalos de acordo com quantidade de consultas. No entanto, o modelo NB2 aparenta estar melhor ajustado aos dados. Apesar de não servir como um bom preditor, pode-se utilizar o resultado para entender o efeito médio. A estimativa da elasticidade, - 0,05687, está em consonância com o resultado obtido pelo estudo exposto em Cameron e Trivedi (2005), que aponta uma elasticidade estimada de - 0,0504[56]. É importante ressaltar que os autores tinham muito mais variáveis relevantes à disposição, uma restrição apontada no início do capítulo.

Uma das possíveis causas de baixo poder preditivo é a quantidade de variáveis relacionadas à condição de saúde dos beneficiários.

5.4 Análise resultado coparticipação

A elasticidade estimada de -0,05687 sugere – conforme apontado no capítulo dois – que ao aumentar o preço em 1%, espera-se que a quantidade consumida reduza em 0,056%. A magnitude da elasticidade é bem baixa. Para dar uma dimensão deste resultado na amostra, ao aumentar o preço da coparticipação em 1%, esperar-se-ia uma redução de 1.075[57] consultas no ano, o que equivaleria a uma redução de custos de aproximadamente R$ 106.000,00. Essa redução desconsidera a receita oriunda do mecanismo de coparticipação, ou seja, considera-se redução de custos somente a que é

[56] Artigo é citado no capítulo 20 do livro de Cameron e Trivedi (2005).
[57] A quantidade total de consultas na amostra selecionada é 1.983.377.

gerada pela redução da quantidade. Destarte, a redução de custos está subestimada. Num novo exemplo, onde a coparticipação é elevada em 20% - que é equivalente a R$3,54 da média do valor da coparticipação -, espera-se que a quantidade demandada reduziria em aproximadamente 19.840, a representar aproximadamente R$1.964.446,00, mantendo a restrição supracitada. É importante lembrar que a elasticidade estimada é constante.

Ao obter a estimativa da elasticidade, deseja-se comparar ela entre os diferentes tipos de produto, isto é, com acesso via núcleo e sem acesso via núcleo, uma vez que o plano C possui um mecanismo distinto para realização de consultas, assim como o tipo de plano pode servir como uma *proxy* da renda – visto que os planos sem núcleo são mais caros do que o plano com núcleo. A tabela seguinte mostra os valores estimados de coparticipação para cada tipo de produto

Tabela 8 – Elasticidades-preço consultas eletivas para diferentes produtos

Produto	Elasticidade-Preço	Erro-Padrão	Significância
Produtos A e B	-0.05752	0.00135	***
Produto C	-0.02296	0.00543	***

Fonte: Elaboração própria

É surpreendente que tenha havido tanta diferença na elasticidade estimada entre os produtos sem núcleo e o produto com núcleo, uma vez que os resultados obtidos por Manning (1987) sugerem que a elasticidade-preço não demonstra-se significativamente diferente entre os distintos níveis de renda, exceto para beneficiários de baixa renda com doença crônica[58]. Dado que não há a variável renda na base de dados, faz-se necessário supor que as constatações feitas por Manning (1987) são válidas e verificar se essa diferença surge por algum comportamento inesperado instigado pelo modelo de regulação. Ao notar que a magnitude do produto C ficou muito abaixo, procura-se entender se a estrutura do produto – isto é, o atendimento via núcleo com clínicos gerais – leva os beneficiários a algum comportamento distinto. Para tanto, é necessário observar o comportamento destes agentes frente a bens substitutos. Pode-se considerar consultas de emergência como bem substituto, visto que os beneficiários podem ir ao hospital para receber atendimentos que deveriam ser feitos em consultório. Ademais, emergências contam, geralmente, com médicos de diversas especialidades, o que

[58] Variáveis que não estavam disponível para este estudo.

eliminaria a necessidade de ir a mais de um lugar e estar sujeito à disponibilidade de agenda de um especialista.

Tabela 9 – Elasticidade-preço consultas de emergência para diferentes produtos

Produto	Elasticidade-Preço	Erro-Padrão	Significância
Produtos A e B	-0.05283	0.00248	***
Produto C	0.05267	0.00932	***
Geral	-0.04543	0.00223	***

Estimado por máxima verossimilhança, com NB 2. Fonte: Elaboração própria

Observa-se que a elasticidade dos beneficiários do produto C para consultas de emergência é positiva, um sinal que vai de encontro ao esperado pela teoria[59], pois este resultado sugere que quanto maior for o preço, maior será a demanda.

Para elucidar uma possível explicação, é necessário considerar o número de visitas ao médico por episódio de doença. Um agente está com algum sintoma e deseja consultar-se para descobrir a causa, ele pode ir tanto (i) ao núcleo quanto (ii) a um pronto-atendimento, em ambos os casos, a coparticipação paga será a mesma. Por fins de notação, denotar-se-á o valor esperado de consultas para o episódio de doença via núcleo de $E_n[y]$, onde y é o número de consultas, e o para o episódio de doença via hospital/pronto-atendimento é denotado por $E_h[y]$. Lembrando que o custo de coparticipação é dado por $co(y)$.

As duas principais diferenças entre o núcleo e a emergência são: (i) médicos especialistas e (ii) necessidade de exames. No núcleo há três especialidades, ao passo que um hospital possui muito mais. O hospital possui uma gama de equipamentos para realização de exames muito mais ampla do que o núcleo (que possui apenas exames muito simples).

Num cenário onde o episódio de doença é diagnosticado e tratado pelo clínico geral, temos que $E_n[y] = E_h[y]$, ou seja, a quantidade de consultas esperada é a mesma para a modalidade emergência/pronto-atendimento e para a modalidade núcleo. Todavia, no caso em que o diagnóstico necessite de um especialista não disponível no núcleo e não seja necessária a realização de algum exame, temos que $E_n[y] = E_h[y] + 1$, pois o beneficiário necessitará dirigir-se a um especialista fora do núcleo, ao passo que no hospital será necessária apenas uma ida (sem deslocamento e sem sujeição à

[59] A exceção são os bens de Giffen, onde o efeito renda supera o efeito substituição.

disponibilidade, embora talvez seja necessária uma longa espera até o atendimento). No caso em que o episódio de doença não requeira um especialista fora do núcleo, mas necessite a realização de um exame não disponível no núcleo, o beneficiário precisará agendar o exame em algum prestador e ir ao núcleo novamente, neste caso, $E_n[y] = E_h[y] + 1$. É importante ressaltar que há custos de deslocamento, custos de tempo despendido e outros custos associados às consultas. No último cenário, se o beneficiário não conseguir agendar o exame e a consulta subsequente em um intervalo de 15 dias, ele necessitará pagar a taxa de coparticipação pela segunda vez. Desta forma, para beneficiários do produto C, é racional ir ao hospital/pronto-atendimento em vez de ir ao núcleo, uma vez que, no melhor dos casos, o valor despendido em coparticipação é o mesmo e a quantidade de deslocamentos para consultas também.

Outros possíveis fatores para este fenômeno são atinentes às características do grupo do produto C, tais como: (a) características da atividade econômica (visto que a maior parte é do segmento empresarial), como, por exemplo, se este grupo contém mais empresas associadas a indústrias pesadas; (b) cidades onde os beneficiários estão mais concentrados, ou seja, se há disponibilidade adequada de prestadores para a população; (c) tempo de espera para agendamento; entre outros possíveis.

Esse fenômeno pode, em parte, explicar o porquê da elasticidade-preço verificada ser menor em módulo para o produto C, pois beneficiários com alta coparticipação buscarão maneiras mais eficientes, ao passo que beneficiários com coparticipação baixa talvez não sejam sensíveis a ponto de traçar uma estratégia dessas. Contudo, um aprofundamento neste tema foge do escopo do presente trabalho.

6 Conclusão

Os resultados obtidos pelo estudo estão em conformidade com o que fora apontado por Deb e Trivedi (2002), isto é, que a estimação do processo de tomada de decisões em duas etapas – a refletir a dinâmica do problema agente-principal – não implica, necessariamente, em um enquadramento econométrico superior, uma vez que o modelo NB2 foi o modelo com melhor ajuste, tanto para os dados empregados neste estudo quanto para a base de dados empregada pelos autores mencionados. Não obstante o argumento de Grossman (1972) tenha ganhado força com os resultados apresentados, a literatura não produziu argumentos suficientes para indicar se uma abordagem aceita ou descarta a estrutura de agente-principal teoricamente suposta.

Um problema que ocorreu na estimação foi a qualidade de ajuste das frequências estimadas (verificado na tabela 6). Isto se deve à indisponibilidade de algumas variáveis cruciais para a determinação da intensidade do uso de recursos – elencadas ao longo do trabalho –, que esboçam a condição de saúde dos beneficiários e que estavam disponíveis em estudos que empregaram a base do RAND. Por consequência, as frequências previstas na literatura mostram-se condizentes com as observadas.

No entanto, a despeito da indisponibilidade de algumas variáveis, a estimativa da elasticidade – que é o cerne deste trabalho – está muito próxima da aferida pelo estudo exposto em Cameron e Trivedi (2005), que aponta uma elasticidade estimada de -0,0504, contra -0,05687.

Este estudo pode ser extrapolado para aplicações em outros segmentos da saúde, cuja natureza da variável dependente é discreta. É esperado que os agentes possuam elasticidades distintas para serviços distintos. Ao segmentar isto, é possível determinar coparticipações ótimas para coibir o excesso de utilização em setores que apresentam esta característica.

7 REFERÊNCIAS

ARROW, K. J. **Aspects of the theory of risk bearing**. The Theory of Risk Aversion. Helsinki: Yrjo Jahnssonin Saatio. Reprinted in: Essays in the Theory of Risk Bearing, Markham Publ. Co., Chicago, 1971, 90–109, 1965.

ARROW, K.J. **Uncertainty and the welfare economics of medical care**. In: American Economic Review, LIII (5), p. 941-973, 1963.

ARROW, K., **The economics of agency**, in: J. Pratt and R. Zeckhauser, eds., Principals and Agents: The Structure of Business (Harvard Business School Press, Cambridge, MA) 37–51, 1985.

BELSLEY, D.; KUH, E.; WELSH, R. **Regression diagnostics: identifying influential data and sources of collinearity.** New York: John Wiley and Sons, 1980.

BERKSON, J. **Application of the logistic function to bio-assay**. Journal of the American Statistical Association 39, p. 357–365, 1944.

BIRCH, M.W. **Maximum likelihood in three-way contingency tables**. Journal of the Royal Statistical Society B25, p. 220–233, 1963.

BLISS, C.I. **The calculation of the dosage-mortality curve**. Annals of Applied Biology 22, p. 134–167, 1953.

BLOMQVIST, A.G. (1997), **Optimal non-linear health insurance**, Journal of Health Economics 16(3):303–321.

BREUSCH, T.S.; PAGAN, A.R. **A simple test for heterocedasticity,** Econometrica, 47, 1287-1294, 1979.

CAMERON, A. C.; TRIVEDI, P.K. **Regression analysis of count data.** Econometric Society Monographs. N. 30. New York: Cambridge University Press, 1998.

CAMERON, A. C.; TRIVEDI, P.K. **Microeconometrics: Methods & Applications**, New York: Cambridge University Press, 2005.

CRAWFORD, I.; LAISNEY, F.; PRESTON, I. **Estimation of household demand systems with theoretically compatible Engel curves and unit value specifications.** Journal of Econometrics (114): p. 221-241, 2003.

CHANDRA, A.,; GRUBER, J.; MCKNIGHT, R.; **Patient cost-sharing and hospitalization off sets in the elderly.** American Economic Review, 100(1): p. 1-24, 2010.

CHANDRA, A.,; GRUBER, J.; MCKNIGHT, R.; **Cost sharing in low income populations.** The American Economic Review, Vol. 100, No. 2, Papers and Proceedings of the One Hundred Twenty Second Annual Meeting of the American Economic Association (May 2010), p. 303-308, 2010.

CUTLER, D.; ZECKHAUSER, R. **Adverse selection in health insurance.** Forum for Health Economics & Policy, v.1, n.1. California: Berkeley Electronic Press, 1998.

CROWDER, M.J. **Maximum likelihood estimation for dependent observations.** Journal of the Royal Statistical Society B, 38, p. 45-53.

de MEZA, D. **Health insurance and the demand for medical care**. Journal of Health Economics, 2 (1). pp. 47-54, 1983

DEATON, A. **Estimation of own and cross-price elasticities from household survey data**. Journal of Econometrics 36(1): p. 7-30, 1987.

DEATON, A. **Quality, quantity, and spatial variation of price.** American Economic Review 78(3): p. 418-430, 1988.

DEATON, A. **Household survey data and pricing policies in developing countries**. The World Bank Economic Review 3(2): p.183-210, 1989.

DEATON, A. **Price elasticities from survey data: extensions and Indonesian results**. Journal of Econometrics 44(3): p. 281-309, 1990.

DEB, P.; TRIVEDI, P.K.. **Demand for medical care by elderly: a finite mixture approach.** Journal of Applied Econometrics, v.12, n. 3, p. 313-36, 1997.

DEB, P.; TRIVEDI, P.K. **The Structure of Demand for Health Care: Latent Class versus Two-Part Models** Journal of Health Economics, 21, p. 601–625, 2002.

ENGLE, R.F. **Wald, likelihood ratio and lagrange multipliers tests in econometrics.** Handbook of Econometrics, v. II, Amsterdam, 1984.

Feldstein, M.S., **Hospital cost inflation: a study of nonprofit price dynamics**, American Economic Review 60:853–872., 1971.

FIRTH, D. **Generalized linear models.** Statistical Theory and Modelling: In Honor of Sir David Cox, FRS, 55 -82, Londres, Inglaterra.

FISHER, R.A. **On the mathematical foundations of theoretical statistics**. Philosophical Transactions of the Royal Society, 222, 309–368, 1922.

FISHER, R.A. **Two new properties of mathematical likelihood**. Proceedings of the Royal Society, A 144, 285–307, 1934.

GIBSON, J.; KIM, B.; ROZELLE, S. **An Empirical Test of Methods for Estimating Price Elasticities from Household Survey Data**. Disponível em http://iis-

db.stanford.edu/pubs/21678/estimating_price_elasticities_from_household_survey_data.pdf. Acesso em dezembro de 2014.

GREENE, W. H. **Econometric Analysis**, 5th ed. Prentice Hall, 2003.

GREENWOOD, M.; YULE, G.U. (1920) **An enquiry into the nature of frequency distributions representative of multiple happenings with particular reference of multiple attacks of disease or of repeated accidents**. J R Statist Soc 83: 25–279

GROSSMAN, Michael. **On the concept of health capital and the demand for health.** The Journal of Political Economy, v.80, n. 2, 1972.

HILL, C.; ADKINS, L. **Collinearity.** A Companion to Theoretical Econometrics. Oxford, 2001.

LINDSEY, J.K. (1997). **Applying generalized linear models**. Springer-Verlag, New York.

MACKINNON, J.; WHITE, H. **Some heterocedasticity consistent covariance matrix estimators with improved finite sample properties.** Journal of Econometrics, 19, p. 305-325, 1985.

MANNING, W.G; NEWHOUSE, J. P.; DUAN, N.H.; KEELER, E.B.; LEIBOWITZ, A. **Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment**. American Economic Review, 77(3): p. 251-77, 1987.

MARSHALL, A. **Principles of cconomics.** 8ª Edição, Nova York, Macmillan Co., 1961.

MCCULLAGH, P. NELDER, J.A. (1989). **Generalized linear models**, 2ed., Chapman and Hall, London.

MCGUIRE, T.G. **Physician agency**, Handbook of Health Economics cap. 9, Elsevier, Amsterdam, 2000.

MULLAHY, J. (1986). **Specification and testing of some modified count data models**. Journal of Econometrics, 33, 341–365.

NELDER, J.A. and WEDDERBURN, R.W.M. (1972). **Generalized linear models**, Journal of the Royal Statistical Society, A, 135 370–384.

NEWEY, W.K.; MCFADDEN, D. **Large sample estimation and hypothesis testing.** Handbook of Econometrics, v. 4, Amsterdam, North-Holland.

NEWHOUSE, J.P.; PHELPS, E. C.; MARQUIS, M. S. **On having your cake and eating it too: econometric problems in estimating the demand for health services.** Journal of Econometric. Vol. 13, p. 365

NEWHOUSE, J.P.; PHELPS, C.E., **New estimates of price and income elasticities of medical care services**, The Role of Health Insurance in the Health Services Sector (National Bureau of Economic Research, New York, c. 7, p. 261-313, 1976.

PHELPS, C.E.; NEWHOUSE, J.P. **Effects of coinsurance of demand for physician services**, RAND, Research Paper Series, No. R-976-OEO, 1972.

PLATÃO. A República. Tradução de Anna Lia A. A. Prado. São Paulo: Martins Fontes, 2006.

POHLMEIER, W.; ULRICH, V. **An econometric model of the two-part making process in the demand for health care.** Journal of Human Resources, v. 30, p. 339-361, 1995.

ROSETT, R.N.; HUANG, L. **The effect of health insurance on the demand for medical care**, Journal of Political Economy 81, p. 281-305, 1973.

SMITH, A. **Investigação sobre a natureza e a causa da riqueza das nações**. Coleção Os Economistas Ed. Abril, 1997.

VARIAN, Hall R.. **Microeconomia – princípios básicos**. Rio de Janeiro: Campus, 7a edição, 2006.

WANG Y,; HUNT K,; NAZARETH, I, et al. **Do men consult less than women? An analysis of routinely collected UK general practice data**. 2013. Disponível em: http://bmjopen.bmj.com/content/3/8/e003320.full

WHITE, H. **A heterocedasticity-consistent covariance matrix estimator and a direct test for heterocedasticity.** Econometrica, 48, p. 817-838, 1980.

WHITE, H. **Maximum likelihood estimation of misspecified models.** Econometrica, 53, p 1-16, 1982.

WOOLDRIDGE, J. M. **Introdução à econometria: Uma abordagem moderna**. São Paulo: Thomson, 2005.

ZECKHAUSER, R.. **Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives**, Journal of Economic Theory 2(1): p. 10-26, 1970.

ZWEIFEL, P. **Supplier induced demand in a model of physician behavious.** *In*: Health, Economics, and Health Economics, p. 245-67, 1981.