2
Metodologia

2.1
Introdução

Neste capítulo, apresentaremos, no nível de detalhe que se fizer necessário, as metodologias utilizadas nesse trabalho, as ferramentas, técnicas e de que maneira nos apropriamos delas e de seus elementos para a obtenção dos resultados da pesquisa.

Na seção 2.2, portanto, faremos uma exposição sobre a regressão logística com resposta binária, técnica utilizada para a obtenção dos resultados mais importantes neste trabalho, descrevendo e definindo, principalmente, os conceitos de probabilidade estimada, intervalo de confiança para essas probabilidades e razões de chance entre as categorias de uma variável. A regressão logística será utilizada para modelar os dados, afim de se investigar o fenômeno da assimetria de informação e de se identificar o que se pode chamar de grupos de risco no que se refere a atividades laborais.

Na seção 2.3 será exposta a Teoria da Resposta ao Item, que, em sua versão não paramétrica, nos auxiliou na construção das variáveis de escala de morbidade e escala de mobilidade física. A primeira será utilizada em uma análise exploratória dos dados, tentando entender os grupos mais prejudicados em termos de estado de saúde, e a segunda, depois de dicotomizada, servirá como variável resposta nos modelos de regressão logística citados no parágrafo anterior.

Na seção 2.4, fica o registro teórico do coeficiente de incerteza, técnica utilizada em alguns momentos da pesquisa, que nos ajudou a entender melhor determinados resultados.

Na seção 2.5, discorremos brevemente sobre o conceito de “Assimetria de Informação”, elegendo as duas formas aqui consideradas, dentre as quais ela pode ser explorada, a saber, “seleção adversa” e “risco moral”. A “Assimetria de Informação” é um fenômeno a ser investigado aqui, através de modelos de regressão logística.

Nas seções 2.6 e 2.7, respectivamente, explicaremos em detalhes a modelagem aplicada aos dados em suas duas abordagens, que denominamos aqui de “Diagnóstico-PS” e “Diagnóstico-MT”, além de seus respectivos interesses de investigar a assimetria de informação e de identificar os grupos de risco nas atividades laborais.

Em ambos os casos, a classe de modelos utilizada é a regressão logística com resposta binária. Na primeira abordagem se pretende investigar se há ou não evidências de Assimetria de Informação, através da identificação ou não de correlação positiva entre a posse de plano privado de saúde e a ocorrência de determinadas doenças, em um estudo similar ao apresentado em Braido e Lins (2006), que fez uso do método de regressão linear para obter seus resultados.

Por ser linear, o método utilizado por Braido e Lins (2006) tem a vantagem de seus coeficientes poderem ser interpretados diretamente, com os parâmetros das variáveis representando exatamente a variação na probabilidade do evento em estudo, no caso, o diagnóstico em doença. Contudo esse método é pouco robusto, à medida que não impõe que essas probabilidades residam entre 0 e 1, como é o caso do método da regressão logística.

Nesta abordagem, será ajustado um modelo para cada uma das doze doenças crônicas consideradas pela PNAD, mais um para cada uma das duas medidas de mobilidade física adotadas. Os modelos serão descritos oportunamente e cada um deles, chamados aqui de modelos “Diagnóstico-PS”, será aplicado à totalidade dos dados (modelo Brasil) e também aos dados referentes aos moradores de cada uma das cinco grandes regiões do país, perfazendo assim, quatorze modelos em seis grandes regiões (Brasil mais cinco), totalizando oitenta e quatro modelos.

Na segunda abordagem, estamos interessados em identificar o que poderiam ser chamados de grupos de risco no que se refere à atividade laboral dos indivíduos, investigando, a probabilidade de ocorrência de diagnóstico positivo em cada uma das doenças, em função da situação no mercado de trabalho (MT), com os modelos aqui chamados de modelos “Diagnóstico-MT”. As variáveis explicativas no foco desta análise são três e dizem respeito ao ramo de atividade econômica do empreendimento no qual o indivíduo trabalha, à ocupação funcional específica do indivíduo em suas atividades e à formalidade no mercado de trabalho, cada uma delas sendo analisadas individualmente em modelos diferentes.

2.2
Regressão Logística com Resposta Binária

A Regressão Logística foi a classe de modelos utilizada em algumas investigações realizadas neste trabalho. Comumente, quando a variável resposta é binária, condição que será observada nesse estudo, essa técnica se propõe a estimar a probabilidade de ocorrência de um evento específico, em função de outras variáveis. Porém, aqui também utilizamos estes modelos com o propósito de se obter uma medida de correlação entre variáveis. Este é o caso em que iremos investigar o fenômeno da "Assimetria de Informação" na contratação de planos privados de saúde, relacionando o diagnóstico em uma série de doenças com a posse ou não desses planos, em suas modalidades individual e privada, controlando pela faixa etária. Ou seja, não procuramos aí uma relação de causalidade entre a posse de planos e a probabilidade de doença, não sugerindo, portanto, que por estarem cobertos por seguros de saúde, os indivíduos tenham maior ou menor probabilidade de adoecer, mas simplesmente, pretendemos descobrir em que casos a variável de diagnóstico de doença está fraca ou fortemente correlacionada com a variável de posse de plano.

Em outra abordagem, pretendemos modelar essas mesmas variáveis de doença, em função de variáveis sobre mercado de trabalho, controlando por sexo, raça e faixa etária. Neste caso sim, o que se pretende é estabelecer relações de causalidade, procurando as posições no mercado de trabalho onde os indivíduos estejam mais propensos a cada uma das doenças estudadas.

2.2.1
O modelo

O objetivo primário da regressão logística é o de estimar parâmetros para variáveis independentes, também chamadas de regressores, com as quais se pretende explicar o comportamento de uma variável resposta. No caso de esta ser binária, assumindo, digamos, "0" e "1", ao receber os valores das variáveis explicativas, o modelo nos retorna, em última instância, uma estimativa para a probabilidade de ocorrência do evento representado nos dados pelo valor "1". Em todos os modelos ajustados neste trabalho, este evento é o diagnóstico positivo em alguma doença crônica, com um modelo para cada doença estudada.

Contudo, a relação estabelecida entre a variável resposta e os regressores não é linear, o que representa uma desvantagem em relação aos modelos lineares, pois a interpretação dos parâmetros não é direta, com o valor de cada um não representando, portanto, a variação na probabilidade estimada, quando se varia o valor da variável explicativa em uma unidade.

A função que descreve a probabilidade através dos regressores, comumente chamada de função logística, e o formato de um modelo logístico de regressão, com, por exemplo, duas variáveis explicativas, X_1 e X_2, e Y como variável resposta, são como se segue:

$$P_i(Y = 1|x_i) = \frac{1}{1 + e^{-\eta_i}}$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Onde x_i é o vetor com os valores das variáveis explicativas para a *i-ésima* observação e P_i é a probabilidade estimada de ocorrência do evento na mesma observação. β_0, β_1 e β_2 são os parâmetros do modelo, com β_0 sendo um intercepto, β_1 e β_2 os coeficientes das variáveis X_1 e X_2, respectivamente, e *i* variando de 1 a *n*, onde *n* é o número de observações. x_{1i} e x_{2i} são os valores da *i-ésima* observação para estas variáveis e η_i é o preditor linear.

Da função logística deriva-se a característica não linear da regressão logística, mas também dela, deriva-se uma grande vantagem destes modelos em relação aos lineares: por construção, os valores estimados para as probabilidades residem necessariamente entre zero e um, o que é conveniente, dado que estamos estimando probabilidades.

No caso específico de termos uma variável explicativa categórica, com *k* categorias de resposta, esta será representada no preditor linear por $k - 1$ variáveis *dummy*, com $k - 1$ parâmetros e a categoria omitida será tida como categoria basal. Os efeitos dos casos basais de todas as variáveis categóricas serão então acumulados no intercepto, que aqui chamamos de β_0. Ou seja, o intercepto pode ser interpretado como o efeito na probabilidade de ocorrência do evento em estudo para os indivíduos classificados como caso basal em todas as variáveis

categóricas do modelo. Ainda que este efeito não seja linear, esta é uma maneira interessante de se interpretar β_0.

2.2.2
Intervalos de Confiança das Probabilidades

Os resultados dos modelos que relacionam o diagnóstico em doenças com a posse ou não de planos, controlando por faixa etária, serão apresentados na forma das estimativas das probabilidades e, através de um procedimento apresentado em Hosmer e Lemeshow (2000), iremos calcular e apresentar também seus intervalos de confiança de 95%.

Esse procedimento se baseia primeiramente na estimativa da variância do preditor linear e, em seguida, no cálculo de seu intervalo de confiança. Os limites inferior e superior deste intervalo são tidos então, como dados de entrada na função logística. Com isso, chegamos aos limites inferior e superior do intervalo de confiança para a probabilidade, no caso, do diagnóstico positivo na doença.

A fórmula apresentada em Hosmer e Lemeshow (2000) para o cálculo da estimativa da variância do preditor linear é como se segue:

$$Var(\hat{\eta}) = \sum_{i=0}^{p} x_i^2 Var(\hat{\beta}_i) + \sum_{i=0}^{p-1} \sum_{j=i+1}^{p} 2x_i x_j Cov(\hat{\beta}_i, \hat{\beta}_j)$$

Onde $p + 1$ é o número de parâmetros no modelo.

A fórmula para o cálculo dos limites do intervalo de confiança de 95% para η já é amplamente explorada pela literatura nessa área, mas nós apresentamos aqui novamente:

$$IC_{lim\,inf}(\eta) = \hat{\eta} - 1{,}96\sqrt{Var(\hat{\eta})}$$
$$IC_{lim\,sup}(\eta) = \hat{\eta} + 1{,}96\sqrt{Var(\hat{\eta})}$$

Onde 1,96 é o valor crítico da normal padrão, para uma confiança de 95%, e $IC_{lim\,inf}(\eta)$ e $IC_{lim\,sup}(\eta)$ são, respectivamente, os limites inferior e superior do intervalo de confiança para η. Estes são inseridos na fórmula da função logística

e obtemos assim os limites do intervalo de confiança de 95% para a probabilidade de ocorrência do evento em estudo.

Todo esse procedimento para o cálculo dos intervalos de confiança das probabilidades foi, neste trabalho, programado em ambiente Visual Basic do Microsoft Excel. Os códigos do programa encontram-se em anexo.

2.2.3
As Interações no Modelo

As interações entre regressores em um modelo permitem que se admita a hipótese de que uma variável não se comporte de maneira igual dentre os indivíduos classificados por diferentes categorias de outra variável.

Ilustrando essa afirmação, consideremos um modelo com duas variáveis explicativas, uma sendo categórica com três níveis, representada pelas *dummies* X_1, e X_2, e outra sendo numérica, representada por X_3, sem interação entre elas. Como o vetor de *dummies* (X_1, X_2) está representando três níveis, ele pode assumir os valores (0;0), para o caso basal, (1;0) e (0;1) para os outros dois níveis. O preditor linear, para cada um dos níveis da variável categórica, fica descrito das seguintes maneiras:

$$\eta_{basal} = \beta_0 + \beta_3 X_3$$
$$\eta_{nível\,2} = \beta_0 + \beta_1 + \beta_3 X_3$$
$$\eta_{nível\,3} = \beta_0 + \beta_2 + \beta_3 X_3$$

Encarando η como uma função de primeiro grau de X_3, com o gráfico, portanto, sendo representado por uma reta, temos, nos três casos, o mesmo coeficiente angular, β_3. A distinção das três equações de reta seria dada, portanto, apenas pelo coeficiente linear. Isto significa que teríamos três retas diferentes para descrever o comportamento do preditor linear, e conseqüentemente da probabilidade estimada, em função da variável X_3, dentre os indivíduos de cada um dos três níveis da variável categórica. Contudo, essas três retas teriam a mesma inclinação!

No caso de um modelo com estas mesmas variáveis e com interação entre elas, a equação geral do preditor linear ficaria da seguinte forma:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3$$

Essa equação, para cada um dos níveis da variável categórica assumiria, portanto, as formas descritas abaixo:

$$\eta_{basal} = \beta_0 + \beta_3 X_3$$
$$\eta_{nível1} = \beta_0 + \beta_1 + \beta_3 X_3 + \beta_4 X_3 = \beta_0 + \beta_1 + (\beta_3 + \beta_4) X_3$$
$$\eta_{nível2} = \beta_0 + \beta_2 + \beta_3 X_3 + \beta_5 X_3 = \beta_0 + \beta_2 + (\beta_3 + \beta_5) X_3$$

Ou seja, as retas que representam o comportamento do preditor linear em função da variável X_3, para cada um dos três níveis da variável categórica, apresentariam coeficientes linear e angular diferentes! Isto permitiria, por exemplo, que uma delas fosse crescente, outra decrescente e outra constante.

Uma alternativa natural para as interações, quando não se pretende desprezar a hipótese de que alguma variável tenha comportamentos diferentes dentre as categorias de outra variável, seria o ajuste de modelos diferentes para cada sub-população definida pelas categorias de uma das variáveis, utilizando as outras como explicativas nesses modelos, que não teriam, portanto, interações entre as variáveis. Esta situação, obviamente, induziria à obtenção de equações diferentes para cada uma das categorias da variável que define as sub-populações para as quais os modelos seriam ajustados.

Esta solução é particularmente interessante, quando se pretende trabalhar com o conceito de razões de chance, objeto de discussão da seção 2.2.4, e que tem sua interpretação facilitada na ausência de interações.

2.2.4
Razões de Chance

Outro conceito da regressão logística explorado nesse trabalho é o de “Razão de Chance” (RC). Este conceito será empregado para apresentar os resultados dos modelos que buscam identificar os grupos de risco em atividades laborais, relacionando as variáveis de mercado de trabalho com as variáveis de diagnóstico de doença, controlando por sexo, raça e faixa etária.

As razões de chance possuem uma interpretação bem simples e interessante. Em se tratando de uma variável explicativa categórica, elas representam o quão maiores são as chances de ocorrência da condição em estudo dentre os indivíduos de uma determinada categoria, em relação às chances de ocorrência da mesma condição dentre os indivíduos da categoria tomada como base.

Primeiramente se calcula as chances de ocorrência da condição dentre os indivíduos de cada categoria, definidas como a divisão entre a probabilidade de ocorrência desta condição, P_i, como já definida neste capítulo, pela probabilidade de não ocorrência, dada por 1 – P_i como abaixo:

$$1 - P_i = 1 - \frac{1}{1+e^{-\eta_i}} = \frac{1+e^{-\eta_i}}{1+e^{-\eta_i}} - \frac{1}{1+e^{-\eta_i}} = \frac{e^{-\eta_i}}{1+e^{-\eta_i}}$$

A divisão de P_i por 1 – P_i fica então, assim:

$$\frac{P_i}{1-P_i} = \frac{\frac{1}{1+e^{-\eta_i}}}{\frac{e^{-\eta_i}}{1+e^{-\eta_i}}} = \frac{1}{e^{-\eta_i}} = e^{\eta_i}$$

Para exemplificar, consideremos o caso de um modelo com duas variáveis explicativas categóricas de dois níveis, X_1, e X_2, ambas assumindo valores “0” ou “1”, com “0” sendo o caso basal. Considerando que os dados não apresentam interações entre as duas variáveis, teremos um parâmetro sendo ajustado para as categorias “1” de cada uma das variáveis, mais um intercepto representando o efeito na probabilidade de ocorrência da condição para os indivíduos classificados com “0” nas duas variáveis. Portanto, o preditor linear fica como se segue:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Sendo assim, a razão de chance para a categoria “1” da variável X_1 fica:

$$RC = \frac{chances\ de\ ocorrência\ da\ condição\ na\ categoria\ 1}{chances\ de\ ocorrência\ da\ condição\ no\ caso\ basal}$$

$$RC = \frac{e^{\beta_0+\beta_1+\beta_2 x_{2i}}}{e^{\beta_0+\beta_2 x_{2i}}} = e^{\beta_1}$$

Ou seja, em relação ao caso basal, a razão de chance para uma determinada categoria de uma variável que não apresente interações com as outras no modelo, é dada pelo exponencial do parâmetro ajustado para esta categoria.

Dizemos então que as chances de ocorrência da condição para os indivíduos classificados com "1" pela variável X_1 é e^{β_1} vezes maior que as chances de ocorrência da condição para os indivíduos no caso basal desta mesma variável.

Contudo, só se pôde chegar a essa conclusão, graças ao cancelamento, durante o desenvolvimento da equação da razão de chance, de alguns termos do preditor linear, o que não teria acontecido se o modelo considerasse interações entre variáveis.

2.3
A Teoria da Resposta ao Item

A metodologia da Teoria da Resposta ao Item propõe modelos que representam a relação entre um construto ou traço latente de um indivíduo e a probabilidade deste indivíduo produzir uma determinada resposta a um item de um questionário. Tanto a versão paramétrica quanto a não paramétrica começam com a definição da função resposta ao item como a probabilidade de acerto no item, dado que o respondente tem, digamos, θ como valor no traço latente. No caso particular de itens com respostas dicotômicas, pode-se definir formalmente a função de resposta ao item *i*, $P_i(\theta)$, como:

$$P_i(\theta) = P(X_i = 1|\theta)$$

Isto significa que o indivíduo com característica θ tem probabilidade $P_i(\theta)$ de acertar o item *i*, considerando que "0" representa uma resposta errada e "1" representa uma resposta correta.

Neste trabalho, estamos interessados no valor θ do indivíduo, que vem a ser o valor atribuído a ele pela escala construída, via TRI, para o traço latente. Portanto, vamos aqui lançar mão da versão não paramétrica desta teoria para a obtenção das escalas de morbidade e de mobilidade física, a serem usadas,

respectivamente, em uma análise exploratória dos dados e como variável resposta para modelos de regressão logística.

As duas versões da teoria se distinguem a partir da definição da função resposta, na medida em que a paramétrica impõe uma fórmula algébrica com parâmetros bem definidos, enquanto que a outra não, exigindo apenas que $P_i(\theta)$ seja monotonamente não decrescente em θ, ou seja, $\theta_a < \theta_b \rightarrow P_i(\theta_a) \leq P_i(\theta_b)$. Esta condição de monotonicidade da função resposta também é a que caracteriza o chamado modelo de homogeneidade monótona, que vamos aplicar neste trabalho.

Além desta, duas condições mais são fundamentais para o modelo, a saber, a unidimensionalidade e a independência local estocástica. A primeira afirma que todos os itens no teste medem o mesmo traço latente. Numa interpretação psicológica, isto significa dizer que todos os itens medem uma mesma habilidade, por exemplo, como resolver frações, ou uma mesma atitude, como por exemplo, a opinião sobre temas controversos. Já numa interpretação matemática, isto significa que necessitamos apenas de um traço latente para a descrição da estrutura dos dados.

A condição de independência local estocástica afirma que a resposta a um determinado item não é afetada pelos outros itens, o que pode soar estranho a princípio, já que se o valor θ do traço latente de um indivíduo é fixo ao longo dos itens e, digamos, relativamente alto, significa que a probabilidade de "acerto" em qualquer item é também alta. Na verdade, essa condição estabelece justamente que este valor θ seja fixo ao longo dos itens do teste, evitando que aconteçam coisas como o que se pode chamar de aprendizado na prática, tornando o indivíduo mais apto a responder uma questão após responder outra, ou ainda, que em questões mais polêmicas ou controversas, o indivíduo dê uma resposta mais socialmente aceita, mesmo que não reflita sua opinião. Em ambos os casos, o valor θ estaria variando de um item para outro.

A qualidade do ajuste do modelo depende do grau com que os dados satisfazem essas condições. Um importante conceito envolvido nessa medida foi introduzido pelo modelo de resposta de Guttman e é o que se chama de "erro de Guttman", definido pela ocorrência de casos onde, dados X_i e X_j, dois itens em um questionário, com $P_i(\theta) < P_j(\theta)$, um indivíduo de valor θ para o traço latente responde corretamente X_i e incorretamente X_j. Nestes casos, dizemos que o

indivíduo acertou o item mais difícil, aquele com menor probabilidade de acerto, e errou o mais fácil, aquele com maior probabilidade de acerto.

Chamemos de P_i e P_j, as probabilidades de acerto nos itens *i* e *j* separadamente e de P_{ij} a probabilidade de acerto simultâneo nos mesmos itens, com $P_i < P_j$ (item *i* mais difícil que item *j*). Estes parâmetros podem ser estimados por meio das médias amostrais, com, por exemplo, P_{ij} sendo a razão entre o número de indivíduos que acertaram os dois itens simultaneamente e o número total de indivíduos. Com isto, podemos resumidamente definir o coeficiente de escalonabilidade H_{ij} do par de itens *i* e *j* como sendo a medida de quanto este par de itens se afasta do escalograma perfeito de Guttman (conjunto de dados onde não se observa a ocorrência de erros de Guttman), da seguinte maneira:

$$H_{ij} = \frac{Cov(X_i, X_j)}{Cov_{max}(X_i, X_j)}$$

Onde $Cov(X_i, X_j)$ é a covariância na população, dada por $Cov(X_i, X_j) = P_{ij} - P_i P_j$, e $Cov_{max}(X_i, X_j)$ é a mesma covariância, sob a condição de ausência completa de erros de Guttman, onde todos os indivíduos que acertam o item mais difícil, acertam também o mais fácil, o que significa dizer $P_i = P_{ij}$. Desta maneira, temos que $Cov_{max}(X_i, X_j) = P_i - P_i P_j$, e que:

$$H_{ij} = \frac{P_{ij} - P_i P_j}{P_i - P_i P_j}, 0 \leq H_{ij} \leq 1$$

Em última instância, o coeficiente H_{ij} vai dar origem ao coeficiente H de escalonabilidade de todo o teste, que em função dos parâmetros amostrais e considerando que a escala inclui *k* itens, é dado por:

$$H = 1 - \frac{\sum_i^k \sum_{j>i}^k P_i - P_{ij}}{\sum_i^k \sum_{j>i}^k (P_i - P_i P_j) + \sum_i^k \sum_{j<i}^k (P_j - P_i P_j)}, 0 \leq H \leq 1.$$

São portanto, testados modelos diferentes, com conjuntos de variáveis diferentes, e o modelo de melhor ajuste é o que possui o coeficiente H mais elevado.

2.4
Uncertainty Coefficient ou "Coeficiente de Incerteza"

O chamado *uncertainty coefficient*, ou "coeficiente de incerteza", é um método que utiliza uma medida de variação proposta por *Theil*, em 1970, para estudar a relação de dependência entre duas variáveis categóricas em uma tabela de contingência.

Considerando Y como variável resposta, com categorias j = 1,...,J, e X como variável explicativa, com categorias i = 1,...,I, a fórmula para o coeficiente de incerteza é como se segue:

$$U = -\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij}ln\left(\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}\right)}{\sum_{j=1}^{J}\pi_{+j}ln\pi_{+j}}$$

Onde:

π_{ij}: probabilidade de se observar os valores *i* para a variável X, e *j* para a variável Y, simultaneamente.

$\pi_{i+} = \sum_{j=1}^{J}\pi_{ij}$: probabilidade de se observar o valor *i* para a variável X.

$\pi_{+j} = \sum_{i=1}^{I}\pi_{ij}$: probabilidade de se observar o valor *j* para a variável Y.

Estes parâmetros podem ser estimados através das médias amostrais.

O coeficiente de incerteza é uma variação do método apresentado em 1954 por *Goodman* e *Kruskal*, o *concentration coefficient*, ou "coeficiente de concentração", também conhecido por *tau de Goodman e Kruskal*. A interpretação dada por eles para esses números é a de que estes representam a redução proporcional na probabilidade de um palpite incorreto para o valor de Y, quando se sabe em que categoria de X nós estamos. Variando de 0 a 1, maiores valores para *U* representam uma dependência maior, com 0 equivalendo a independência total.

A dificuldade desses métodos reside em saber o quão grande precisa ser o valor dos coeficientes para que se possa rejeitar a hipótese de independência, já que eles não estão relacionados a uma função de densidade de probabilidade convencional, como a qui-quadrado, por exemplo. Desta maneira, essas medidas se tornam particularmente interessantes quando estamos tratando de diferentes situações e temos, portanto, a possibilidade de tomar uma como referência e compará-la com as demais. Mas mesmo não sendo este o caso, o "coeficiente de incerteza" ainda consegue dar uma interpretação bem interessante para a relação entre variáveis categóricas.

O leitor interessado encontra referência mais profunda em Agresti (1990).

2.5
A Assimetria de Informação

No presente contexto, entende-se por assimetria de informação, a diferença entre os níveis de acesso a informação, por parte dos diferentes agentes envolvidos em uma transação, como a de um contrato de seguro de saúde.

A seleção adversa, uma das modalidades de assimetria de informação, se manifesta quando a seguradora, sem saber, possui em seu portfólio de segurados, indivíduos que omitem o fato de serem portadores de determinados problemas crônicos. O segurado em questão, consciente de seu problema, busca a cobertura de um plano privado para assegurar o pagamento de um tratamento ou procedimento mais caro que por ventura se faça necessário, mas não fornece toda a informação a respeito de seu estado de saúde, sabedor de que se assim o fizer, provavelmente lhe será cobrado um prêmio mais elevado.

Justamente por esse aspecto da iniciativa do segurado em buscar um plano privado de saúde, se considera que a seleção adversa ocorre fundamentalmente em contratos de seguro individuais, ou seja, não empresariais, já que esta última modalidade, institucionalizada, normalmente vem embutida em um pacote de benefícios oferecido pelo empregador. Neste caso, se poderia dizer que o segurado estaria assumindo uma postura passiva na contratação do seguro, desconfigurando a iniciativa de buscar a cobertura do plano de saúde.

Já o risco moral, outra modalidade de assimetria de informação, se caracteriza pelo comportamento de risco de indivíduos, que, tendo a consciência

de estarem segurados por planos privados de saúde, deixam de tomar medidas preventivas e a acabam por requisitarem mais vezes os médicos, laboratórios e/ou hospitais credenciados pelo seu plano. Em uma analogia com o seguro de carro, seria como se o segurado, por estar coberto, não se preocupasse em estacionar seu carro em locais considerados perigosos ou dirigisse sem cuidado, sabendo que, em caso de sinistro, o prejuízo será coberto pelo seguro.

Conscientes desse movimento, muitas operadoras de planos de saúde instauram programas de saúde preventiva dentre seus segurados, incentivando-os a hábitos mais saudáveis e diminuindo assim a probabilidade de eventuais gastos com consultas, exames, cirurgias, etc. que poderiam ser evitados. Outra estratégia adotada pelas seguradoras é a instituição de co-pagamento, inibindo o uso desnecessário de serviços de saúde.

Um dos objetivos deste trabalho é investigar indícios de *onde* e *como* esses fenômenos podem acontecer, minimizando a diferença entre os níveis de acesso à informação no mercado de planos privados de saúde no Brasil.

2.6
Os Modelos Diagnóstico-PS

Estes modelos relacionam a variável binária de ocorrência de diagnóstico positivo de doença, onde 1 representa “presença da doença” e 0 representa “ausência da doença”, com as seguintes variáveis:

- Posse de plano de saúde (gerada a partir da combinação das variáveis v1332 e v1321, originais dos dados da PNAD-2003).
- Faixa etária (gerada a partir da variável v8005, original dos dados da PNAD-2003).

A variável de posse de plano de saúde é categórica de três níveis, a saber: “sem plano”, “com plano individual” e “com plano empresarial”, onde “sem plano” é o caso basal. Naturalmente, já que se trata de modelagem envolvendo questões de saúde, prevemos também o controle pela idade através de variáveis *dummy* para as faixas etárias de “até 17 anos”, “de 18 a 29 anos”, “de 30 a 39 anos”, “de 40 a 49 anos”, “de 50 a 59 anos”, “de 60 a 69 anos” e “70 ou mais anos”, em compatibilidade com as faixas pelas quais a legislação brasileira de

planos de saúde permite que se faça distinção de prêmios. A idade do indivíduo está registrada na variável v8005 da PNAD-2003.

A variável de posse/tipo de plano de saúde foi gerada a partir das informações sobre a origem de seu financiamento, encontradas nos dados da PNAD-2003 sob a forma da variável de nome "v1332". Definiu-se como empresarial, os planos de saúde financiados "somente pelo empregador do titular", "pelo titular através do trabalho atual" ou "pelo titular através do trabalho anterior". Por sua vez, foram classificados como individuais, os planos financiados "pelo titular diretamente ao plano", "por outro morador do domicílio" ou "por não morador". Esta é a mesma classificação de planos de saúde utilizada em Braido e Lins (2006).

O modelo de regressão logística, para cada região ou Brasil como um todo, vai, portanto, estimar a probabilidade de ocorrência de diagnóstico positivo de doença em função da condição ou não de posse de plano de saúde, e de qual tipo, controlando pela faixa etária do indivíduo. Para isso o preditor linear vai conter, além do intercepto, dois coeficientes para as três categorias da variável de posse de plano, onde "sem plano" é o caso basal, seis coeficientes para as sete faixas etárias, onde "até 17 anos" é o caso basal e mais doze coeficientes representando as interações entre as duas variáveis, num total de vinte e um parâmetros. O modelo fica definido, portanto, da seguinte maneira:

$$Y = \begin{cases} 1, & com\ P_i(Y = 1|x_i) \\ 0, & com\ 1 - P_i(Y = 1|x_i) \end{cases}$$

Onde:

$$P_i(Y = 1|x_i) = \frac{1}{1 + e^{-\eta_i}}$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + + \beta_{20} x_{20i}$$

Neste caso, Y é a variável de diagnóstico de doença em questão e $x_i = (x_{1i}; x_{2i}; \ldots; x_{20i})$ representa o vetor com os valores, na *i-ésima* observação, de cada uma das *dummies* utilizadas no modelo (duas para a variável de plano de saúde, seis para a variável de faixa etária e mais doze para as interações entre

essas variáveis); $\hat{\beta}_k$, com *k* variando de 0 a 20, representa o valor ajustado pelo modelo para o *n-ésimo* parâmetro.

Através do procedimento descrito na seção 2.2, e apresentado em mais detalhes em Hosmer e Lemeshow (2000), serão calculados, também, os intervalos de confiança de 95% para essas probabilidades.

O pacote estatístico utilizado para o ajuste do modelo foi o STATA 10 e o cálculo dos intervalos de confiança para as estimativas das probabilidades foi programado no ambiente Visual Basic do Microsoft Excel. Não se observou, em determinadas grandes regiões, a ocorrência de indivíduos de determinados perfis que tivessem determinadas doenças. Sendo assim, não foi possível a estimação dos parâmetros para as interações nesses casos, e os resultados não serão apresentados.

Importante esclarecer que não estamos preocupados aqui com a capacidade preditiva do modelo para probabilidade de ocorrência de diagnóstico positivo das doenças, em função da cobertura ou não por planos de saúde, muito menos sugerir que há uma relação de causalidade entre as variáveis explicativas e a variável resposta. O que se pretende é simplesmente utilizar as saídas do modelo como uma medida de correlação entre as variáveis do preditor linear, mesmo que este assim o seja chamado, e a variável dependente, tornando possível a análise dos fenômenos descritos pela assimetria de informação.

Dessa maneira, sob a hipótese de que não há diferença relevante entre as coberturas dos planos individual e empresarial (ou seja, a variável abrangência da cobertura do plano não é relevante para o modelo), a assimetria de informação fica caracterizada pela correlação positiva entre a ocorrência da doença e a posse de plano de saúde. Quando esta correlação não faz distinção entre os dois tipos de cobertura, diz-se apenas que há a observância de risco moral. Quando a correlação é mais forte para os indivíduos cobertos por planos individuais, diz-se que há seleção adversa, pois a iniciativa de procurar a cobertura se fez presente.

Esta correlação se torna estatisticamente significante quando o intervalo de confiança para essas probabilidades não se interceptam. Ou seja, diz-se que há a assimetria de informação, sob a forma do risco moral, ao nível de 5%, quando o limite inferior do intervalo de confiança para a estimativa da probabilidade de ocorrência da doença dentre os cobertos por plano de saúde, tanto individual quanto empresarial, é maior do que o limite superior do mesmo intervalo de

confiança para os indivíduos sem plano. E se, além disso, o limite inferior do intervalo de confiança para os indivíduos com plano individual for maior do que o limite superior deste intervalo para os indivíduos com plano empresarial, diz-se que há também seleção adversa.

É possível ainda que esta última ocorra sem que o risco moral se faça presente. Este é o caso quando a estimativa da probabilidade de doença para os cobertos por planos individuais é maior que a estimativa da mesma probabilidade para os sem plano, sem que os respectivos intervalos de confiança se interceptem, não ocorrendo com o mesmo para os indivíduos cobertos por planos empresariais. Na Ilustração 1 encontram-se exemplos dessas situações, e suas respectivas descrições, em uma esquematização gráfica, facilitando o entendimento de cada uma delas.

Desta maneira, os intervalos de confiança serão utilizados como evidências para a rejeição ou não de H_0, em favor de H_a, onde H_0 e H_a, no caso dos planos de saúde individuais, são as seguintes hipóteses formais:

H_0: *A probabilidade de doença para os indivíduos com pano privado individual de saúde e a probabilidade de doença para os indivíduos sem plano privado de saúde são iguais.*

H_a: *Essas probabilidades são diferentes.*

O análogo vale para comparamos as probabilidades de doença para os indivíduos com planos privados empresariais de saúde e para os indivíduos sem plano privado de saúde. Rejeitaremos então H_0, quando os intervalos de confiança da probabilidade de doença para os indivíduos com plano não interceptar o mesmo intervalo para os indivíduos sem plano, admitindo a hipótese de que estas probabilidades são diferentes. Conclui-se portanto que, já que os intervalos de confiança de 95% não possuem pontos em comum, temos evidências suficientes para rejeitar H_0, em favor de H_a.

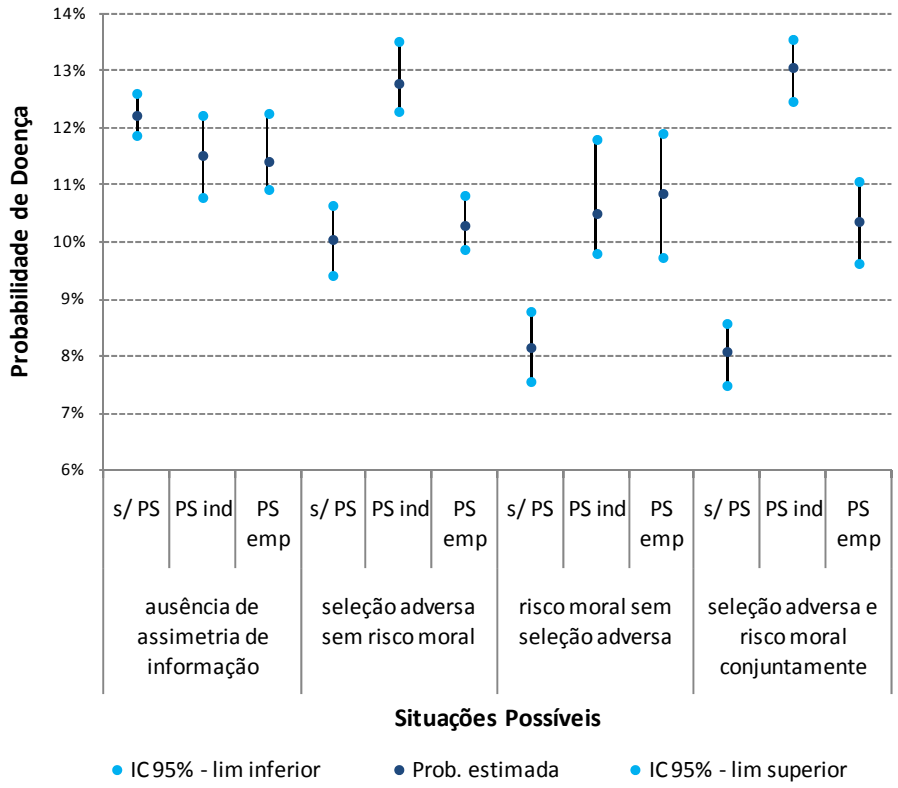

Ilustração 1: Esquema diagramático a ser utilizado para inferir sobre a ocorrência de seleção adversa e/ou risco moral.

2.7
Os Modelos Diagnóstico-MT

Os modelos aqui apresentados relacionam as mesmas variáveis binárias de diagnóstico de doenças, onde 1 representa "presença da doença" e 0 representa "ausência da doença", com as seguintes variáveis:

- Ramo de atividade econômica do empreendimento no qual o indivíduo trabalha (v4809 nos dados da PNAD-2003). Presente apenas nos modelos para estudo de seus efeitos sobre as probabilidades de doença.
- Ocupação funcional do indivíduo em suas atividades (v4810 nos dados da PNAD-2003). Presente apenas nos modelos para estudo de seus efeitos sobre as probabilidades de doença.
- Situação do indivíduo quanto à formalidade no mercado de trabalho (gerada a partir da combinação das variáveis v4706 e v9059, originais dos dados da PNAD-2003). Presente apenas nos modelos para estudo de seus efeitos sobre as probabilidades de doença.

- Faixa etária, como descrita na seção anterior e presente em todos os modelos.
- Sexo (v0302 nos dados da PNAD-2003), presente em todos os modelos.
- Raça (v0404 nos dados da PNAD-2003), presente em todos os modelos.

O controle pelo sexo é feito através de uma variável *dummy* para o caso "feminino", e pela raça, através de quatro variáveis *dummy* para as cinco raças admitidas, a saber, "indígena", "branca", "preta", "amarela" e "parda", em que "branca" é o caso basal. Originalmente, nos dados da PNAD-2003, a variável de raça admite também a categoria "ignorada", porém, por motivos de ausência, para muitas das doenças, e em todas as grandes regiões, de indivíduos assim classificados pela variável e que tivessem a doença em questão, essa categoria não foi considerada pela modelagem e as observações de "ignorada" na variável raça foram excluídas de todas as amostras para ajuste dos modelos. Além desses casos, que foram excluídos a priori, para algumas outras combinações de grande região e doença, também não se observou a ocorrência de respostas positivas para a variável resposta em indivíduos de determinados perfis nas variáveis explicativas. Estas observações foram excluídas automaticamente, pelo próprio programa, das amostras para ajuste dos respectivos modelos. Estes também são exemplos de casos de tabelas de contingência com células vazias, já citados na seção anterior.

Aqui, não consideramos as interações entre as variáveis, sendo assim, além do intercepto, do parâmetro para o sexo feminino, dos seis parâmetros para as faixas etárias e dos quatro parâmetros para as raças, o preditor linear, em cada um dos três casos de caracterização do mercado de trabalho, vai conter portanto, mais tantos parâmetros quantas forem as categorias das variáveis de mercado de trabalho em questão, menos um (caso basal da variável de MT).

Estes modelos ficam então, definidos da seguinte maneira:

$$Y = \begin{cases} 1, & com\ P_i(Y = 1|x_i) \\ 0, & com\ 1 - P_i(Y = 1|x_i) \end{cases}$$

Onde:

$$P_i(Y = 1|x_i) = \frac{1}{1 + e^{-\eta_i}}$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + + \beta_p x_{pi}$$

Neste caso, Y , x_i e η_i são definidos da mesma maneira que na seção anterior, onde se explicou os modelos “Diagnóstico-PS”, mudando apenas o número de parâmetros no modelo, com p sendo igual a 21, 18 e 13, para os casos dos modelos com a variável de ramo de atividade econômica do empreendimento, ocupação funcional do indivíduo e formalidade no mercado de trabalho, respectivamente.

A não ser pelos casos de células vazias nas tabelas de contingência, ou por *missing values* em alguma das variáveis envolvidas, onde as observações são excluídas, as amostras para as quais os modelos serão ajustados são as mesmas utilizadas para o ajuste dos modelos “Diagnóstico-PS”, ou seja, Brasil (totalidade dos dados), mais as cinco grandes regiões do país. Porém, neste caso, serão ajustados três modelos para cada uma das combinações de grande região e doença, para que possamos analisar separadamente os efeitos de cada uma das três variáveis de mercado de trabalho sobre cada uma das doenças, em cada uma das grandes regiões. O total é, portanto, de 254 regressões.

Neste contexto sim, admitimos no modelo, uma relação de causalidade entre as variáveis explicativas e a variável resposta, querendo entender quais as atividades e funções laborais que estão mais propensas a determinadas doenças.

A primeira variável de mercado de trabalho considerada, encontrada na base de dados original da PNAD-2003 com o nome de “v4809”, diz respeito à atividade econômica do empreendimento no qual o indivíduo trabalha. Esta variável é categórica nominal e admite treze categorias, a saber: “agrícola”, “indústria de transformação”, “outras atividades industriais”, “construção”, “comércio e reparação”, “alojamento e alimentação”, “transporte, armazenagem e comunicação”, “administração pública”, “educação, saúde e serviços sociais”, “serviços domésticos”, “outros serviços coletivos, sociais, pessoais”, “outras atividades” e “atividades mal definidas ou não declaradas”. A categoria tomada como caso basal foi “comércio e reparação”, que é a segunda maior em termos de número de observações dentro da variável. A categoria que abrangia o maior

número de trabalhadores é a categoria “agrícola”, porém, não a definimos como caso basal com a intenção de focar o estudo nas áreas urbanas.

A segunda variável de mercado de trabalho foca no indivíduo, considerando as ocupações ou funções especificas exercidas em suas atividades laborais. Encontrada na base original da PNAD-2003 sob o nome “v4810”, essa variável é também categórica nominal e admite as dez seguintes categorias: “dirigentes em geral”, “profissionais das ciências e das artes”, “técnicos de nível médio”, “trabalhadores de serviços administrativos”, “trabalhadores dos serviços”, “vendedores e prestadores de serviços do comércio”, “trabalhadores agrícolas”, “trabalhadores da produção de bens e serviços de reparação e manutenção”, “membros das forças armadas e auxiliares” e “ocupações mal definidas ou não declaradas”. Dentre estas, a escolhida como caso basal foi a categoria “trabalhadores da produção de bens e serviços de reparação e manutenção”, por se tratar da que tinha o maior número de observações (em estatística, este conceito de valor com o maior número de ocorrências é conhecido como moda da variável). O critério de definir, como caso basal, as categorias com grande número de observações, permite que as estimativas de seus parâmetros sejam mais precisas, o que é muito importante, já que o caso basal compõe, juntamente com cada um dos outros parâmetros, o efeito de cada categoria em questão sobre o valor da variável resposta. Em absolutamente todas as outras variáveis categóricas nominais consideradas para modelagem neste trabalho, o critério de escolha do caso basal foi o de tornar a sua comparação com as outras categorias e a interpretação dos resultados a mais adequada possível.

Também por motivos de células vazias nas tabelas de contingência, as categorias “atividades mal definidas ou não declaradas” e “indústria de transformação”, da variável de atividade econômica do empreendimento, e as categorias “ocupações mal definidas ou não declaradas” e “membros das forças armadas, da variável de ocupação funcional do indivíduo, foram desconsideradas e mais observações foram excluídas de todas as amostras para ajuste dos modelos. Sendo assim, nos preditores lineares dos modelos em que forem consideradas, essas variáveis serão representadas, respectivamente, por sete e dez *dummies*, ou seja, o número original de categorias de cada variável, menos 3, em função das categorias excluídas e a do caso basal.

A terceira variável de mercado de trabalho se refere à situação do indivíduo quanto à formalidade ou não no mercado de trabalho. Mais simples, esta variável admite três categorias, “formalidade” (caso basal), “informalidade” e “conta própria ou empregador” e, assim como as outras, é categórica nominal. Esta será representada nos preditores lineares por variáveis dummies para as categorias “informalidade” e “conta própria ou empregador”. Sua construção se deu a partir das variáveis de posição na ocupação exercida pelo indivíduo (“v4706” na PNAD-2003) e de contribuição ou não, através de seu trabalho, com institutos de previdência municipal, estadual ou federal (“v9059” na PNAD-2003). Considerou-se em situação de formalidade, todos aqueles que disseram contribuir com tais institutos ou que, em suas ocupações laborais, assumiam posição de “empregado com carteira”, “militar”, “funcionário público estatutário” ou “trabalhador doméstico com carteira”. Na informalidade foram alocados os que não contribuíam com os institutos de previdência citados e que, em suas ocupações laborais, assumiam posição de “outros empregados sem carteira”, “empregados sem declaração de carteira”, “trabalhador doméstico sem carteira”, “trabalhador doméstico sem declaração de carteira”, “trabalhador na construção para o próprio uso”, “não remunerado” ou “sem declaração - não aplicável”. Já os trabalhadores por conta própria ou empregadores, assim o foram considerados por não contribuírem com esses institutos de previdências e por, em suas ocupações laborais, dizerem assumir posição justamente de “conta própria” ou “empregador”.

Para toda esta classe de modelos, os resultados serão apresentados sob a forma de razões de chance para as diferentes categorias das três variáveis de estudo. Já que as variáveis demográficas estão presentes nos preditores lineares apenas para efeito de controle, e como o interesse é entender a influência da situação no mercado de trabalho sobre as probabilidades de doença, serão apresentados apenas os resultados das razões de chance para as três variáveis foco.

O conceito de razões de chance foi apresentado na seção 2.2.4.