LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO

ECONOMETRIA APLICADA E PREVISÃO

TRABALHO FINAL DE MESTRADO

DISSERTAÇÃO

ANÁLISE DE MODELOS DE REGRESSÃO BINARIA COM EVENTOS RAROS

JOÃO MANUEL FRANCO VIERA

OUTUBRO-2019

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO EM
ECONOMETRIA APLICADA E PREVISÃO

TRABALHO FINAL DE MESTRADO
DISSERTAÇÃO

ANÁLISE DE MODELOS DE REGRESSÃO BINARIA COM EVENTOS RAROS

JOÃO MANUEL FRANCO VIERA

ORIENTAÇÃO:

PROFESSORA DOUTORA ESMERALDA RAMALHO

OUTUBRO-2019

Análise de Modelos de Regressão Binária com Eventos Raros

João Vieira[1]
Orientadora: Profª Drª Esmeralda Ramalho[2]

Resumo

Nesta dissertação serão abordadas duas estratégias para lidar com eventos raros em variáveis dependentes binárias: a seleção amostral com base na variável dependente e a aplicação de modelos flexíveis. Enquanto que na primeira estratégia a proporção de eventos raros é inflacionada artificialmente na amostra usada para estimação, na segunda estratégia utilizam-se modelos cuja capacidade de descrição dos dados é potencialmente superior aos modelos binários standard (tipicamente o *logit* e o *probit*). Para estudar qual das opções será melhor empiricamente foi realizado um estudo de simulação avaliando vários aspetos destes modelos, como a capacidade preditiva e a dificuldade computacional na sua implementação. Neste estudo observou-se que a estimação dos modelos flexíveis era mais instável, apresentando mais dificuldade na sua implementação. Por outro lado, quando se realiza seleção amostral é necessário ter uma boa dimensão amostral para se notarem os efeitos desta seleção na probabilidade de uma observação pertencer ao conjunto dos uns. Tendo em conta os resultados do estudo de simulação, é recomendado utilizar a seleção amostral com correção, para ter em conta a alteração da probabilidade dos eventos. Recomenda-se, adicionalmente, que a amostra utilizada para a estimação contenha entre 20% a 35% de uns para não se perder informação que possa estar contida no zeros, caso a amostra tenha uma dimensão pequena/média, ou 50% caso se tenha uma amostra com uma grande dimensão.

Palavras-Chave: *Choise-Based Sampling*, Eventos Raros, Funções *Link*, *Links* Assimétricos, *Logit*, *Probit*, Seleção Amostral, Simulação.

[1] Contacto: joaovieira@aln.iseg.ulisboa.pt
[2] ISEG Lisbon School of Economics & Management-Departamento de Matemática e REM. Contacto: eramalho@iseg.ulisboa.pt

Análise de Modelos de Regressão Binária com Eventos Raros

João Vieira[3]
Advisor: Profª Drª Esmeralda Ramalho[4]

Abstract

In this dissertation we will address two strategies for dealing with rare events that occur in binary dependent variables: sample selection based on the dependent variable and the application of flexible models. While the first strategy changes the proportion of rare events artificially, the second strategy uses models that, potentially, describe the data more properly than the standard binary models (for example, the *logit* and the *probit*). In order to study which of the options will be empirically better, a simulation study was conducted, evaluating several aspects of this models, such as predictive ability and computational difficulty in their execution. In this study, the estimation of flexible models was more unstable, presenting more difficulty in their implementation. On the other hand, when sample selection is performed, it is required a large sample size to observe the effects of this selection on the probability of an observation belonging to the set of events. Taking into account the results of our simulation, it is recommended to use sample selection with correction, to account for the change in the probability of an event. In addition, it is recommended that the sample used for estimation contains between 20% to 35% of occurrences to avoid loosing information that may be contained in the zeros if the sample has a small/medium size, or 50% if a large sample is available.

Keywords: Assymetric Links, Choise-Based Sampling, Rare Events, Link Functions, *Logit*, *Probit*, Sample Selection, Simulation.

[3] Contact: joaovieira@aln.iseg.ulisboa.pt
[4] ISEG Lisbon School of Economics & Management-Mathematics Department and REM. Contact: eramalho@iseg.ulisboa.pt

Dedicado aos meus pais

Agradecimentos

Gostaria de agradecer à minha orientadora pela sua disponibilidade, por sugerir-me este interessante tema, por todas as correções e conselhos dados ao longo da realização desta dissertação. Quero também agradecer ao corpo docente do mestrado em Econometria Aplicada e Previsão, pela disponibilidade demonstrada em ajudar os alunos e por todo o conhecimento transmitido durante a parte curricular. Por último quero agradecer aos meus pais por toda a paciência e apoio durante os meus estudos.

"There is nothing more difficult to take in hand, more perilous to conduct, or more uncertain in its success, than to take the lead in the introduction of a new order of things."
Niccolo Machiavelli

Conteúdo

Siglas

AMPE Average Marginal Probability Effect.

AUC Area Under The Curve.

BFGS Broyden–Fletcher–Goldfarb–Shanno.

BS Brier Score.

CBS Choice-Based Sampling.

GAM Generalized Additive Models.

GEV Generalized Extreme Value Distribution.

GMM Generalized Method of Moments.

IRLS Iteratively Reweighted Least Squares.

MPE Marginal Probability Effect.

WESML Weighted Exogenous Sample Maximum Likelihood Estimator.

RESET Regression Specification Error Test.

ROC Receiver Operating Characteristic.

SLID Survey of Labour and Income Dynamics.

Capítulo 1

Introdução

If you really want to escape the things that harass you, what you're needing is not to be in a different place but to be a different person.
Lucius Annaeus Seneca

Os eventos raros são dos fenómenos mais difíceis e interessantes para modelar e prever, sendo estudados em várias áreas do conhecimento. Na área das ciências económicas, a sua modelação tem um interesse especial na análise e gestão de risco, onde se pode destacar, por exemplo, o cálculo das probabilidades de incumprimento (*probabilities of default*) para a indústria financeira, ou o cálculo das probabilidades de ruína para a industria seguradora (Rubino e Tuffin, 2009).

Weiss (2004) identificou duas classes de problemas ligados com eventos raros, a raridade absoluta e a raridade relativa. O primeiro conceito diz respeito a casos em que o número de exemplos associados com a classe minoritária é pequeno em sentido absoluto. O segundo conceito diz respeito a casos em que o número de exemplos associados à classe minoritária é pequeno comparativamente com as outras classes (Ogundimu, 2019).

Aqui será abordado o problema da raridade em variáveis binárias pressupondo a raridade relativa, onde a classe dos uns (ocorrências/eventos) será a classe rara e a classe dos zeros (controlos/não ocorrências) será a classe comum. Dentro da raridade relativa em dados binários não existe consenso em relação à percentagem a partir da qual uma classe é considerada rara. Por exemplo King e Zeng (2001) consideram que uma classe é rara quando o número de ocorrências é dezenas e centenas de vezes menor do que a frequência dos zeros. Nesta dissertação adotar-se-á esta definição. No entanto,

os métodos e conclusões que irão ser apresentados continuarão a ser validos no caso de raridade absoluta.

Esta dissertação terá a seguinte estrutura. No capítulo dois será feita a revisão de literatura começando por introduzir a notação e apresentando de forma geral a regressão binomial. Seguidamente apresentam-se os modelos standard, explicam-se os problemas de ignorar o baixo número de ocorrências e apresentam-se duas soluções para ultrapassar os mesmos. A primeira consiste em realizar seleção amostral para obter informação adicional, inflacionando a percentagem de eventos raros na amostra, o que permite melhorar a capacidade preditiva dos modelos e a eficiência dos mesmos. A segunda consiste em utilizar modelos com parâmetros extra. Estes parâmetros extra irão trazer maior flexibilidade, permitindo um melhor ajustamento aos dados.

No capítulo três, apresentar-se-á, em primeiro lugar, a metodologia usada para realizar o estudo de simulação, sendo posteriormente realizada a análise dos resultados obtidos. A metodologia envolverá a aplicação dos estimadores apresentados no capítulo dois aos dados canadenses do Inquérito a Dinâmica do Trabalho e Rendimento (*Survey of Labour and Income Dynamics*), doravante SLID, sendo também apresentada a estratégia usada para estimação e as medidas acessórias utilizadas para avaliar estes modelos. No início da análise de resultados serão relatados, de forma genérica, os sinais dos coeficientes das variáveis, as estimativas para os seus desvios padrão, os *p-values* obtidos para a inferência estatística bem como as medidas de avaliação utilizadas. Posteriormente, será realizada uma análise mais alargada destes resultados e avaliados os efeitos de algumas variáveis sobre a probabilidade de um evento.

No quarto, e último, capítulo será apresentada a conclusão. Nesta secção é apontado um conjunto de recomendações que os investigadores devem ter em atenção quando lidam com eventos raros em dados binários. Adicionalmente, identificam-se algumas questões merecedoras de investigação futura baseadas na análise de resultados.

Capítulo 2

Revisão de Literatura

When dealing with people, remember you are not dealing with creatures of logic, but with creatures bristling with prejudice and motivated by pride and vanity.

Dale Carnegie

Seja n a dimensão de uma dada amostra $\{Y, X\}$, com X e Y definidos em $\mathcal{X} \times \mathcal{Y}$, onde $Y = (y_1, y_2, \ldots, y_n)^T$ é um vetor $n \times 1$ de observações independentes com $y_i \in \{0, 1\}$, onde $i = \{1, \cdots, n\}$, definida como a variável de interesse binária e X é uma matriz $n \times (k+1)$ que contem as linhas de $\boldsymbol{x}_i^T$, na qual $\boldsymbol{x}_i^T$ representa a transposta do vetor $\boldsymbol{x}_i = (1, x_1, x_2, \ldots, x_k)^T$ com dimensões $(k+1) \times 1$ das variáveis explicativas, logo x_{ij} representa a i-ésima observação da variável j. Considere-se também $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ um vetor $(k+1) \times 1$ dos coeficientes que se pretendem estimar e seja a componente sistemática dada por $\eta_i = \boldsymbol{x}_i^T \beta$.

Dado que y_i segue uma distribuição de Bernoulli, $y_i | \boldsymbol{x}_i \sim Ber(\pi)$, pertencente à família de distribuições exponenciais $\mathcal{F}$, onde $\pi = p(\boldsymbol{x}) \equiv P(y_i = 1|X)$. O nosso interesse é estimar os coeficientes de regressão bem como a probabilidade $p(\boldsymbol{x})$ (probabilidade de uma ocorrência) em função dos regressores.

É possível definir $p(\boldsymbol{x}) = G(\eta_i)$, onde $G(\cdot)$ é uma função distribuição, pelo que $G^{-1}(\cdot)$ será a função *link*, onde $G(\cdot) : \mathbb{R} \mapsto [0; 1]$ é estritamente crescente e duplamente diferenciável. As funções *link* fornecem a relação linear entre a componente sistemática e a média da distribuição, transformando a variável dependente.

A estimação destes modelos, independentemente da distribuição assumida para

$G(\cdot)$, é geralmente feita por máxima-verosimilhança, maximizando 2.1 no espaço dos parâmetros, utilizando um algoritmo de otimização[1]. O logaritmo natural da função de máxima-verosimilhança para uma amostra aleatória é dado por:

$$\ell(\beta|\{Y,X\}) = \sum_{i=1}^{n} y_i \, ln\left[G(\eta_i)\right] + (1-y_i) \, ln\left[1-G(\eta_i)\right] \tag{2.1}$$

Em termos de previsão, classifica-se um evento como pertencente à classe dos uns se $\widehat{\pi}_i > \theta$, caso contrário classifica-se como pertencente à classe dos zeros, onde θ é a fronteira de decisão. Geralmente fixa-se $\theta = 1/2$, sendo esta escolha ótima se as probabilidades estimadas definirem a classe que se deve atribuir a cada evento considerando iguais custos de má classificação (Palepu, 1986; Cramer, 1999; Hosmer e Lemeshow, 2000; Kleinbaum e Klein, 2010).

Como estamos perante funções não lineares, os coeficientes das variáveis não medem diretamente o efeito marginal. Para calcular estes efeitos é necessário fazer uso da equação (2.2) para obter o Efeito Marginal na Probabilidade (*Marginal Probability Effect*), doravante MPE:

$$MPE_j = \Delta p(\boldsymbol{x}_i) = G(\boldsymbol{x}_i^T\beta + \Delta x_{ij}\beta_j) - G(\boldsymbol{x}_i^T\beta) \tag{2.2}$$

Esta equação interpreta-se como uma mudança discreta na probabilidade associada com uma mudança discreta no regressor j na quantidade Δx_{ij}. Este efeito varia naturalmente de indivíduo para indivíduo, pois é função dos regressores.

Para variáveis continuas o MPE é dado pela derivada, fazendo uso da regra da cadeia obtendo-se:

$$MPE_j = \frac{\partial p(\boldsymbol{x}_i)}{\partial x_{ij}} = \frac{dG(x_i^T\beta)}{d(x_i^T\beta)}\frac{\partial x_i^T\beta}{\partial x_{ij}} = \nabla_{\eta_i} G(\eta_i)\beta_j = g(\boldsymbol{x}_i^T\beta)\beta_j \tag{2.3}$$

Assim, uma mudança na probabilidade de obter uma ocorrência devido a uma variação no regressor numa quantidade $\Delta\boldsymbol{x}_{ij}$ pode ser aproximada por $\Delta\pi_{ij} \approx \left[g(\boldsymbol{x}_i^T\beta)\beta_j)\right]\Delta x_{ij}$. Note-se que (2.3) será maximiza quando $\eta = 0$ dado que a função densidade atinge o seu máximo nesse ponto. Na figura A.3 encontra-se um exemplo gráfico para os modelos binários *standard*. Nesta figura também se pode ver que o sinal do efeito da variável j é o mesmo de β_j, dado que $g(\eta_i) > 0 \forall \eta_i$.

Partindo da equação (2.3) é possível derivar algumas quantidades de interesse.

[1] Veja-se Mai et al. (2014) para uma apresentação detalhada dos mesmos.

Para esta dissertação será utilizado o efeito marginal esperado, $E_x\left[g(\boldsymbol{x}_i^T\beta)\right]\beta_j$, que pode ser estimado por:

$$\widehat{AMPE}_j = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i^T\beta)\beta_j \tag{2.4}$$

Onde AMPE representa o Efeito Marginal Médio na Probabilidade (*Average Marginal Probability Effect*). Esta equação implica calcular a probabilidade marginal para cada indivíduo e depois calcular a sua média (Leeper, 2017).

A escolha de $G(\cdot)$ não pode ser arbitrária, pois a seleção errada da distribuição causa enviesamento nas estimativas dos parâmetros, tanto assintoticamente como em amostras finitas(Czado e Santner, 1992).

2.1 Modelos *Standard*

Tradicionalmente o *logit* e o *probit*, a apresentar a seguir, são os modelos categóricos mais conhecidos. No entanto, existe um conjunto mais alargado de especificações para $G(\cdot)$ que já são relativamente conhecidas que se passa a apresentar.

Se for assumido que $G(\cdot)$ é a distribuição acumulada da $N(0,1)$, isto é,

$$G(\eta_i) := \Phi(\eta_i) = \frac{1}{\sqrt{2\,\pi}}\int_{-\infty}^{\eta_i} \mathrm{e}^{-\frac{1}{2}t^2}dt \tag{2.5}$$

obtemos o modelo *probit*. Se assumirmos que $G(\cdot)$ é a distribuição Logística$(0,1)$, ou seja,

$$G(\eta_i) := \Lambda(\eta_i) = \frac{1}{1+e^{-\eta_i}} \tag{2.6}$$

obtemos o *logit*. No contexto dos modelos lineres generalizados (Nelder e Wedderburn, 1972), para a distribuição Bernoulli, o logit é o link natural\canónico. Alternativamente, se $G(\cdot)$ for a distribuição acumulada da Cauchy$(0,1)$,

$$G(\eta_i) = \frac{1}{\pi}\arctan\left(\eta_i\right) + \frac{1}{2}\,, \tag{2.7}$$

obtemos o modelo *cauchit*. Estas distribuições são simétricas em torno da sua média.

Para além destes três modelos simétricos deve-se também apresentar as duas distribuições Gumbel que estão na base dos dois únicos modelos assimétricos desta secção. A distribuição Gumbel acumulada do valor máximo *standard*, ou simplesmente

distribuição Gumbel *standard* acumulada, é dada por

$$G(\eta_i) = exp\left[-exp(-\eta_i)\right] \tag{2.8}$$

e a sua utilização produz o modelo *loglog*. A distribuição acumulada Gumbel do valor mínimo *standard*, ou distribuição Gumbel *standard* complementar acumulada, é dada por

$$G(\eta_i) = 1 - exp\left[-exp(\eta_i)\right] \tag{2.9}$$

e a sua utilização produz o modelo *cloglog*. Nas figuras A.1 e A.2 encontram-se o gráfico da distribuição acumulada (f.d.a.) da probabilidade e da função densidade da probabilidade (f.d.p.), respetivamente, para estes cinco modelos. Nestes gráficos pode ver-se que o *cauchit* tem as caudas mais pesadas, seguido pelo *logit* e *probit*, ou seja, o probit aproxima-se dos eixos mais rapidamente que os outros dois modelos. O *cloglog* e o *loglog* comportam-se de forma simétrica em relação à média, $g_{loglog}(\eta_i) = g_{cloglog}(-\eta)$.

A diferença entre os modelos simétricos e assimétricos encontra-se nas caudas. Nos modelos assimétricos as funções distribuição acumuladas aproximam-se dos extremos a taxas diferentes, o que não acontece nos modelos simétricos (Hardin et al., 2007).

Os estimadores dos coeficientes dos modelos *standard*, apesar de consistentes e eficientes, estarão enviesados na presença de eventos raros. A $P(y_i = 1|X)$ (ocorrências) é sub-estimada, logo $P(y_i = 0|X)$ é sobre-estimada). No entanto, isto não implica que $p(\boldsymbol{x})$ deixe de assumir valores em todo o seu domínio (Cramer, 1999; King e Zeng, 2001).

Para o *logit*, Owen (2007) provou que quando se mantém constante uma classe de y e aumenta de forma irrestrita o número de observações da outra classe, $\beta_0 \to -\infty$ enquanto os coeficientes dos restantes regressores tendem para um limite inferior. No entanto, nada se pode concluir para os restantes modelos.

As estimativas obtidas por máxima verosimilhança, para os modelos *standard*, apresentam uma fórmula fechada para a variância (assintótica), assumindo a boa especificação da densidade de $y_i|\boldsymbol{x}_i$, dada por:

$$Avar(\beta) = \left[\sum_{i=1}^{N} -E[H_i(\beta)]\right]^{-1} \tag{2.10}$$

onde o valor esperado da Hessiana para uma qualquer função $G(\eta_i)$ é dado por:

$$E[H_i(\beta)] = -\frac{\left[\nabla_{\eta_i} G(\eta_i)\right]^2 \boldsymbol{x}_i \boldsymbol{x}_i^T}{G(\eta_i)\left[1 - G(\eta_i)\right]} \tag{2.11}$$

Como a primeira derivada dos termos da equação (2.11) afetados pelos eventos extremos é uma função positiva, ou seja,

$$\frac{\partial}{\partial G(\eta_i)}\left[-\frac{\left[\nabla_{\eta_i} G(\eta_i)\right]^2}{G(\eta_i)\left[1 - G(\eta_i)\right]}\right] = -\frac{\left[g(\eta_i)\right]^2\left[1 - 2G(\eta_i)\right]}{\left[G(\eta_i)\right]^2\left[1 - G(\eta_i)\right]^2} < 0, \forall G(\eta_i) \in (0; 0,5) \tag{2.12}$$

a variância decresce com o aumento da probabilidade de um evento condicional nos regressores[2], para uma dimensão amostra fixa.

Por outro lado a variância será mínima quando a parte da sua fórmula que é afetada pela escolha de $G(\cdot)$ na Hessiana for máxima, o que acontece quando $G(\eta_i) = 0,5$. Ao adicionar-se mais observações da classe com menos observações na amostra, a variância assintótica dos modelos de regressão binária tende a reduzir-se, pelo que essa classe é mais informativa que a sua contraparte. Este *"trade-off"* continuará até que se atinja uma situação de estabilidade, $P(y_i = 1|X) = P(y_i = 0|X) = \frac{1}{2}$, onde a variância é mínima, considerando a dimensão da amostra fixa.

A situação onde se alteram as proporções de uns e zeros na amostra relativamente à população, onde o número de eventos e dos controlos são iguais é um caso especial[3], será o primeiro caso a abordar na próxima subsecção.

[2] Este resultado advém da generalização, para qualquer função $G(\cdot)$, dos resultados apresentados em Westphal (2013) para o modelo *logit*.

[3] A situação onde existe igual número de controlos e ocorrências é designado de amostragem de proporções iguais (*equal shares sampling*).

2.2 Modelos que Procuram Incorporar Eventos Raros

2.2.1 Modelos Baseados em Seleção Amostral

Defina-se a função densidade conjunta da população, $f(\mathbf{x}, y) = f(y|\boldsymbol{x}, \beta)f(\boldsymbol{x})$, onde $f(y|\boldsymbol{x}, \beta)$, para variáveis binárias será a função densidade Bernoulli de parâmetro π, depende de um parâmetro desconhecido β. Defina-se $\mathcal{C}_s$, com $s = 1, 2, \cdots, T$, como um conjunto finito mutuamente exclusivo designado por estrato de onde as observações que irão fazer parte da amostra a ser estudada são retiradas. Segundo Imbens e Lancaster (1996) é possível definir três casos de interesse para eventos raros binários, considerando que existe amostragem aleatória dentro de cada estrato:

- Se $s = 1$ e $\mathcal{C}_1 = \mathcal{X} \times \mathcal{Y}$ estamos na presença de uma amostra aleatória, logo não existe enviesamento de seleção.
- Se $\mathcal{C}_s = \mathcal{X}_s \times \mathcal{Y}$, com $\mathcal{X}_s \subset \mathcal{X}$, a estratificação da amostra é baseada apenas nas variáveis explicativas, temos estratificação exógena pura.
- Se $\mathcal{C}_s = \mathcal{X} \times \mathcal{Y}_s$, com $\mathcal{Y}_s \subset \mathcal{Y}$, a estratificação na amostra é baseada apenas na variável dependente, temos estratificação endógena pura .

A diferença entre uma amostra aleatória e uma amostra não aleatória, como é o caso da amostragem estratificada, encontra-se no facto da probabilidade de uma unidade ser selecionada ao acaso numa amostra não aleatória ser diferente da probabilidade de selecionar essa mesma unidade na população (Imbens e Lancaster, 1996; Cameron e Trivedi, 2005).

Na estratificação exógena pura, desde que $\boldsymbol{x}$ seja exógeno em relação a y, a função densidade conjunta amostral é dada por $f^s(y, \boldsymbol{x}|\beta) = f(y|\boldsymbol{x}, \beta)f^s(\boldsymbol{x})$, onde a densidade $f^s(\boldsymbol{x})$ não depende de β. Sob estas condições o estimador não é afetado pela estratificação da amostra, pelo que os estimadores *standard* dos parâmetros serão consistentes e eficientes. Isto deve-se ao *kernel* do logaritmo da verosimilhança depender apenas de $f(y|\boldsymbol{x}, \beta)$ e não precisar que a distribuição $f^s(\boldsymbol{x})$ seja parametrizada ou estimada (Manski e McFadden, 1981). Note-se que se for possível arranjar uma, ou mais variáveis explicativas com grande capacidade preditiva da probabilidade de observar um evento raro, é possível estratificar a amostra obtendo mais observações desse segmento

da população sem invalidar a inferência. Esta abordagem pode não ser possível, pois a existência de variáveis explicativas com grande capacidade explicativa pode não existir. Alem disso, em principio, não é possível saber quais são esses regressores sem primeiro obter a amostra e analisar os dados (Winkelmann e Boes, 2006).

Na estratificação endógena a distribuição de y na amostra difere da distribuição de y na população. A função densidade conjunta amostral é dada por $f^s(y, \boldsymbol{x}|\beta) = f(y|\boldsymbol{x}, \beta)\frac{f^s(y)}{f(y|\boldsymbol{x})}f(\boldsymbol{x})$. Como tal a estimação de máxima verosimilhança baseada apenas em $f(y|\boldsymbol{x}, \beta)$ será inconsistente porque negligencia $f(y|\boldsymbol{x})$, o que pode alterar totalmente, quantitativamente e qualitativamente, as conclusões nos estudos empíricos (Cameron e Trivedi, 2005; Winkelmann e Boes, 2006; Abadie et al., 2018).

Para variáveis binárias, ou multinomiais, a estratificação baseada em y é apelidada de Amostragem Baseada na Escolha (*Choice-Based Sampling*), doravante CBS. Quando cada classe da variável dependente é um diferente estrato estamos perante CBS puro (Imbens, 1992).

A utilização de uma amostra com seleção na variável dependente é uma ideia atrativa. Os custos para obter uma amostra com seleção endógena são geralmente baixos. A informação trazida por uma observação adicional de uma ocorrência é maior que a informação adicional de um novo controlo, sendo o caso em que o número de controlos é igual ao número de ocorrências considerado ótimo, ou perto de ótimo, para a maioria das situações. Além disso, na análise de eventos raros, uma amostra aleatória não terá um número de ocorrências suficientes para uma análise estatística efetiva, pois a informação presente nessa amostra é bastante pequena, levando a estimativas dos parâmetros relativamente imprecisas. (Cosslett, 1981a; Xie e Manski, 1989; Imbens, 1992; King e Zeng, 2001).

Seja $Q(s|\beta)$ a probabilidade de selecionar uma observação da população do estrato $\mathcal{C}_s$ dada por:

$$Q_s := Q(s|\beta) = P((Y, X) \in \mathcal{C}_s) = \int_{\mathcal{C}_s} f(y|\boldsymbol{x}, \beta)f(\boldsymbol{x})dyd\boldsymbol{x} \tag{2.13}$$

A distribuição conjunta amostral de (S, Y, X), é dada por (2.14), onde S é o indicador do estrato de onde a observação foi retirada e H_s é a probabilidade de na amostra selecionar uma observação de $\mathcal{C}_s$, e está na base dos estimadores que iremos apresentar a seguir:

$$f(s, y, \boldsymbol{x}|\beta) = \frac{H_s}{Q_s} f(y|\boldsymbol{x}, \beta) f(\boldsymbol{x}) = f^s(y|\boldsymbol{x}, \beta) f^s(\boldsymbol{x}) \tag{2.14}$$

A maximização direta da função de verosimilhança com base em (2.14) envolve a distribuição de Q_s, dada por (2.13), que, por sua vez, requer que se especifique $f(x)$. Manski e Lerman (1977) apresentam uma solução simples e fácil implementar que permite a obtenção das estimativas dos parâmetros com base na equação (2.14).

Suponha-se que existe informacao adicional sobre H_s e Q_s, onde Q_s é a fração de elementos selecionados na população e H_s será o seu equivalente no estrato s da amostra. A proposta de Manski e Lerman (1977) para uma variável binária, com $\omega_1 = Q_1/H_1 = \tau/\bar{y}$ para $y_i = 1$ e $\omega_0 = (1 - Q_1)/(1 - H_1) = (1 - \tau)/(1 - \bar{y})$ para $y_i = 0$ consiste em maximizar

$$\ell(\beta|\{Y, X\}, \omega_i) = \sum_{i=1}^{n} \omega_1 y_i \, ln\left[G(\eta_i)\right] + \omega_0 (1 - y_i) \, ln\left[1 - G(\eta_i)\right] , \tag{2.15}$$

onde $\bar{y}$ é a proporção de uns na amostra e τ é a proporção de uns na população.

A função objetivo (2.15) corresponde a uma simples ponderação do logaritmo da verosimilhança, sendo conhecido como Estimador de Máxima Verosimilhança Ponderada de Amostragem Exógena (*Weighted Exogenous Sample Maximum Likelihood Estimator*), doravante WESML. Esta ponderação faz o estimador comportar-se como se estivéssemos numa amostra aleatória. No entanto, (2.15) não é formalmente o logaritmo de uma verosimilhança, dado que (2.14) não implica que $f^s(y|\boldsymbol{x}, \beta) = f(y|\boldsymbol{x}, \beta)^{Q_s/H_s}$, pelo que é necessário utilizar as variâncias robustas de Huber. O WESML será consistente mas não eficiente (Huber, 1967; Cameron e Trivedi, 2005; Freedman, 2006).

Para o modelo *logit*, dado que este modelo pertence à classe dos modelos com constante multiplicativa (Hsieh et al., 1985), prova-se que as estimativas para β em CBS com ou sem aplicação da correção para seleção amostral são equivalentes, à exceção da constante. A constante, quando não é aplicada a correção para a seleção amostral, pode ser retificada posteriormente. Isto implica que utilizar o WESML ou corrigir a constante *a posteriori* leva a estimativas consistentes e assintoticamente normais, sendo o último caso também assintoticamente eficiente (Xie e Manski, 1989).

Alternativamente, é possível utilizar o Método do Momentos Generalizado (*Generalized Methods of Moments*), doravante GMM. Esta solução foi apresentada por Imbens (1992) para variáveis categóricas, sendo posteriormente generalizada em Imbens

e Lancaster (1996) para variáveis dependentes contínuas. Este estimador considera que não se conhecem as probabilidades Q_S e H_S, sendo considerados parâmetros que se devem estimar conjuntamente com os coeficientes das variáveis de interesse. As estimativas obtidas via GMM serão consistentes e eficientes. No entanto, este estimador é mais difícil de implementar e produz estimativas pontuais parcialmente iguais às obtidas pelo WESML, divergindo apenas nas estimativas das variâncias, pelo que o GMM não será abordado nesta dissertação[4].

Para variáveis binárias com eventos raros esta metodologia pode ser aplicada de varias formas. Primeiramente é necessário obter informação acerca de τ, esta informação pode ser obtida, por exemplo, de estudos realizados previamente, estatísticas/relatórios oficiais publicados por bancos/seguradoras ou pelo governo. Se não for possível obter informação sobre a percentagem de uns na população utilizando estas fontes pode-se retirar uma amostra aleatória da população que se pretende estudar obtendo-se assim uma estimativa para τ.

Obtida uma estimativa para o valor de τ é necessário inflacionar o número de uns que será utilizado para a estimação dos parâmetros, pode-se então retirar uma amostra da população onde são recolhido mais eventos na população que a sua real percentagem. Caso o investigador tenha retirado/acesso a uma amostra aleatória pode inflacionar o número de uns na amostra retirando de forma aleatória eventos da população e juntando-os à amostra utilizada para a estimação. Se não for possível retirar mais eventos da população a inflação do numero de uns pode ser realizada retirando controlos da amostra aleatória. Note-se que em todos os casos τ vai diferir de $\bar{y}$, ter-se-á $\tau < \bar{y}$, no entanto, no segundo caso existe um ganho de informação pois aumenta-se a amostra utilizada para a estimação, enquanto que no último caso existira perda de informação dado que se retiram zeros. Para este estudo de simulação adotar-se-á este último procedimento, dado que se pretende ver o resultado do balanceamento da amostra nas estimativas dos parâmetros, sendo o caso em que se retiram controlos mais relevante do ponto vista empírico. Uma explicação mais alargada da adoção deste procedimento bem como o seu enquadramento no estudo de simulação pode ser visto na sub-secção 3.1.

[4] Existem outras possibilidades para além dos dois estimadores apresentados, veja-se por exemplo Prentice e Pyke (1979), Manski e McFadden (1981), Cosslett (1981a) e Cosslett (1981b).

2.2.2 Modelos Baseados em Formas Flexíveis

Como já foi mencionado, a utilização de distribuições simétricas pode ser bastante restritiva e obviamente em algumas aplicações não ser adequada, principalmente quando as duas categorias da variável binárias não estão equilibradas. Esta má especificação leva a um enviesamento na estimativa dos parâmetros, bem como nas probabilidades previstas, mesmo assintoticamente tal como já mencionado(Czado e Santner, 1992), tendo especial relevo quando a má especificação envolve a assimetria da distribuição, mesmo que a componente sistemática esteja bem especificada (Koenker e Yoon, 2009)

Apresenta-se a seguir um conjunto de funções *link* que permitem lidar diretamente com a assimetria no número de ocorrências comparativamente com o número de controlos. Para introduzir mais flexibilidade adiciona-se um ou mais parâmetros na distribuição, ou utiliza-se uma distribuição assimétrica. Em ambos os casos é possível obter distribuições simétricas como casos especiais dependendo da forma como esse(s) parâmetro(s) extra são introduzidos, ou da distribuição assimétrica usada. A utilização destes modelos permite, à partida, um melhor ajustamento aos dados, dada a alta versatilidade das suas especificações.

Quando se fixa esse(s) parâmetro(s) adicionais, atribuindo-lhe(s) um valor, a estimação é extremamente simples, mas sem grande interesse. Quando se estima(m) conjuntamente com os coeficientes dos regressores, as funções log-verosimilhança não são separáveis sendo necessário obter simultaneamente as estimativas para todos os parâmetros. Seguindo Taylor et al. (1996) o caso em que estima ζ, ou outro parâmetro adicional, será chamado de incondicional e o caso em se fixam os parâmetros adicionais de condicional. Tal como na secção anterior, as estimativas dos parâmetros serão obtidas por máxima-verosimilhança.

Aranda-Ordaz (1981), sugere a utilização da seguinte função *link* assimétrica:

$$G^{-1}(\pi;\zeta) = ln\left[\frac{(1-p(\boldsymbol{x}))^{-\zeta}-1}{\zeta}\right], \zeta > 0 \tag{2.16}$$

Esta função *link* tem como caso especial o *cloglog* quando $\lim_{\zeta\to 0} G^{-1}(\pi;\zeta)$ e o *logit*, quando $\zeta = 1$. No entanto, a assimetria e a *kurtosis* não variam com os valores dos regressores, apenas a média e a variância. Na figura A.4 é apresentada a função distribuição acumulada Aranda-Ordaz bem como a sua função densidade da probabilidade, respetivamente, com a mesma componente sistemática para dois valores diferentes de ζ.

Caron e Polpo (2009) sugerem a utilização da distribuição Weibull, dada a sua simplicidade e flexibilidade, pois acomoda caudas simétricas e assimétricas (Caron et al., 2018). A função distribuição acumulada é dada por

$$G(x;\zeta) = 1 - exp\left[-\left\{\frac{(x-\mu)}{\alpha}\right\}^{\zeta}\right]\mathbf{I}_{(x>\mu)} \tag{2.17}$$

onde $\alpha > 0$, $\mu \in \mathbb{R}$ é parâmetro de localização, $\zeta > 0$ controla a forma das caudas da distribuição e $\mathbf{I}_{(x>\mu)}$ é a função indicadora para $x > \mu$.

Pode-se definir a forma funcional para este modelo como

$$G(\eta_i) = 1 - exp\left[-\eta_i^{\zeta}\right] \tag{2.18}$$

onde $E(y_i|x_i) = G(\eta_i;\zeta)$, $\zeta > 0$ e $\eta_i > 0$. Nesta parametrização, a restrição $\eta > 0$ não é problemática pois β_0 vai desempenhar o papel de parâmetro de localização e constante na componente sistemática, pelo que a sua interpretação difere dos restantes modelos (Caron e Polpo, 2009). Desta forma, evita-se o problema de identificabilidade e obtém-se um modelo mais parcimonioso (Caron et al., 2018). Estes últimos autores definem também a função Weibull refletida como

$$G(\eta_i) = exp\left[-\eta_i^{\zeta}\right] \tag{2.19}$$

Quando $\zeta \to \infty$, obtemos a função de distribuição acumulada de Gumbel de valor máximo ou mínimo, dependendo se estamos a utilizar a distribuição Weibull ou Weibull refletida, respetivamente. Na figura A.5 é apresentada a função Weibull e Weibull refletida para $\zeta = 0,5$ e $\zeta = 1,5$ com a mesma componente sistemática.

Usando os resultados em Rinne (2008) obtém-se

$$\begin{aligned} G(\eta_i) &= 1 - exp[-(0.90144 + 0.27787\eta_i)^{3.60235}] \approx \Phi(\eta_i) \\ G(\eta_i) &= 1 - exp[-(0.89864 + 0.16957\eta_i)^{3.60235}] \approx \Lambda(\eta_i) \end{aligned}, \tag{2.20}$$

pelo que a distribuição Weibull pode também aproximar os modelos *probit* e *logit*. Logo estas duas funções de distribuição têm como casos especiais os modelos *standard*, podendo ajustar modelos onde a função *link* é assimétrica ou simétrica.

A figura A.6 apresenta a aproximação do modelo Weibull ao *logit* e *probit*, respetivamente, verificando-se um bom ajuste deste modelo às outras duas funções *link*

mencionadas. Apesar disso, o modelo Weibull aproxima-se melhor do *probit*.

Considera-se de seguida os modelos baseados na função potência. Seguindo Bazán et al. (2017) define-se genericamente que Θ segue uma distribuição potência, $\Theta \sim P(\mu, \sigma, \zeta)$, onde a sua função densidade e função distribuição são dadas respetivamente por,

$$f_P(\theta|\mu,\sigma,\zeta) = \frac{\zeta}{\sigma}\, g\Big(\frac{\theta-\mu}{\sigma}\Big)\left[G\Big(\frac{\theta-\mu}{\sigma}\Big)\right]^{\zeta-1} \tag{2.21}$$

$$F_P(z) = \left[G\Big(\frac{\theta-\mu}{\sigma}\Big)\right]^{\zeta} \tag{2.22}$$

Temos, genericamente, que Θ segue uma distribuição potência reversa, $\Theta \sim PR(\mu, \sigma, \zeta)$, onde a sua função densidade e função distribuição são dadas respetivamente por,

$$f_{PR}(\theta|\mu,\sigma,\zeta) = \frac{\zeta}{\sigma}\, g\left[-\Big(\frac{\theta-\mu}{\sigma}\Big)\right]\left[G\Big(-\Big(\frac{\theta-\mu}{\sigma}\Big)\Big)\right]^{\zeta-1} \tag{2.23}$$

$$F_{PR}(\theta) = 1 - \left[G\Big(-\Big(\frac{\theta-\mu}{\sigma}\Big)\Big)\right]^{\zeta} \tag{2.24}$$

De forma equivalente, Θ segue uma distribuição potência complementar, $\Theta \sim PC(\mu, \sigma, \zeta)$, onde a sua função densidade e função distribuição são dadas respetivamente por,

$$f_{PC}(\theta|\mu,\sigma,\zeta) = \frac{\zeta}{\sigma}\, g\Big(\frac{\theta-\mu}{\sigma}\Big)\left[1-G\Big(\frac{\theta-\mu}{\sigma}\Big)\right]^{\zeta-1} \tag{2.25}$$

$$F_{PC}(\theta) = 1 - \left[1-G\Big(\frac{\theta-\mu}{\sigma}\Big)\right]^{\zeta} \tag{2.26}$$

Onde $G(\cdot)$ representa uma qualquer função distribuição acumulada contínua e $g(\cdot)$ representa a sua derivada, com $\zeta \in \mathbb{R}^+$ para as as funções distribuição pertencerem ao intervalo $[0; 1]$. O parâmetro ζ controla a assimetria da distribuição e esta relacionado com a proporção de ocorrências na amostra. É esperado que, para $\zeta > 1$, existam mais eventos do que controlos, já quando $\zeta < 1$ o oposto deverá acontecer. Este parâmetro na função potência complementar é esperado que se comporte de forma contrária ao modelo potência (Abanto-Valle et al., 2014; Lemonte e Bazán, 2018)[5]. Na figura A.7 e

[5] Originalmente as funções potência e potência complementar foram propostas por Ramalho et al.

A.8 encontra-se uma exposição gráfica da distribuição potência e potência complementar com base da distribuição logística. Note-se que a distribuição potência complementar e potência reversa são equivalentes quando $G(\cdot)$ é uma distribuição simétrica. Para demonstrar esta regularidade basta verificar que $G(-a) = 1 - G(a), \forall a \in \mathbb{R}$.

Dada a especificação para a distribuição de base é possível obter os modelos previamente apresentados nas outras secções como casos especiais, bem como outros modelos já conhecidos na literatura, como por exemplo o *Scobit* e o *Generalized Skew Normal*, propostos por Chen et al. (1999) e Gupta e Gupta (2004) respetivamente. Estes modelos têm uma grande versatilidade, pois apresentam a possibilidade de utilizar como base numa distribuição simétrica enquanto se mantém uma grande flexibilidade na forma da assimetria.

A utilização da regressão binária com funções potência e potência reversa não deixa de ter críticas, Jiang et al. (2013) apontam que que a utilização das funções potência ou reversa potência apenas conseguem captar a assimetria numa direção, sendo a assimetria na outra direção limitada.

2.3 Notas Finais

A estimação dos parâmetros $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ e ζ, baseada na função de verosimilhança para os modelos flexíveis numa amostra aleatória é dada por:

$$\ell(\beta, \zeta | \{Y, X\}) = \sum_{i=1}^{n} \Big(y_i \, ln\big[G(\eta_i, \zeta)\big] + (1 - y_i) \, ln\big[1 - G(\eta_i, \zeta)\big] \Big) \tag{2.27}$$

A diferença entre (2.1) e (2.27) está na distribuição assumida para a probabilidade $P(y_i = 1|X)$, na ultima equação esta probabilidade depende, não só das variáveis mas também de um parâmetro extra ζ. No caso de ser pretendida a estimação não condicional em ζ, as funções *score* obtidas das primeiras derivadas de (2.27) em ordem ζ e β não são ortogonais, dado que a matriz de Informação de Fisher não é diagonal, pelo que a estimação tem de ser realizada conjuntamente, tal como mencionado anteriormente. No entanto, tal como no caso dos modelos standard os estimadores serão consistes e eficientes, assumindo como satisfeitas as usuais condições de regularidade, podendo as

(2011). Posteriormente Abanto-Valle et al. (2014) apresentam a função potência reversa, no entanto estes autores também apresentaram a função potência como sendo de sua autoria, o que não é o caso.

variâncias assintóticas continuar a ser estimadas usando um dos estimadores habituais para a variância[6].

Na literatura existem outros modelos propostos para eventos raros. Por exemplo, Calabrese e Osmetti (2015) apontam que a utilização de um modelo linear nos regressores é um pressuposto que não se verifica tipicamente nas aplicações empíricas, estes autores sugerem a utilização conjunta da distribuição GEV[7] com os Modelos Aditivos Generalizados (GAM)[8], ou Maalouf et al. (2011) que sugerem a utilização de *kernels* para melhor separação entre os zeros e os uns. Alternativamente, é possível criar mais modelos adicionando mais parâmetros à função *link* como forma de controlar outros aspetos, tal como a assimetria e forma das caudas da distribuição, como é o caso do *pregibit* (Pregibon, 1980) e, mais recentemente, da função *link* proposta por Taneichi et al. (2014), que se pode facilmente demonstrar como sendo uma função potência com base na distribuição Aranda-Ordaz assimétrica.

Existe também literatura que sugere a criação de modelos mistos usando duas ou mais das funções *link* apresentadas, veja-se por exemplo Jiang et al. (2013). Estes modelos permitem, teoricamente, um melhor ajuste aos dados, mas apresentam um baixo grau de parcimoniosidade e grande dificuldade de implementação computacional.

[6] Veja-se Czado e Santner (1992) para uma apresentação mais extensa.

[7] Tanto a distribuição Gumbel como a distribuição Weibull pertencem à família da distribuição GEV.

[8] Um modelo GAM pode definir-se como $G^{-1}(p(\boldsymbol{x})) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_{k+1}(x_{k+1})$, onde f_i representa uma qualquer função paramétrica, semi-paramétrica ou não paramétrica.

Capítulo 3

Estudo de Simulação

Your faith in yourself should overcome your insecurities. Imagine yourself a king and you will become a king.

Robert Greene

3.1 Dados e Metodologia

O estudo de simulação, que ilustra o comportamento dos modelos considerados anteriormente, utiliza os dados disponibilizados em Wen e Gordon (2014), que foram retirados do SLID, para o ano de 1999 a 2005. Pretende-se estudar os efeitos de um conjunto de variáveis explicativas na probabilidade de um indivíduo se encontrar empregado por conta própria. Um indivíduo é considerado como empregado por conta própria se a sua principal fonte de rendimento for proveniente de trabalho por conta própria e não obtiver rendimento oriundo da agricultura. A variável dependente assume valor um se o indivíduo i é empregado por conta própria e zero caso contrário. A tabela B.1 apresenta a descrição das variáveis disponíveis que foram usadas por Wen e Gordon (2014). Depois de realizado um tratamento inicial da base de dados, que coincide com Wen e Gordon (2014), obtêm-se 142329 observações das quais 8677 são ocorrências, cerca de 6% do total da amostra. No entanto, as observações para o ano 2005 serão usadas como amostra de teste, pelo que serão apenas utilizadas 121930 observações, das quais 7358 são ocorrências, mantendo-se aproximadamente a mesma proporção de uns.

Nas tabelas B.2 e B.3 encontram-se as estatísticas resumidas para os seis anos a usar para o treino dos modelos bem como as estatísticas para o total da amostra, respetivamente. Nestes quadros é possível observar que os trabalhadores por conta própria têm, em média, maiores rendimentos (rendimentos de investimentos e ganhos de capital) e maior variação nos rendimentos comparativamente com os trabalhares empregados por conta de outrem. Isto reflete a noção de que trabalhar por conta própria é mais arriscado. No entanto, estes indivíduos passam menos tempo desempregados, têm uma escolaridade menor e mais filhos que a sua contraparte. Estas conclusões correspondem às conclusões apresentadas por Wen e Gordon (2014) para a totalidade da amostra.

Dado que se pretende comparar modelos oriundos de duas formas diferentes lidar com seleção amostral (modelos apresentados na subsecção 2.2.1) e outra sem seleção (modelos apresentados na secção 2.1 e na subsecção 2.2.2) será necessário controlar não só o número de observações, como a percentagem de eventos na amostra usada para estimar os modelos e a forma como a estratificação endógena é realizada.

Adotou-se então o seguinte esquema metodológico para estudar os modelos. Serão utilizados três valores diferentes para a dimensão amostral ($n = 1000$, $n = 5000$, $n = 10000$) com amostragem aleatória. Cada experiência envolverá 1000 replicas com reposição, sendo reportado o valor médio das estimativas pontuais para os coeficientes dos regressores, bem como para respetivo desvio padrão. Também serão estimados os modelos flexíveis e *standard* usando todas as observações para os anos de 1999 a 2004 como forma de verificar a evolução das estimativas dos parâmetros, das medidas de avaliação da capacidade preditiva e do teste de especificação realizado. As medidas de avaliação da capacidade preditiva bem como o teste de especificação utilizado serão apresentados posteriormente nesta sub-secção.

Para os modelos com seleção amostral serão usadas as mesmas dimensões amostrais e cinco valores para a percentagem de uns na amostra utilizada para a estimação (6%, 15%, 25%, 35%, 50%), sendo a seleção amostral realizada sobre a amostra aleatória retirada dos dados do SLID de 1999 a 2004, isto é, utilizando apenas a amostra de treino. Ou seja, suponha-se que na população em estudo temos uma percentagem de eventos de $6,2\%$, o investigador retira uma amostra aleatória com mil observações ($n = 1000$), desta amostra obtém-se uma estimativa para τ, por exemplo $\hat{\tau} = 0,06$ (note-se que dada a existência de variabilidade amostral a estimativa para τ pode não representar a real percentagem de eventos na população, mas dado o numero de repetições é esperado que

$\bar{\tau} = 0,062$), este investigador decide que pretende ter 15% de uns na amostra utilizada para a estimação ($\bar{y} = 0,15$), logo essa amostra utilizada para a estimação terá de ter 60 eventos e 400 controlos. Cada experiência envolverá também 1000 replicas com reposição sendo reportado o valor médio dos coeficientes e valor médio do médio do respetivo desvio-padrão.

Para uma diferenciação mais fácil entre as amostras, os dados do SLID para o anos de 1999 e 2004 serão denominados simplesmente por população de treino, as amostras aleatórias retiradas da população de treino serão denominadas de amostras de treino aleatórias e o conjunto de observações retiradas da amostra de treino aleatória com seleção amostral serão denominadas por amostras de treino com seleção amostral.

Na figura 3.1 é apresentado de forma esquemática o método de amostragem realizado, quais os modelos utilizados para cada amostra de treino e o método utilizado para realizar seleção amostral.

Esta decisão metodológica advém do que se espera durante a realização de um estudo empírico. O investigador reúne os dados e depois de verificar que os eventos em estudo apresentam uma frequência baixa tem de optar entre realizar seleção amostral baseada na variável dependente, utilizar formas flexíveis para a função *link* ou desconsiderar a raridade dos eventos. Alternativamente, tal como previamente mencionado, pode dar-se o caso em que o investigador sabe *a priori* qual a percentagem de ocorrências na população. Neste caso, durante a recolha da amostra o investigador pode dar mais ênfase na recolha de ocorrências, realizando estratificação endógena.

Para avaliar a capacidade preditiva dos modelos apresentados serão utilizadas duas medidas de comparação, o Brier *Score* (BS) e a Área sob a Curva Característica (*Area Under The Curve*), doravante AUC, obtida da curva do Operador do Recetor Característico *(Receiver Operating Characteristic*), doravante ROC, calculadas para as estimativas pontuais médias.

O Brier *Score*, equação 3.1, foi desenvolvido por Brier (1950) e é utilizado extensivamente na avaliação da probabilidade de uma previsão (Elliott e Timmermann, 2013).

$$BS = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\pi}_i - Y_i)^2 \tag{3.1}$$

O Brier *Score* assume valores entre zero e dois sendo que quanto mais baixo este *score* melhor a qualidade do modelo, quando $BS = 0$ temos o caso de previsão perfeita.

Para apresentação da AUC, defina-se primeiro a curva ROC. Seja a função

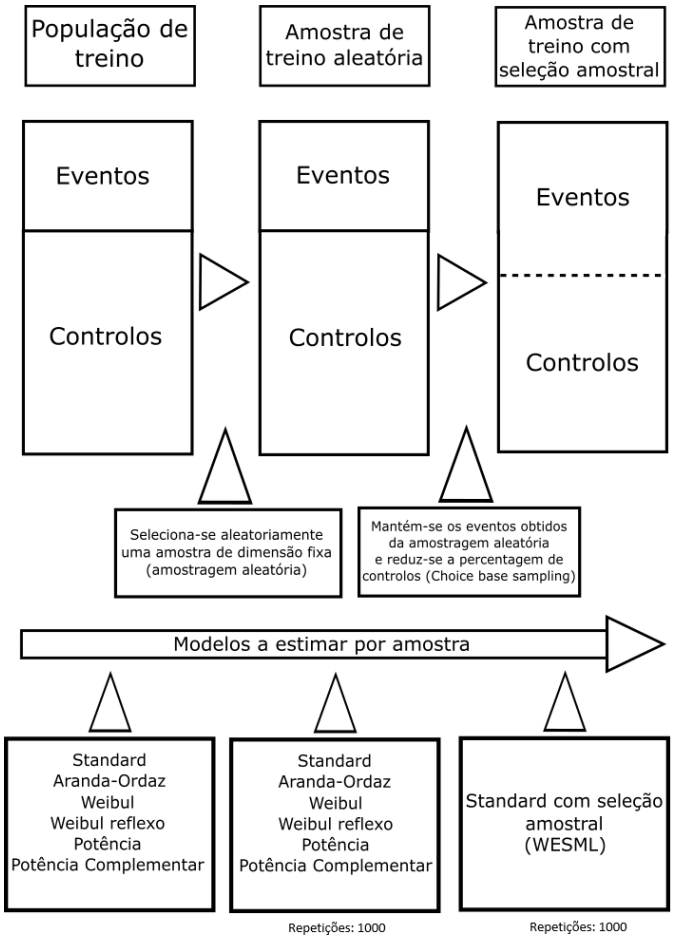

Figura 3.1

densidade para os elementos corretamente classificados como uns dada por $f_1(\hat{\pi})$ e a função densidade para os elementos erroneamente classificados como eventos dada por $f_0(\hat{\pi})$. A Taxa de Verdadeiros Positivos (*True Positive Rate*), doravante TPR, e a Taxa de Falsos Positivos (*False Positive Rate*), doravante FPR, podem ser escritas em função da fronteira de decisão como:

$$\text{TPR}(\theta) = \int_{\theta}^{\infty} f_1(\hat{\pi})\, d\hat{\pi} \tag{3.2}$$

$$\text{FPR}(\theta) = \int_{\theta}^{\infty} f_0(\hat{\pi})\, d\hat{\pi} \tag{3.3}$$

A curva ROC consiste na representação gráfica, de forma paramétrica, de $\text{TPR}(\theta)$ *versus* $\text{FPR}(\theta)$ com θ a variar entre zero e um. Esta curva tem especial interesse pois

permite visualizar graficamente a capacidade preditiva do modelo para várias fronteiras de decisão e comparar com as curvas resultantes de outros modelos.

Para sintetizar toda a informação disponível na curva ROC geralmente usa-se a AUC, que corresponde a área sob a curva ROC. Esta área pode calcular-se como:

$$AUC = \int_0^1 \text{TPR}(\text{FPR}^{-1}(\theta))\, d\theta \tag{3.4}$$

A AUC é uma medida da capacidade do modelo separar os eventos de interesse dos não eventos. Um modelo sem capacidade discriminatória tem um valor perto de 1/2, enquanto um valor perto de 1 sugere excelente capacidade preditiva (Bradley, 1997; Hanley e McNeil, 1982; Fawcett, 2006; Ogundimu, 2019).

Para garantir a boa especificação dos modelos estimados será utilizado o Teste de Erro de Especificação em Regressão (*Regression Specification Error Test*) (Ramsey, 1969), doravante RESET. Para realizar este teste serão usadas as segundas e terceiras potências do índice linear estimado, dado que este teste apresenta uma potência superior às outras versões deste teste para os modelos *standard* (Ramalho e Ramalho, 2012), utilizando o rácio de verosimilhanças. Serão também reportados os *p-values* médias para este teste.

A estimação dos parâmetros para os modelos *standard* com e sem seleção será realizada utilizando os Mínimos Quadrados Iterativamente Ponderados (*Iteratively Reweighted Least Squares*), doravante IRLS. Este método é uma simplificação do algoritmo de Broyden-Fletcher-Goldfarb-Shanno (BFGS) proposto por Fletcher (1987), usando a matriz Hessiana esperada (Hardin et al., 2007).

Para os modelos potência e potência complementar foi primeiramente realizada a transformação $\zeta = e^{\varphi}$, dado que $\zeta \in \mathbb{R}^+$, com $\varphi \in \mathbb{R}$, fazendo-se uso do método delta (Doob, 1935) para se obter as estimativas da variância. As estimações dos parâmetros para os modelos flexíveis foram realizadas usando o procedimento iterativo de verosimilhança perfilada de Fischer (Fisher, 1956) usando iterativamente o algoritmo IRLS para obter as estimativas para os coeficientes das variáveis e o algoritmo de Pesquisa em Linha (Line Search)[1] para obter as estimativas de φ para os modelos potencia e potencia complementar e ζ para os restantes modelos flexíveis. Foi usada uma tolerância de 1×10^4 e um número máximo de iterações de 1×10^5, sendo as estimativas iniciais para β obtidas fazendo a estimação condicional em ζ, sendo usado o valor de ζ que mais

[1] Veja-se Dennis Jr. e Schnabel (1996) para uma apresentação mais detalhada deste algoritmo

se aproxima das distribuições apresentadas para os modelos *standard*, usando pesos iguais para todas as variáveis.

Nesta aplicação, com exceção do WESML, será utilizada a inversa da matriz de Informação de Fisher avaliada nas estimativas pontuais obtidas por máxima-verosimilhança para os parâmetros de interesse como estimador para a variância desses mesmos parâmetros. Esta escolha deve-se às boas propriedades deste estimador em amostras de pequena e média dimensão (Greene, 2018), bem como à facilidade na sua computação usando o *software* R (R Core Team, 2013).

Para o WESML a variância robusta será estimada por $\widehat{Avar(\beta)}_{rob} = H^{-1}(\hat{\beta})M(\hat{\beta})H^{-1}(\hat{\beta})$, tal como expresso em White (1982), onde[2]:

$$H(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \omega_i \, ln\big[f(y_i|\boldsymbol{x}_i, \hat{\beta})\big]}{\partial\hat{\beta}\,\partial\hat{\beta}^T} \tag{3.5}$$

$$M(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{\partial \omega_i \, ln\big[f(y_i|\boldsymbol{x}_i, \hat{\beta})\big]}{\partial\hat{\beta}}\right]\left[\frac{\partial \omega_i \, ln\big[f(y_i|\boldsymbol{x}_i, \hat{\beta})\big]}{\partial\hat{\beta}^T}\right]^T \tag{3.6}$$

3.2 Análise de Resultados

Os resultados reportados no apêndice C têm pelo menos uma taxa de convergência de 80% para todos os modelos. No caso em que esta percentagem não é atingida a média das estimativas pontuais para os coeficientes dos regressores não são apresentados. Para todos os resultados apresentados as estimativas pontuais para os coeficientes das variáveis *dummy* de localização não serão reportados. Adicionalmente, o sobrescrito nas estimativas dos *betas* significa que este coeficiente é estatisticamente significativo aos níveis usuais de significância[3] 1%, 5% e 10%, tal como é habitual.

Para os modelos *standard* observa-se uma divergência no resultado do teste RESET para $n = 1000$ e $n = 5000$, á exceção do *cauchit* para esta ultima dimensão amostral, em comparação com as estimativas para a população de treino, aos níveis habituais de significância, pois não rejeitamos a boa especificação destes modelos para as dimensões amostrais mais pequenas. Já para dimensões amostrais superiores esta divergência esbate-se, dada a evolução das estimativas do *p-value* em direção a zero. Observa-se o mesmo comportamento para os modelos flexíveis, nos casos em que foi

[2] Veja-se Zeileis (2006) para mais referências sobre a implementação deste estimador no software R.

[3] A utilização destes níveis de significância tem sofrido várias críticas, veja-se Benjamin et al. (2018) onde é sugerido a redução dos níveis de significância para $0,5\%$.

possível obter as estimativas do *p-value*. Esta tendência é bastante conhecida e reportada na literatura e deve-se ao facto de quando as estimativas são tão precisas (grande amostra) que qualquer desvio em relação a hipótese alternativa é estatisticamente significativo[4].

Quando se realiza seleção amostral o teste RESET não rejeita a nula aplicando-se ou não a correção para a seleção amostral, aos níveis habituais de significância, para os casos em que foi possível obter estimativas para o *p-value*. Para os modelos com seleção amostral corrigida o teste apresenta valores bastante próximos, mas consoante n aumenta e/ou $H_S \to 1/2$ o *p-value* do RESET aumenta. Quando a correção não é aplicada, os valores também aumentam nas mesmas condições, mas não tanto como como no caso anterior. Isto indica que o RESET não se comporta de forma correta dado que não deteta a inconsistência nos modelos onde seleção amostral não é corrigida.

As medidas de avaliação da qualidade de ajustamento apresentam valores a rondar $0,06$ e $0,6$ para o BS e AUC, respetivamente, independentemente do modelo utilizado ou da seleção amostral, principalmente o BS, que se mantém constante para quase todas as dimensões amostrais utilizadas. Isto indica falta de capacidade de avaliação da capacidade preditiva dos modelos por parte destas medidas.

Assim, não há neste contexto um referencial objetivo em termos de análise de especificação e de capacidade preditiva. A percentagem de uns (6%) poderá não ser ainda extrema ao ponto dos problemas identificados com eventos raros se tornarem aparentes. Dado que os modelos *standard* não foram rejeitados, a não ser no caso de em que a amostra é muito grande, (para este caso os resultados do teste RESET também são questionáveis, tal como explicado anteriormente), considerar-se-á que descrevem os dados e serão utilizados como base de comparação com os modelos com seleção e modelos flexíveis.

Avaliando a significância estatística de forma genérica para as variáveis explicativas observa-se que o número de variáveis estatisticamente significativas para os modelos *standard*, flexíveis e com seleção amostral, considerando a mesma dimensão amostral e independentemente do valor de H_S utilizado, é relativamente igual. Apenas com uma amostra mais pequena ($n = 1000$) e $H_S = 0,5$ se notou uma melhoria significativa no WESML em comparação com os modelos standard com a mesma dimensão amostral.

Avaliando o sinal dos coeficientes das variáveis explicativas estatisticamente significativas, observa-se uma concordância entre todos os modelos.

[4] Veja-se Lin et al. (2013) para uma discussão sobre este problema.

Apresenta-se agora uma análise mais profunda dos resultados obtidos para os diversos modelos. Para $n = 1000$ não foi possível estimar nenhum dos modelos flexíveis e para os modelos *standard* e WESML também não convergiram de forma sistemática. Note-se os valores inadmissíveis para o *cloglog* e *cauchit* para $H_S = 0,35$ e $H_S = 0,5$, com seleção amostral não corrigida, apesar de no último modelo mencionado o coeficiente afetado não ser estatisticamente significativo. Estes valores indicam que apesar de o *software* R ter reportado convergência na estimação, esta poderá não ter ocorrido, ou ocorreu num máximo local.

Nos modelos *standard* observa-se que não foi possível obter estimativas para o *cauchit* bem como uma estimativa do *p-value* para o teste RESET no *cloglog*. Para as dimensões amostrais superiores estas limitações em termos de capacidade de estimação vão diminuindo.

Observando a evolução da constante para os modelos standard com seleção amostral corrigida verifica-se que $\beta_0 \to -\infty$ consoante $H_S \to 1/2$ com a exceção de quando $n = 10000$. Quando a seleção amostral não é corrigida os valores da constante mantêm-se sempre abaixo (mais perto de zero) do que o valor das estimativas obtidas para a constante com seleção amostral corrigida, e para $n \geqslant 5000$ as estimativas são sempre inferiores as proporcionadas pela amostra aleatória.

Comparando as estimativas para o *logit* com seleção amostral observa-se que consoante Q_S se afasta de H_S, para um n constante, a divergência das estimativas pontuais para os coeficientes de referência aumenta. Adicionalmente, constata-se que conforme a dimensão amostral aumenta esta divergência esbate-se, mantendo o valor de H_S fixo. Também se verifica que as estimativas da variância para as variáveis explicativas do modelo *logit* obtidas do WESML são superiores ás ostentadas pelo *logit* sem correção para seleção amostral, revelando maior eficiência do último estimador[5].

Analisa-se agora o AMPE, tabela C.4, para todos os modelos estimados na população de treino e para os modelos com seleção amostral corrigida e não corrigida para $n = 10000$ com $H_S = 1/2$. Observa-se que nos modelos com seleção e sem a aplicação da correção os AMPEs são claramente superiores aos obtidos por todos os outros modelos. Estes resultados refletem a inconsistência dos coeficientes. Note-se que no caso do *logit*, essa inconsistência está restrita à constante, e ainda assim as distorções são importantes.

[5] Estes resultados são baseados e estão de acordo com as conclusões presentes em (Xie e Manski, 1989).

Observando os resultados dos AMPEs para o WESML nota-se que estes são ligeiramente superiores aos obtidos pelos modelos *standard*. Também se observa que as estimativas para os coeficientes para o WESML com $n = 10000$ e $H_s = 0,5$ são superiores (quando o coeficiente é negativo o WESML apresenta uma estimativa para esse coeficiente mais perto zero). Isto observa-se para outros valores de H_s, com $n = 10000$, mas acontece para a maioria das variáveis para as outras dimensões amostrais. Estes resultados são indicativos de que a redução do número de controlos na amostra de treino aleatória, para além de afetar os desvios padrão estimados, também afeta positivamente o efeito na probabilidade (em direção a um).

Capítulo 4

Conclusão

If you would know who controls you
see who you may not criticise.
Marcus Claudius Tacitus

Na presença de uma amostra com eventos raros o investigador não deve ignorar esta peculiaridade, pois a utilização dos modelos *standard* subestima as probabilidades associadas a cada evento enviesando-as para zero. A decisão entre utilizar modelos flexíveis ou modelos *standard* com seleção amostral pode parecer trivial, dado todos os resultados apresentados até agora. No entanto serão expostos mais fatores para uma melhor fundamentação desta decisão.

Primeiramente observa-se que os vários critérios de comparação apresentam valores muito próximos entre os vários modelos apresentados, pelo que se conclui que estas medidas não têm grande capacidade para avaliar a capacidade preditiva em amostras com eventos raros.

Os modelos flexíveis, apesar da sua atratividade teórica, pois utilizam toda a amostra recolhida e apresentam uma maior flexibilidade (maior aderência potencial aos dados), ostentam uma maior complexidade tanto em termos de interpretabilidade das estimativas dos coeficientes e parcimoniosidade, tempo de estimação e na dimensão amostral necessária para obter as estimativas dos coeficientes, sendo necessária uma amostra de média/grande dimensão, atendendo ao número de parâmetros a estimar, para ser possível obter estimativas pontuais. A utilização de valores diferentes como valores de partida para o algoritmo de otimização não garante que o mesmo convirja nas mesmas estimativas, fazendo os resultados divergir.

Para além disso, estes modelos ainda se encontram sub-estudados, principalmente

os modelos potência e potência complementar, sendo necessária mais investigação em relação às propriedades da inferência estatística realizadas nestes modelos. Note-se também que pode dar-se o caso em que dois modelos flexíveis não sejam rejeitados nos testes de boa especificação. Nesta situação o investigador poderá optar por utilizar um teste de hipóteses não encaixadas. Isto levantaria mais problemas computacionais, podendo os algoritmos utilizados não convergir, ou convergir no ponto errado.

Os modelos com seleção amostral são facilmente estimáveis, mas têm o revés de ser necessário a eliminação de controlos da amostra recolhida pelo investigador, ou a recolha de mais eventos diretamente na população (caso se saiba as percentagens de uns e zeros na população) selecionando-se aleatoriamente um grupo de zeros. Atendendo ao caso da seleção amostral, não existe na literatura muita informação sobre como esta deve de ser realizada numa amostra aleatória. As recomendações existentes, tanto quanto é do meu conhecimento, são para o caso em que se faz seleção amostral na população. Recomenda-se, se possível, a recolha de uma amostra aleatória sendo realizada a seleção amostral posteriormente, devendo esta ser realizada como apresentado na subsecção 3.1: utilizar todos os uns na amostra e selecionar um conjunto de zeros aleatoriamente dos zeros disponíveis, de forma a satisfazer as percentagens pretendidas destas observações na amostra usada para a estimação. Esta recomendação permite obter estimativas para a percentagem de uns na população, caso estas sejam desconhecidas, como já mencionado. Segundo este procedimento permite controlar e verificar a estabilidade dos resultados, evitando problemas de variabilidade amostral que possam produzir resultados anómalos. Observando os resultados do teste RESET para os modelos com seleção amostral não corrigida, conclui-se que este teste não consegue captar a inconsistência do estimador em CBS sem correção, tal como já indicado.

Tomando em atenção as várias considerações aqui levantadas, recomenda-se, para este conjunto de dados, a utilização dos modelos *standard* ou do WESML. Para uma amostra com maior grau de raridade recomenda-se a utilização de seleção amostral aplicando o estimador WESML, dado que mesmo com $Q_S = 6\%$ já se nota um pequeno aumento na probabilidade de um evento. Adicionalmente recomenda-se, que na presença de uma grande amostra se faça $H_S = 1/2$, dado que a informação perdida da redução do número de zeros será desprezível. Caso o investigador não tenha acesso a uma amostra com uma grande dimensão, para o número de variáveis presentes, recomenda-se que $H_S \in [0,2; 0,35]$ de forma a minimizar a informação perdida da redução do número de controlos e evitar problemas de convergência no algoritmo utilizado.

Bibliografia

Abadie, A, MM Chingos e MR West (2018). Endogenous stratification in randomized experiments. *Review of Economics and Statistics* **100**(4), 567–580.

Abanto-Valle, CA, JL Bazán e AC Smith (2014). State space mixed models for binary responses with skewed inverse links using JAGS. *Rio de Janeiro, Brazil*, 18.

Aranda-Ordaz, FJ (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68**(2), 357–363.

Bazán, JL, F Torres-Avilés, AK Suzuki e F Louzada (2017). Power and reversal power links for binary regressions: an application for motor insurance policyholders. *Applied Stochastic Models in Business and Industry* **33**(1), 22–34.

Benjamin, DJ, JO Berger, M Johannesson, BA Nosek, EJ Wagenmakers, R Berk, KA Bollen, B Brembs, L Brown, C Camerer et al. (2018). Redefine statistical significance. *Nature Human Behaviour* **2**(1), 6.

Bradley, AP (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* **30**(7), 1145–1159.

Brier, GW (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review* **78**(1), 1–3.

Calabrese, R e SA Osmetti (2015). Improving forecast of binary rare events data: a GAM-based approach. *Journal of Forecasting* **34**(3), 230–239.

Cameron, AC e PK Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Caron, R e A Polpo (2009). Binary data regression: Weibull distribution. Em: *AIP Conference Proceedings*.

Caron, R, D Sinha, D Dey e A Polpo (2018). Categorical data analysis using a skewed Weibull regression model. *Entropy* **20**(3), 176.

Chen, MH, DK Dey e QM Shao (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association* **94**(448), 1172–1186.

Cosslett, SR (1981a). «Structural analysis of discrete data with econometric applications». Em: ed. por C Manski e D McFadden. MIT Press. Cap. Ecient estimation of discrete-choice models, pp. 51–111.

Cosslett, SR (1981b). Maximum likelihood estimator for choice-based samples. *Econometrica* **49**(5), 1289–1316.

Cramer, JS (1999). Predictive performance of the binary logit model in unbalanced samples. *Journal of the Royal Statistical Society: Series D (The Statistician)* **48**(1), 85–94.

Czado, C e TJ Santner (1992). The effect of link misspecification on binary regression inference. *Journal of statistical planning and inference* **33**(2), 213–231.

Dennis Jr., JE e RB Schnabel (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. Vol. 16. Siam.

Doob, JL (1935). The limiting distributions of certain statistics. *The Annals of Mathematical Statistics* **6**(3), 160–169.

Elliott, G e A Timmermann (2013). *Handbook of economic forecasting*. Vol. 2. Elsevier.

Fawcett, T (2006). An introduction to ROC analysis. *Pattern recognition letters* **27**(8), 861–874.

Fisher, RA (1956). *Statistical Methods and Scientific Inference*. Oliver e Boyd.

Fletcher, R (1987). *Practical methods of optimization*. John Wiley & Sons.

Freedman, DA (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician* **60**(4), 299–302.

Greene, WH (2018). *Econometric Analysis*. Pearson.

Gupta, RC e RD Gupta (2004). Generalized skew normal model. *Test* **13**(2), 501–524.

Hanley, JA e BJ McNeil (1982). The meaning and use of the area under a receiver operating characteristic ROC curve. *Radiology* **143**(1), 29–36.

Hardin, JW, JW Hardin, JM Hilbe e J Hilbe (2007). *Generalized linear models and extensions*. Stata press.

Hosmer, DW e S Lemeshow (2000). *Applied logistic regression*. Wiley New York.

Hsieh, DA, CF Manski e D McFadden (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association* **80**(391), 651–662.

Huber, PJ (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Em: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pp.221–233.

Imbens, GW (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica* **60**(5), 1187–1214.

Imbens, GW e T Lancaster (1996). Efficient estimation and stratified sampling. *Journal of Econometrics* **74**(2), 289–318.

Jiang, X, DK Dey, R Prunier, AM Wilson e KE Holsinger (2013). A new class of flexible link functions with application to species co-occurrence in cape floristic region. *The Annals of Applied Statistics* **7**(4), 2180–2204.

King, G e L Zeng (2001). Logistic regression in rare events data. *Political Analysis* **9**(2), 137–163.

Kleinbaum, DG e M Klein (2010). *Logistic Regression: A Self-Learning*. Springer.

Koenker, R e J Yoon (2009). Parametric links for binary choice models: A Fisherian-Bayesian colloquy. *Journal of Econometrics* **152**(2), 120–130.

Leeper, TJ (2017). Interpreting regression results using average marginal effects with R's margins. *Available at the comprehensive R Archive Network (CRAN)*.

Lemonte, AJ e JL Bazán (2018). New links for binary regression: an application to coca cultivation in Peru. *Test* **27**(3), 597–617.

Lin, M, HC Lucas Jr. e G Shmueli (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research* **24**(4), 906–917.

Maalouf, M, TB Trafalis e I Adrianto (2011). Kernel logistic regression using truncated newton method. *Computational management science* **8**(4), 415–428.

Mai, AT, F Bastin e M Toulouse (2014). *On Optmization Algorithms for Maximum Likelihood Estimation*. Rel. téc. CIRRELT, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport.

Manski, CF e SR Lerman (1977). The estimation of choice probabilities from choice based samples. *Econometrica* **45**(8), 1977–1988.

Manski, CF e D McFadden (1981). *Structural analysis of discrete data with econometric applications*. Ed. por CF Manski e D McFadden. MIT press Cambridge, MA.

Nelder, JA e RWM Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135**(3), 370–384.

Ogundimu, E (2019). Prediction of default probability by using statistical models for rare events. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Owen, AB (2007). Infinitely imbalanced logistic regression. *Journal of Machine Learning Research* **8**(Apr), 761–773.

Palepu, KG (1986). Predicting takeover targets: A methodological and empirical analysis. *Journal of accounting and economics* **8**(1), 3–35.

Pregibon, D (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **29**(1), 15–24.

Prentice, RL e R Pyke (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**(3), 403–411.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramalho, E e J Ramalho (2012). Alternative versions of the RESET test for binary response index models: a comparative study. *Oxford bulletin of economics and statistics* **74**(1), 107–130.

Ramalho, E, J Ramalho e JMR Murteira (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys* **25**(1), 19–68.

Ramsey, JB (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)* **31**(2), 350–371.

Rinne, H (2008). *The Weibull distribution: A Handbook*. Chapman e Hall/CRC.

Rubino, G e B Tuffin (2009). An Introduction to Monte Carlo Methods and Rare Event Simulation. Em: *QEST*.

Taneichi, N, Y Sekiya e J Toyama (2014). A new family of parametric links for binomial generalized linear models. *Journal of the Japan Statistical Society* **44**(2), 119–133.

Taylor, JMG, AL Siqueira e RE Weiss (1996). The cost of adding parameters to a model. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(3), 593–607.

Weiss, GM (2004). Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter* **6**(1), 7–19.

Wen, JF e DV Gordon (2014). An empirical model of tax convexity and self-employment. *Review of Economics and Statistics* **96**(3), 471–482.

Westphal, C (2013). *Logistic regression for extremely rare events: The case of school shootings*. Rel. téc. Joint Discussion Paper Series in Economics.

White, H (1982). Maximum likelihood estimation of misspecified models. *Econometrica*.

Winkelmann, R e S Boes (2006). *Analysis of microdata*. Springer Science & Business Media.

Xie, Y e CF Manski (1989). The logit model and response-based samples. *Sociological Methods & Research* **17**(3), 283–302.

Zeileis, A (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software* **16**(9).

Apêndice A

Gráficos

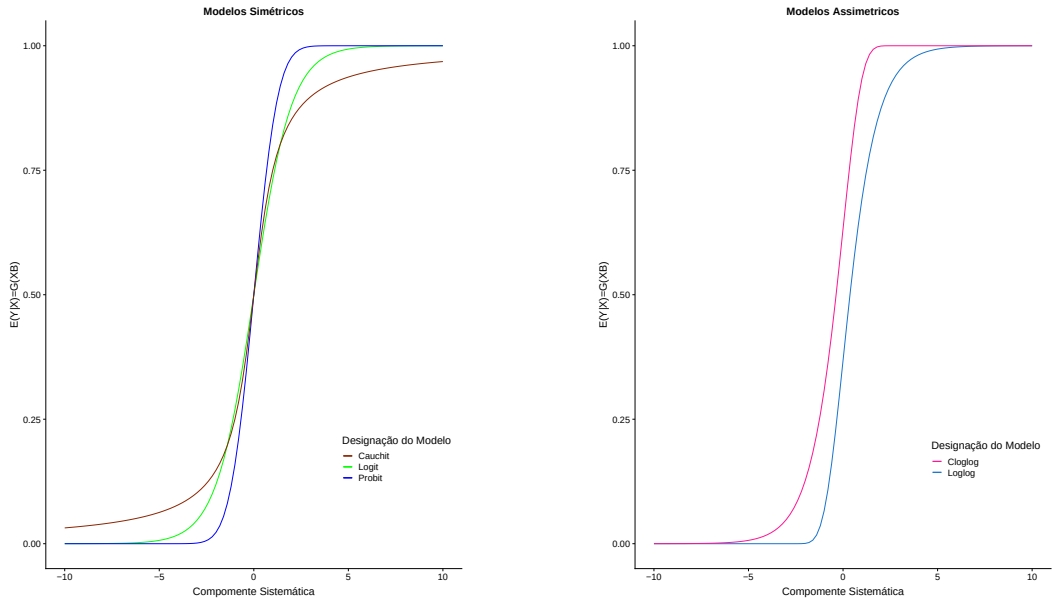

Figura A.1: Modelos *standard* com componente sistemática dada por: $\eta = 1 + 2x$.

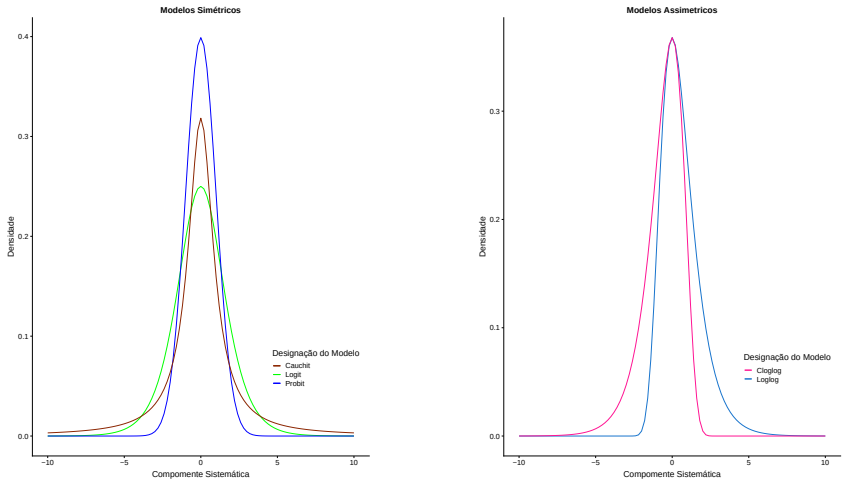

Figura A.2: Função distribuição da probabilidade, com componente sistemática dada por: $\eta = 1 + 2x$.

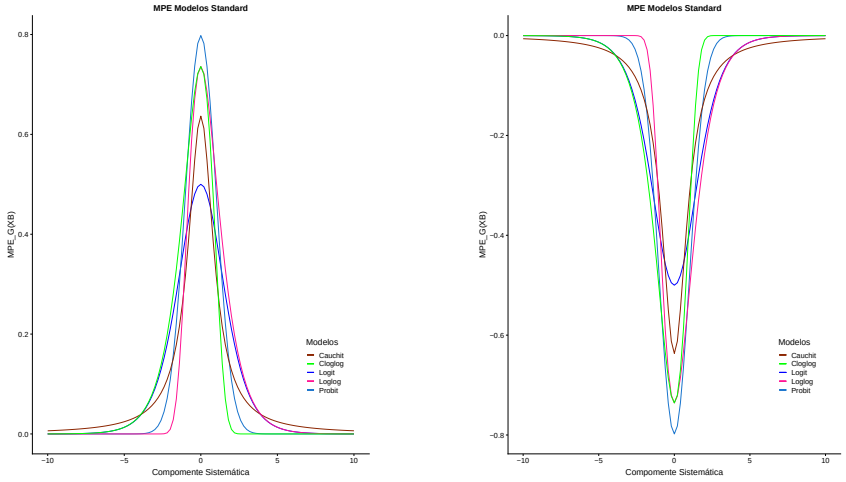

Figura A.3: MPE para os modelos *standard* com componentes sistémicas dadas por $\eta = 1 + 2x$ e $\eta = 1 - 2x$ respetivamente.

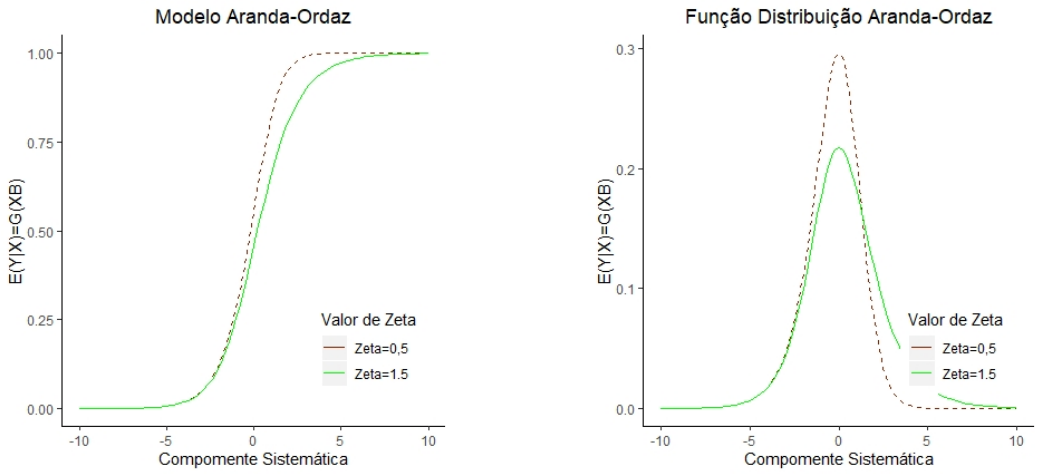

Figura A.4: Função distribuição acumulada Aranda-Ordaz e função distribuição (da probabilidade) Aranda-Ordaz com a componente sistémica dada por $\eta = 1 + 2x$ e com $\zeta = 1/2$ e $\zeta = 3/2$

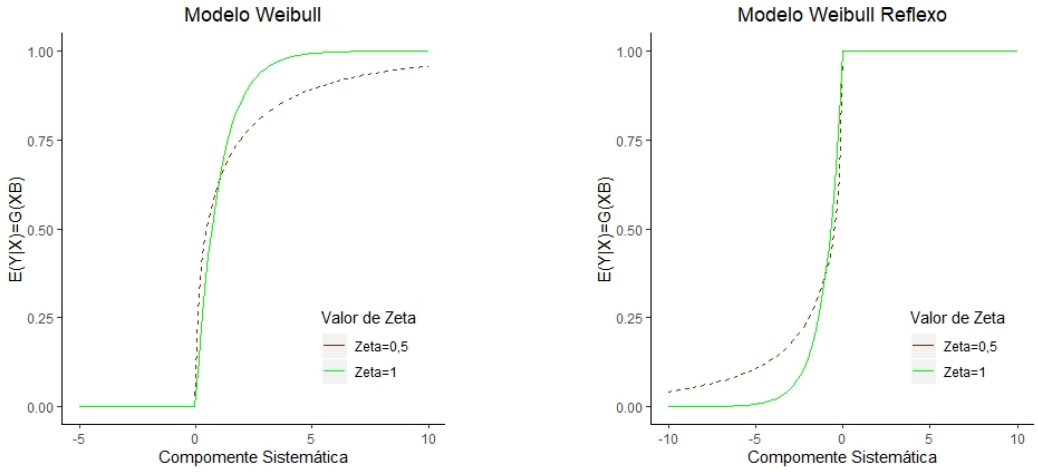

Figura A.5: Função distribuição Weibull e Weibull Reflexo com componentes sistémicas dadas por $\eta = 1 + 2x$ e $\eta = 1 - 2x$ respetivamente

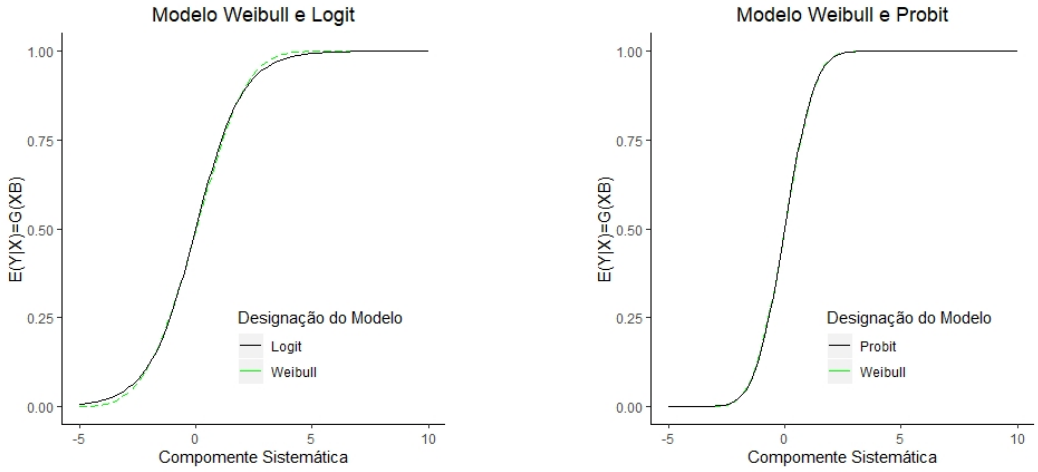

Figura A.6: Aproximações do modelos Weibull ao *logit* e *probit*, respetivamente.

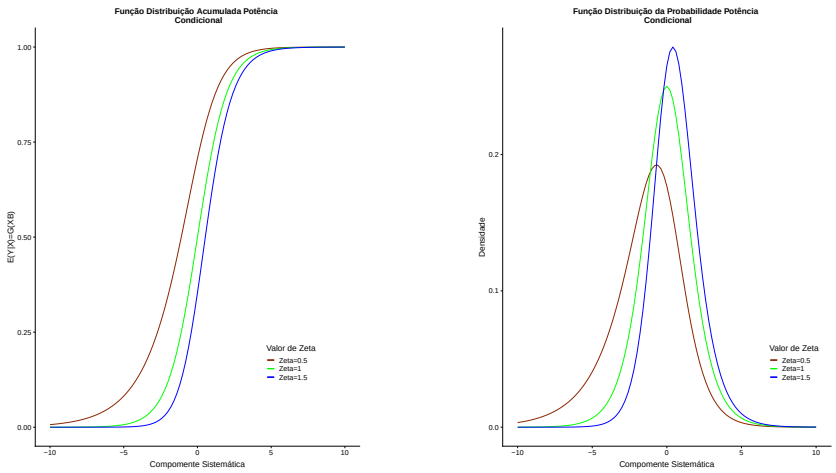

Figura A.7: Função distribuição acumulada potência e função densidade da probabilidade potência condicionada em ζ, com componente sistemática dada por: $\eta = 1 + 2x$.

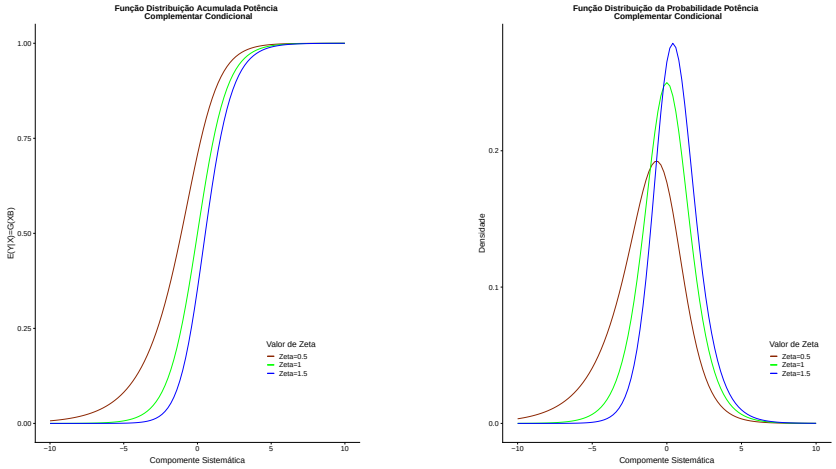

Figura A.8: Função distribuição acumulada potência complementar e função densidade da probabilidade potência complementar condicionada em ζ, com componente sistemática dada por: $\eta = 1 + 2x$.

Apêndice B

Estatísticas descritivas e descrição das variáveis

Tabela B.1: Descrição das variáveis

Variável	Definição
Self-Employed	Variável *dummy* = 1 se a principal forma de rendimento advém de rendimentos obtidos por conta própria
Rendimento de Investimentos	Rendimento obtido de Investimentos em milhares de euros
Ganhos de Capital	Rendimentos oriundos de aumento no valor de um ativo em milhares de euros
Escolaridade	número de anos de escolaridade
Idade	Idade do indivíduo em dezenas de anos
Doenças	Variável *dummy* = 1 se o indivíduo sofre de doenças crómicas.
Homem	Variável *dummy* = 1 se o indivíduo sofre é do sexo masculino.
Casado	Variável *dummy* = 1 se o indivíduo é casado.
número de filhos	número de filhos no agregado familiar.
Semanas desempregado	número de semanas que o indivíduo se encontrou desempregado no ano.
Operário	Variável *dummy* = 1 se o indivíduo for da classe operaria.
Serviços	Variável *dummy* = 1 se o indivíduo trabalha no setor terciário
Dummys de localização	Oito variáveis de localização que correspondem à residência do indivíduo i. Cada variável *dummy* = 1 se o indivíduo i reside nessa província.

Tabela B.2: Estatísticas resumidas para a amostra de treino

	Self-Employd=1		Self-Employd=0	
Variáveis	Média	Desvio padrão	Média	Desvio padrão
Rendimento de Investimentos	1.546	7.175	1.028	5.677
Ganhos de Capital	0.742	7.059	0.476	5.819
Escolaridade	13.491	3.376	13.666	3.004
Idade	4.429	0.999	4.014	1.117
Doenças	0.148	0.355	0.136	0.343
Homem	0.62	0.486	0.524	0.499
Casado	0.81	0.393	0.701	0.458
número de filhos	0.846	1.097	0.77	1.041
Semanas desempregado	0.375	2.881	1.144	4.855
Operário	0.307	0.461	0.272	0.445
Serviços	0.242	0.428	0.219	0.413
Total	7358		114572	
Proporção	0.06		0.94	

Tabela B.3: Estatísticas resumidas para toda a amostra

	Self-Employd=1		Self-Employd=0	
Variáveis	Média	Desvio padrão	Média	Desvio padrão
Rendimento de Investimentos	1.555	7.26	1.041	5.695
Ganhos de Capital	0.721	6.74	0.476	5.701
Escolaridade	13.508	3.352	13.662	2.981
Idade	4.451	1.009	4.033	1.124
Doenças	0.155	0.362	0.141	0.348
Homem	0.62	0.485	0.523	0.499
Casado	0.807	0.395	0.7	0.458
Nùmero de filhos	0.826	1.092	0.762	1.038
Semanas desempregado	0.377	2.909	1.134	4.826
Operário	0.309	0.462	0.27	0.444
Serviços	0.235	0.424	0.216	0.412
Total	8677		133652	
Proporção	0.061		0.939	

Apêndice C

Resultados

C.1 Resultados dos Modelos *Standard*

	n =1000					n =5000				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.082	-0.263	-	-0.259	-0.046	-0.015*	-0.036*	-0.485	-0.035*	-0.009
	(2.454)	(12.961)	-	(12.911)	(0.916)	(0.009)	(0.021)	(0.322)	(0.021)	(0.006)
Homem	0.221	0.432	-	0.409	0.168	0.166**	0.336**	1.125**	0.325**	0.12***
	(0.145)	(0.293)	-	(0.277)	(0.106)	(0.065)	(0.135)	(0.51)	(0.129)	(0.046)
N. filhos	0.034	0.072	-	0.069	0.023	0.048	0.103*	0.382**	0.1*	0.032
	(0.066)	(0.132)	-	(0.124)	(0.049)	(0.03)	(0.061)	(0.185)	(0.058)	(0.022)
R. Investimentos	0.005	0.009	-	0.009	0.004	0.007*	0.014	0.033	0.01	0.007**
	(0.012)	(0.853)	-	(0.821)	(0.009)	(0.004)	(0.354)	(1.866)	(0.341)	(0.003)
G. Capital	-0.018	-0.038	-	-0.037	-0.013	0.003	0.005	0.011	0.004	0.003
	(0.029)	(0.065)	-	(0.063)	(0.02)	(0.003)	(0.006)	(0.014)	(0.005)	(0.003)
Escolaridade	0.076	0.171	-	0.175	0.049	-0.09*	-0.176*	-0.42	-0.165*	-0.069*
	(0.155)	(0.316)	-	(0.298)	(0.115)	(0.051)	(0.099)	(0.272)	(0.093)	(0.038)
Escolaridade^{2}	-0.002	-0.004	-	-0.004	-0.001	0.004**	0.007*	0.017*	0.007**	0.003***
	(0.005)	(0.011)	-	(0.01)	(0.004)	(0.002)	(0.004)	(0.01)	(0.003)	(0.001)
Idade	0.422	0.8	-	0.738	0.328	0.623***	1.348***	6.559***	1.292***	0.429***
	(0.46)	(0.949)	-	(0.897)	(0.331)	(0.207)	(0.449)	(2.286)	(0.433)	(0.143)
Idade^{2}	-0.03	-0.055	-	-0.05	-0.024	-0.053**	-0.114**	-0.585**	-0.109**	-0.036**
	(0.055)	(0.112)	-	(0.105)	(0.04)	(0.024)	(0.052)	(0.243)	(0.05)	(0.017)
Doenças	-0.232	-0.479	-	-0.452	-0.166	-0.033	-0.068	-0.157	-0.065	-0.022
	(0.189)	(0.388)	-	(0.368)	(0.137)	(0.081)	(0.167)	(0.516)	(0.159)	(0.059)
Casado	0.248	0.496	-	0.466	0.186	0.069	0.151	0.624	0.147	0.046
	(0.172)	(0.361)	-	(0.347)	(0.122)	(0.076)	(0.161)	(0.71)	(0.155)	(0.053)
Operário	0.137	0.252	-	0.219	0.109	0.126	0.262	0.798	0.245	0.089
	(0.172)	(0.342)	-	(0.321)	(0.128)	(0.078)	(0.16)	(0.549)	(0.153)	(0.056)
Serviços	0.231	0.437	-	0.402	0.182	0.269***	0.545***	1.545***	0.513***	0.196***
	(0.174)	(0.347)	-	(0.326)	(0.129)	(0.077)	(0.157)	(0.515)	(0.15)	(0.056)
Constante	-3.802***	-7.268**	-	-7.103**	-2.681**	-2.943***	-5.862***	-21.634***	-5.798***	-1.947***
	(1.446)	(2.983)	-	(2.819)	(1.052)	(0.536)	(1.132)	(5.578)	(1.085)	(0.381)
Reset-pvalues	0.475	0.09	-	-	0.18	0.134	0.496	0	0.54	0.552
B. Score	0.061	0.103	-	0.061	0.061	0.06	0.06	0.06	0.06	0.06
AUC	0.594	0.594	-	0.593	0.595	0.631	0.631	0.632	0.631	0.631

	n =10000					Toda a amostra				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.021***	-0.053***	-0.705**	-0.052***	-0.013***	-0.022***	-0.053***	-0.902***	-0.053***	-0.013***
	(0.007)	(0.018)	(0.313)	(0.018)	(0.004)	(0.002)	(0.005)	(0.112)	(0.005)	(0.001)
Homem	0.171***	0.348***	1.13***	0.332***	0.124***	0.178***	0.365***	1.245***	0.349***	0.129***
	(0.045)	(0.094)	(0.362)	(0.091)	(0.032)	(0.013)	(0.027)	(0.118)	(0.026)	(0.009)
N. filhos	0.043**	0.093**	0.344**	0.09**	0.03**	0.033***	0.072***	0.292***	0.069***	0.022***
	(0.021)	(0.043)	(0.136)	(0.041)	(0.015)	(0.006)	(0.012)	(0.043)	(0.012)	(0.004)
R. Investimentos	0.008***	0.014	0.031	0.012	0.006***	0.004***	0.008	0.014	0.008	0.003***
	(0.003)	(0.264)	(1.424)	(0.256)	(0.002)	(0.001)	(0.076)	(0.377)	(0.074)	(0.001)
G. Capital	0.002	0.005	0.011*	0.004	0.002	0.001	0.003	0.007**	0.003***	0.001
	(0.002)	(0.004)	(0.006)	(0.004)	(0.002)	(0.001)	(0.002)	(0.003)	(0.001)	(0.001)
Escolaridade	-0.1***	-0.189***	-0.427***	-0.177***	-0.079***	-0.097***	-0.181***	-0.333***	-0.168***	-0.077***
	(0.035)	(0.067)	(0.157)	(0.063)	(0.027)	(0.01)	(0.019)	(0.052)	(0.018)	(0.008)
Escolaridade^2	0.004***	0.007***	0.017***	0.007***	0.003***	0.004***	0.007***	0.013***	0.006***	0.003***
	(0.001)	(0.002)	(0.006)	(0.002)	(0.001)	(0)	(0.001)	(0.002)	(0.001)	(0)
Idade	0.591***	1.288***	6.387***	1.244***	0.4***	0.439***	0.951***	4.639***	0.924***	0.299***
	(0.144)	(0.311)	(1.572)	(0.3)	(0.099)	(0.041)	(0.089)	(0.447)	(0.086)	(0.028)
Idade^2	-0.05***	-0.109***	-0.559***	-0.105***	-0.033***	-0.034***	-0.075***	-0.373***	-0.073***	-0.023***
	(0.017)	(0.036)	(0.166)	(0.035)	(0.012)	(0.005)	(0.01)	(0.047)	(0.01)	(0.003)
Doenças	-0.012	-0.033	-0.256	-0.032	-0.005	-0.02	-0.046	-0.198	-0.045	-0.013
	(0.057)	(0.116)	(0.394)	(0.111)	(0.041)	(0.017)	(0.035)	(0.126)	(0.033)	(0.012)
Casado	0.082	0.174	0.553	0.168	0.057	0.128***	0.266***	1.067***	0.256***	0.091***
	(0.053)	(0.112)	(0.492)	(0.108)	(0.037)	(0.015)	(0.033)	(0.167)	(0.032)	(0.011)
Operário	0.109**	0.224**	0.581	0.212**	0.078*	0.08***	0.17***	0.702***	0.164***	0.056***
	(0.055)	(0.113)	(0.386)	(0.108)	(0.04)	(0.016)	(0.033)	(0.122)	(0.032)	(0.011)
Serviços	0.279***	0.557***	1.362***	0.525***	0.205***	0.22***	0.44***	1.257***	0.419***	0.162***
	(0.053)	(0.109)	(0.364)	(0.104)	(0.039)	(0.016)	(0.032)	(0.121)	(0.031)	(0.011)
Constante	-2.731***	-5.484***	-20.782***	-5.465***	-1.765***	-2.393***	-4.746***	-18.034***	-4.771***	-1.54***
	(0.369)	(0.776)	(3.761)	(0.742)	(0.265)	(0.106)	(0.222)	(1.08)	(0.213)	(0.076)
Reset-pvalues	0.001	0.004	0.00	0.006	0.178	0	0	0	0	0
B. Score	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
AUC	0.635	0.635	0.636	0.635	0.635	0.639	0.64	0.641	0.64	0.64

C.2 Resultados dos Modelos com Seleção Amostral

Tabela C.1: Estimativas para $n = 1000$ com $H_S = 15\%$ e $H_S = 25\%$, respetivamente.

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.082***	-0.253***	-	-0.252***	-0.046***	-0.092	-0.26	-38118.4	-0.253	-1.808*
	(0.027)	(0.064)	-	(0.062)	(0.017)	(2.909)	(12.193)	(12154634.017)	(12.373)	(0.946)
Homem	0.246	0.481	-	0.454	0.187*	0.277	0.483	0.746	0.426	0.07
	(0.155)	(0.315)	-	(0.298)	(0.112)	(0.174)	(0.316)	(0.53)	(0.281)	(0.145)
N. filhos	0.034	0.075	-	0.076	0.022	0.035	0.068	0.133	0.067	-0.025
	(0.071)	(0.144)	-	(0.135)	(0.052)	(0.08)	(0.143)	(0.214)	(0.126)	(0.068)
R. Investimentos	0.008	0.014	-	0.013	0.006	0.009	0.015	0.009	0.011	0.006
	(0.015)	(0.88)	-	(0.848)	(0.012)	(0.019)	(0.899)	(1.759)	(0.82)	(0.017)
G. Capital	-0.019	-0.044	-	-0.043	-0.013	-0.022	-0.042	-0.134	-0.041	-0.015
	(0.025)	(0.058)	-	(0.056)	(0.017)	(0.041)	(0.079)	(0.25)	(0.074)	(0.032)
Escolaridade	0.115	0.254	-	0.257	0.076	0.13	0.244	0.729	0.244	0.053
	(0.163)	(0.34)	-	(0.322)	(0.119)	(0.19)	(0.349)	(0.692)	(0.311)	(0.16)
Escolaridade^2	-0.003	-0.007	-	-0.007	-0.002	-0.003	-0.006	-0.022	-0.006	-0.003
	(0.006)	(0.012)	-	(0.011)	(0.004)	(0.007)	(0.012)	(0.024)	(0.011)	(0.006)
Idade	0.507	0.972	-	0.9	0.387	0.607	1.066	1.163	0.926	0.015
	(0.473)	(0.98)	-	(0.925)	(0.34)	(0.54)	(1.008)	(1.822)	(0.9)	(0.438)
Idade^2	-0.042	-0.078	-	-0.071	-0.033	-0.052	-0.089	-0.07	-0.076	0.016
	(0.058)	(0.118)	-	(0.111)	(0.042)	(0.064)	(0.119)	(0.204)	(0.105)	(0.053)
Doenças	-0.209	-0.435	-	-0.417	-0.148	-0.228	-0.424	-0.952	-0.385	0.069
	(0.21)	(0.43)	-	(0.408)	(0.152)	(0.226)	(0.417)	(0.791)	(0.375)	(0.186)
Casado	0.263	0.523	-	0.49	0.198	0.299	0.536	1.105	0.471	0.282*
	(0.185)	(0.392)	-	(0.377)	(0.13)	(0.204)	(0.384)	(0.89)	(0.351)	(0.164)
Operário	0.128	0.245	-	0.221	0.098	0.148	0.262	0.455	0.221	0.275
	(0.188)	(0.376)	-	(0.354)	(0.139)	(0.208)	(0.374)	(0.566)	(0.329)	(0.177)
Serviços	0.258	0.497	-	0.466	0.199	0.293	0.507	0.777	0.443	0.106
	(0.194)	(0.387)	-	(0.362)	(0.143)	(0.209)	(0.379)	(0.599)	(0.333)	(0.176)
Constante	-4.232***	-8.2***	-	-8.023***	-2.976***	-4.222**	-7.491**	-13.022**	-7.097**	-1.524
	(1.499)	(3.158)	-	(2.994)	(1.077)	(1.748)	(3.261)	(6.444)	(2.904)	(1.447)
Reset-pvalues	0.91	0.899	-	0.851	0.92	-	-	0.277	0.29	0.493
B. Score	0.061	0.061	-	0.061	0.061	0.07	0.07	0.068	0.07	0.073
AUC	0.591	0.591	-	0.59	0.593	0.591	0.591	0.58	0.59	0.609

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.08***	-0.255***	-	-0.25***	-0.045**	-0.094	-11.812	-50417.3	-0.257	-0.062
	(0.029)	(0.069)	-	(0.065)	(0.019)	(2.686)	(13.648)	(22154963.575)	(14.266)	(1.09)
Homem	0.267	0.523	-	0.494	0.204	0.319	0.272	0.63	0.435	0.314*
	(0.172)	(0.35)	-	(0.328)	(0.125)	(0.199)	(0.345)	(0.451)	(0.288)	(0.182)
N. filhos	0.03	0.071	-	0.075	0.019	0.029	-0.052	0.085	0.058	0.021
	(0.08)	(0.161)	-	(0.151)	(0.058)	(0.091)	(0.157)	(0.184)	(0.129)	(0.085)
R. Investimentos	0.008	0.015	-	0.015	0.006	0.009	0.023	0.013	0.011	0.008
	(0.017)	(0.948)	-	(0.912)	(0.014)	(0.022)	(0.952)	(1.5)	(0.824)	(0.022)
G. Capital	-0.017	-0.037	-	-0.036	-0.011	-0.022	-0.022	-0.104	-0.036	-0.018
	(0.034)	(0.071)	-	(0.066)	(0.024)	(0.051)	(0.094)	(0.19)	(0.08)	(0.044)
Escolaridade	0.145	0.311	-	0.309	0.099	0.177	0.184	0.438	0.292	0.152
	(0.179)	(0.369)	-	(0.345)	(0.133)	(0.213)	(0.376)	(0.541)	(0.314)	(0.198)
Escolaridade^2	-0.004	-0.009	-	-0.009	-0.003	-0.005	-0.009	-0.014	-0.009	-0.004
	(0.006)	(0.013)	-	(0.012)	(0.005)	(0.007)	(0.013)	(0.019)	(0.011)	(0.007)
Idade	0.5	0.935	-	0.843	0.394	0.688	0.286	0.953	0.915	0.666
	(0.516)	(1.06)	-	(0.994)	(0.376)	(0.607)	(1.079)	(1.477)	(0.902)	(0.54)
Idade^2	-0.039	-0.07	-	-0.06	-0.032	-0.058	0.006	-0.064	-0.074	-0.057
	(0.063)	(0.128)	-	(0.12)	(0.047)	(0.073)	(0.128)	(0.169)	(0.106)	(0.066)
Doenças	-0.294	-0.636	-	-0.625	-0.203	-0.327	0.223	-0.97	-0.517	-0.278
	(0.233)	(0.484)	-	(0.462)	(0.169)	(0.254)	(0.449)	(0.663)	(0.382)	(0.225)
Casado	0.279	0.561	-	0.529	0.21	0.337	0.613	0.936	0.481	0.319
	(0.206)	(0.441)	-	(0.423)	(0.143)	(0.232)	(0.413)	(0.667)	(0.36)	(0.201)
Operário	0.193	0.388	-	0.367	0.142	0.229	0.684*	0.528	0.351	0.209
	(0.212)	(0.425)	-	(0.397)	(0.157)	(0.239)	(0.409)	(0.486)	(0.334)	(0.224)
Serviços	0.29	0.553	-	0.516	0.228	0.354	0.346	0.688	0.482	0.359
	(0.218)	(0.438)	-	(0.41)	(0.161)	(0.239)	(0.413)	(0.533)	(0.34)	(0.221)
Constante	-4.436***	-8.535**	-	-8.293***	-3.162***	-4.509**	-4.376	-8.648*	-6.894**	-3.863**
	(1.636)	(3.395)	-	(3.183)	(1.202)	(1.959)	(3.477)	(4.994)	(2.893)	(1.796)
Reset-pvalues	0.883	0.849	-	0.801	0.86	0.297	-	0.298	0.303	-
B. Score	0.061	0.061	-	0.062	0.061	0.093	0.096	0.086	0.091	0.095
AUC	0.591	0.591	-	0.591	0.593	0.591	0.612	0.582	0.589	0.593

Tabela C.2: Estimativas para $n = 1000$ com $H_S = 35\%$ e $H_S = 50\%$, respetivamente.

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.086**	-0.263***	-	-0.256***	-0.05*	-0.107	-0.284	-27956.5	-22765612543.92***	-0.08
	(0.035)	(0.079)	-	(0.071)	(0.03)	(3)	(14.38)	(10839399.298)	(366.101)	(1.04)
Homem	0.276	0.551	-	0.531	0.206	0.336	0.549	0.523	5487490030.876***	0.35*
	(0.184)	(0.379)	-	(0.356)	(0.134)	(0.216)	(0.367)	(0.409)	(2339.296)	(0.21)
N. filhos	0.031	0.081	-	0.091	0.016	0.025	0.042	0.01	66129675336.661***	0.012
	(0.087)	(0.177)	-	(0.166)	(0.063)	(0.1)	(0.168)	(0.181)	(1192.156)	(0.098)
R. Investimentos	0.013	0.026	-	0.025	0.01	0.013	0.023	0.017	-449580012.081***	0.014
	(0.02)	(0.985)	-	(0.95)	(0.016)	(0.025)	(0.978)	(1.329)	(6213.611)	(0.028)
G. Capital	-0.019	-0.04	-	-0.038	-0.013	-0.027	-0.046	-0.05	-15588512036.135***	-0.024
	(0.038)	(0.079)	-	(0.074)	(0.03)	(0.058)	(0.102)	(0.136)	(894.215)	(0.055)
Escolaridade	0.149	0.319	-	0.316	0.103	0.189	0.313	0.288	8445865991.404***	0.175
	(0.199)	(0.412)	-	(0.384)	(0.147)	(0.236)	(0.408)	(0.52)	(1959.214)	(0.233)
Escolaridade^2	-0.004	-0.009	-	-0.009	-0.003	-0.005	-0.008	-0.008	681870943.146***	-0.004
	(0.007)	(0.015)	-	(0.014)	(0.005)	(0.008)	(0.014)	(0.018)	(72.053)	(0.008)
Idade	0.449	0.837	-	0.75	0.358	0.657	1.088	1.156	-165823731728.642***	0.698
	(0.563)	(1.161)	-	(1.089)	(0.409)	(0.668)	(1.157)	(1.389)	(7729.01)	(0.632)
Idade^2	-0.034	-0.06	-	-0.05	-0.029	-0.055	-0.092	-0.105	29123102389.796***	-0.061
	(0.069)	(0.14)	-	(0.131)	(0.051)	(0.08)	(0.137)	(0.16)	(948.453)	(0.077)
Doenças	-0.305	-0.638	-	-0.607	-0.217	-0.353	-0.615	-0.868	-115855712747.249***	-0.332
	(0.247)	(0.508)	-	(0.48)	(0.179)	(0.275)	(0.474)	(0.587)	(3544.342)	(0.258)
Casado	0.279	0.543	-	0.499	0.216	0.361	0.605	0.774	78079439322.002***	0.375
	(0.219)	(0.467)	-	(0.445)	(0.154)	(0.252)	(0.437)	(0.57)	(2534.171)	(0.233)
Operário	0.222	0.462	-	0.445	0.159	0.259	0.44	0.502	-5196463649.572***	0.242
	(0.233)	(0.469)	-	(0.437)	(0.172)	(0.263)	(0.442)	(0.472)	(3026.159)	(0.263)
Serviços	0.295	0.574	-	0.544	0.228	0.363	0.594	0.634	18785146464.44***	0.391
	(0.239)	(0.481)	-	(0.449)	(0.176)	(0.263)	(0.445)	(0.491)	(2898.641)	(0.257)
Constante	-4.396**	-8.504**	-	-8.283**	-3.132**	-4.396**	-7.256*	-7.03	-207080532944.071***	-4.019*
	(1.814)	(3.758)	-	(3.502)	(1.334)	(2.165)	(3.756)	(4.718)	(19592.633)	(2.11)
Reset-pvalues	0.856	0.829	-	0.755	0.899	0.39	-	0.492	0.413	-
B. Score	0.062	0.062	-	0.062	0.062	0.119	0.117	0.086	0.364	0.123
AUC	0.587	0.587	-	0.588	0.587	0.585	0.586	0.582	0.534	0.586

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.089	-0.271	-	-0.265	-	-0.115***	-0.297	-22056.7	-3.9E+09	-0.092
	(0.046)	(0.096)	-	(0.086)	-	(3.226)	(15.462)	(9687214.147)	(148.371)	(1.26)
Homem	0.295**	0.608**	-	0.589**	-	0.358***	0.58**	0.58**	1368038372.921**	0.401**
	(0.211)	(0.436)	-	(0.406)	-	(0.24)	(0.403)	(0.42)	(774.166)	(0.249)
N. filhos	0.045	0.116	-	0.128	-	0.033	0.054	0.062	-8.1E+09	0.018
	(0.101)	(0.205)	-	(0.189)	-	(0.112)	(0.186)	(0.187)	(381.733)	(0.118)
R. Investimentos	0.012	0.026	-	0.026	-	0.009	0.016	0.028	-1E+10	0.008
	(0.021)	(1.051)	-	(0.993)	-	(0.029)	(25.603)	(1.097)	(2002.573)	(0.033)
G. Capital	-0.007*	-0.016*	-	-0.016**	-	-0.013	-0.02	-0.023	2.1E+10	-0.013
	(0.035)	(0.076)	-	(0.069)	-	(0.057)	(0.099)	(0.137)	(277.281)	(0.058)
Escolaridade	0.179	0.369	-	0.355	-	0.241*	0.388*	0.468*	-32911579845.642*	0.257
	(0.216)	(0.447)	-	(0.412)	-	(0.259)	(0.442)	(0.501)	(605.021)	(0.272)
Escolaridade^2	-0.005*	-0.011**	-	-0.011**	-	-0.007*	-0.011**	-0.014*	1423171598.458*	-0.007**
	(0.008)	(0.016)	-	(0.015)	-	(0.009)	(0.015)	(0.017)	(22.393)	(0.009)
Idade	0.363***	0.669***	-	0.568***	-	0.593*	0.997***	1.175***	-99823723969.001***	0.666***
	(0.671)	(1.385)	-	(1.285)	-	(0.763)	(1.304)	(1.494)	(2519.79)	(0.771)
Idade^2	-0.023**	-0.038**	-	-0.026**	-	-0.047	-0.08**	-0.104**	16045325545.653**	-0.056**
	(0.081)	(0.166)	-	(0.153)	-	(0.091)	(0.155)	(0.173)	(306.971)	(0.094)
Doenças	-0.236	-0.488	-	-0.454	-	-0.27	-0.477	-0.52	2.8E+08	-0.267
	(0.28)	(0.574)	-	(0.537)	-	(0.308)	(0.518)	(0.533)	(1117.242)	(0.315)
Casado	0.305	0.617	-	0.575	-	0.399	0.668	0.71	5.5E+10	0.446
	(0.248)	(0.532)	-	(0.503)	-	(0.279)	(0.473)	(0.53)	(827.159)	(0.276)
Operário	0.229	0.495	-	0.494	-	0.25*	0.424	0.286	1.1E+10	0.228
	(0.266)	(0.537)	-	(0.496)	-	(0.292)	(0.487)	(0.495)	(976.573)	(0.311)
Serviços	0.277***	0.547***	-	0.526***	-	0.331***	0.547***	0.511***	-11931042465.052***	0.38***
	(0.271)	(0.552)	-	(0.51)	-	(0.291)	(0.488)	(0.518)	(1011.579)	(0.304)
Constante	-4.489***	-8.672***	-	-8.372***	-	-4.491	-7.36***	-8.085***	230409745321.588***	-4.459**
	(2.037)	(4.206)	-	(3.895)	-	(2.412)	(4.135)	(4.847)	(6361.831)	(2.496)
Reset-pvalues	-	0.698	-	0.604	-	-	-	0.406	-	0.34
B. Score	0.064	0.065	-	0.068	-	0.163	0.1651	0.086	0.383	0.175
AUC	0.59	0.59	-	0.593	-	0.589	0.588	0.582	0.569	0.585

Tabela C.3: Estimativas para $n = 5000$ com $H_S = 15\%$ e $H_S = 25\%$, respetivamente.

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.014	-0.034	-0.532	-0.034	-0.008	-0.015	-0.033	-0.215	-0.043*	-0.01
	(0.011)	(0.026)	(0.406)	(0.025)	(0.007)	(0.01)	(0.022)	(0.14)	(0.023)	(0.007)
Homem	0.168**	0.343**	1.268**	0.331**	0.12**	0.189**	0.343**	0.519**	0.34**	0.153**
	(0.067)	(0.139)	(0.553)	(0.134)	(0.047)	(0.077)	(0.142)	(0.259)	(0.135)	(0.062)
N. filhos	0.044	0.095	0.358*	0.093	0.029	0.047	0.092	0.185*	0.032	0.035
	(0.031)	(0.064)	(0.199)	(0.061)	(0.022)	(0.036)	(0.065)	(0.106)	(0.062)	(0.029)
R. Investimentos	0.007	0.014	0.018	0.012	0.006	0.008	0.015	0.025	0.012	0.008
	(0.005)	(0.373)	(1.804)	(0.358)	(0.004)	(0.005)	(0.373)	(0.879)	(0.386)	(0.005)
G. Capital	0.003	0.005	0.013	0.004	0.003	0.003	0.005	0.008	0.003	0.004
	(0.004)	(0.007)	(0.014)	(0.006)	(0.004)	(0.004)	(0.008)	(0.012)	(0.007)	(0.005)
Escolaridade	-0.093*	-0.183*	-0.51*	-0.172*	-0.07*	-0.105*	-0.186*	-0.269*	-0.271***	-0.09*
	(0.056)	(0.109)	(0.282)	(0.103)	(0.042)	(0.061)	(0.108)	(0.161)	(0.092)	(0.052)
Escolaridade^2	0.004**	0.007*	0.021*	0.007*	0.003	0.004**	0.008**	0.011*	0.01***	0.004**
	(0.002)	(0.004)	(0.011)	(0.004)	(0.002)	(0.002)	(0.004)	(0.006)	(0.003)	(0.002)
Idade	0.654***	1.418***	7.083***	1.368***	0.448***	0.748***	1.45***	3.394***	0.943**	0.573***
	(0.213)	(0.461)	(2.412)	(0.444)	(0.148)	(0.242)	(0.47)	(1.144)	(0.425)	(0.187)
Idade^2	-0.057**	-0.123**	-0.646**	-0.119**	-0.039**	-0.065**	-0.127**	-0.309**	-0.076	-0.05**
	(0.026)	(0.054)	(0.26)	(0.052)	(0.018)	(0.029)	(0.055)	(0.123)	(0.049)	(0.023)
Doenças	-0.037	-0.08	-0.22	-0.076	-0.024	-0.041	-0.081	-0.133	-0.017	-0.029
	(0.087)	(0.178)	(0.567)	(0.17)	(0.063)	(0.096)	(0.176)	(0.288)	(0.163)	(0.078)
Casado	0.075	0.162	0.624	0.156	0.051	0.084	0.162	0.328	0.227	0.064
	(0.082)	(0.174)	(0.772)	(0.168)	(0.057)	(0.089)	(0.168)	(0.349)	(0.161)	(0.07)
Operário	0.121	0.254	1.016*	0.24	0.086	0.138	0.261	0.518*	0.325**	0.109
	(0.083)	(0.171)	(0.576)	(0.164)	(0.06)	(0.092)	(0.169)	(0.29)	(0.158)	(0.075)
Serviços	0.269***	0.544***	1.598***	0.515***	0.196***	0.305***	0.556***	0.879***	0.64***	0.251***
	(0.084)	(0.17)	(0.57)	(0.162)	(0.061)	(0.091)	(0.166)	(0.284)	(0.154)	(0.074)
Constante	-2.992***	-5.964***	-22.562***	-5.916***	-1.982***	-2.761***	-5.154***	-10.548***	-3.487***	-1.905***
	(0.556)	(1.162)	(5.682)	(1.109)	(0.402)	(0.632)	(1.202)	(2.794)	(1.047)	(0.507)
Reset-pvalues	0.761	0.661	-	0.996	0.787	0.097	0.092	0.062	0.213	0.133
B. Score	0.06	0.06	0.06	0.06	0.06	0.066	0.066	0.066	0.065	0.066
AUC	0.63	0.631	0.632	0.631	0.63	0.63	0.63	0.631	0.636	0.629

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.012	-0.029	-0.29	-0.028	-0.007	-0.041***	-0.028	-0.09	-0.033	-0.009
	(0.011)	(0.026)	(0.233)	(0.026)	(0.008)	(0.011)	(0.022)	(0.067)	(0.022)	(0.009)
Homem	0.18**	0.368**	1.163**	0.353**	0.13***	0.288***	0.373**	0.415**	0.347***	0.193**
	(0.07)	(0.148)	(0.545)	(0.142)	(0.05)	(0.086)	(0.151)	(0.197)	(0.133)	(0.075)
N. filhos	0.038	0.086	0.495**	0.084	0.025	0.012	0.076	0.124	0.056	0.031
	(0.033)	(0.069)	(0.218)	(0.066)	(0.024)	(0.04)	(0.07)	(0.085)	(0.061)	(0.036)
R. Investimentos	0.008	0.016	0.06	0.015	0.006	0.007	0.016	0.023	0.014	0.009
	(0.005)	(0.399)	(1.717)	(0.382)	(0.004)	(0.006)	(0.397)	(0.614)	(0.361)	(0.006)
G. Capital	0.004	0.006	-0.005	0.004	0.004	-0.006	0.006	0.012	0.005	0.005
	(0.004)	(0.008)	(0.021)	(0.007)	(0.004)	(0.005)	(0.009)	(0.015)	(0.008)	(0.006)
Escolaridade	-0.102*	-0.207*	-0.752**	-0.198*	-0.074	-0.098	-0.205*	-0.232*	-0.195**	-0.11*
	(0.061)	(0.121)	(0.308)	(0.113)	(0.046)	(0.069)	(0.118)	(0.132)	(0.096)	(0.065)
Escolaridade^2	0.004**	0.008**	0.03**	0.008**	0.003	0.003	0.008**	0.009*	0.008**	0.004**
	(0.002)	(0.004)	(0.012)	(0.004)	(0.002)	(0.003)	(0.004)	(0.005)	(0.004)	(0.002)
Idade	0.698***	1.514***	7.779***	1.469***	0.477***	0.497*	1.564***	2.415***	1.191***	0.715***
	(0.225)	(0.488)	(2.594)	(0.472)	(0.156)	(0.269)	(0.492)	(0.819)	(0.433)	(0.225)
Idade^2	-0.062**	-0.134**	-0.707**	-0.131**	-0.042**	-0.041	-0.14**	-0.221**	-0.103**	-0.063**
	(0.027)	(0.058)	(0.279)	(0.056)	(0.019)	(0.032)	(0.057)	(0.089)	(0.05)	(0.027)
Doenças	-0.029	-0.061	-0.09	-0.059	-0.019	0.082	-0.064	-0.056	-0.049	-0.026
	(0.093)	(0.19)	(0.556)	(0.182)	(0.067)	(0.108)	(0.186)	(0.224)	(0.163)	(0.096)
Casado	0.083	0.179	0.796	0.173	0.056	0.073	0.178	0.251	0.177	0.081
	(0.087)	(0.186)	(0.827)	(0.179)	(0.06)	(0.099)	(0.177)	(0.254)	(0.158)	(0.085)
Operário	0.124	0.262	1.224**	0.25	0.087	0.174*	0.269	0.36	0.255	0.126
	(0.088)	(0.182)	(0.618)	(0.174)	(0.063)	(0.103)	(0.178)	(0.221)	(0.157)	(0.091)
Serviços	0.282***	0.572***	1.899***	0.545***	0.205***	0.267***	0.59***	0.665***	0.57***	0.306***
	(0.089)	(0.183)	(0.607)	(0.174)	(0.065)	(0.102)	(0.177)	(0.222)	(0.154)	(0.091)
Constante	-3.042***	-6.048***	-23.635***	-5.994***	-2.023***	-1.765**	-4.794***	-7.024***	-4.007***	-1.938***
	(0.594)	(1.242)	(6.079)	(1.185)	(0.428)	(0.708)	(1.271)	(2.013)	(1.09)	(0.618)
Reset-pvalues	0.754	0.666	0.545	0.0.782	0.777	0.0.167	0.160	0.134	0.231	0.236
B. Score	0.06	0.06	0.061	0.06	0.06	0.083	0.082	0.081	0.081	0.083
AUC	0.629	0.629	0.628	0.629	0.63	0.629	0.629	0.631	0.633	0.628

Tabela C.4: Estimativas para $n = 5000$ com $H_S = 35\%$ e $H_S = 50\%$, respetivamente.

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.013	-0.031	-0.324	-0.031	-0.008	-0.015	-0.03	-0.073	-0.027	-0.011
	(0.011)	(0.026)	(0.28)	(0.026)	(0.008)	(0.012)	(0.023)	(0.052)	(0.02)	(0.01)
Homem	0.18**	0.365**	0.999*	0.349**	0.131**	0.228**	0.379**	0.348**	0.317**	0.218**
	(0.074)	(0.157)	(0.577)	(0.151)	(0.053)	(0.093)	(0.159)	(0.174)	(0.132)	(0.087)
N. filhos	0.035	0.08	0.49**	0.079	0.022	0.037	0.066	0.092	0.064	0.028
	(0.035)	(0.073)	(0.23)	(0.07)	(0.025)	(0.043)	(0.073)	(0.077)	(0.06)	(0.041)
R. Investimentos	0.007	0.013	0.062	0.012	0.005	0.008	0.013	0.017	0.01	0.008
	(0.005)	(0.437)	(1.98)	(0.416)	(0.004)	(0.006)	(0.428)	(0.559)	(0.348)	(0.006)
G. Capital	0.005	0.01	0.013	0.009	0.004	0.005	0.009	0.011	0.006	0.006
	(0.006)	(0.011)	(0.023)	(0.01)	(0.004)	(0.007)	(0.012)	(0.015)	(0.008)	(0.008)
Escolaridade	-0.112*	-0.227*	-0.848**	-0.217*	-0.082	-0.135*	-0.226*	-0.228*	-0.178*	-0.133*
	(0.068)	(0.134)	(0.362)	(0.125)	(0.051)	(0.076)	(0.127)	(0.126)	(0.099)	(0.077)
Escolaridade^2	0.005**	0.009*	0.035***	0.009*	0.003	0.005*	0.009*	0.009*	0.007*	0.005*
	(0.002)	(0.005)	(0.013)	(0.005)	(0.002)	(0.003)	(0.005)	(0.005)	(0.004)	(0.003)
Idade	0.668***	1.438***	7.317***	1.392***	0.459***	0.877***	1.523***	1.786***	1.309***	0.772***
	(0.239)	(0.516)	(2.68)	(0.498)	(0.166)	(0.291)	(0.512)	(0.678)	(0.439)	(0.259)
Idade^2	-0.059**	-0.127**	-0.667**	-0.123**	-0.04**	-0.078**	-0.137**	-0.161**	-0.118**	-0.069**
	(0.029)	(0.062)	(0.287)	(0.059)	(0.02)	(0.035)	(0.06)	(0.075)	(0.051)	(0.031)
Doenças	-0.015	-0.032	0.15	-0.03	-0.009	-0.019	-0.038	-0.033	-0.033	-0.009
	(0.099)	(0.203)	(0.586)	(0.195)	(0.072)	(0.117)	(0.197)	(0.203)	(0.162)	(0.112)
Casado	0.101	0.221	1.432	0.216	0.068	0.123	0.215	0.254	0.185	0.106
	(0.091)	(0.196)	(0.968)	(0.19)	(0.063)	(0.107)	(0.184)	(0.218)	(0.157)	(0.097)
Operário	0.123	0.259	1.273*	0.248	0.085	0.152	0.265	0.295	0.213	0.133
	(0.092)	(0.192)	(0.672)	(0.184)	(0.066)	(0.111)	(0.187)	(0.198)	(0.155)	(0.105)
Serviços	0.283***	0.575***	2.108***	0.549***	0.204***	0.354***	0.596***	0.588***	0.479***	0.335***
	(0.095)	(0.194)	(0.634)	(0.185)	(0.069)	(0.11)	(0.187)	(0.201)	(0.153)	(0.104)
Constante	-2.937***	-5.805***	-22.681***	-5.758***	-1.953***	-2.47***	-4.268***	-4.94***	-4.085***	-1.821**
	(0.631)	(1.312)	(6.341)	(1.249)	(0.457)	(0.766)	(1.331)	(1.692)	(1.111)	(0.714)
Reset-pvalues	0.842	0.816	0.596	0.87	0.869	0.313	0.308	0.17	0.41	0.323
B. Score	0.06	0.06	0.061	0.06	0.06	0.104	0.104	0.102	0.103	0.104
AUC	0.63	0.63	0.628	0.63	0.629	0.629	0.629	0.632	0.629	0.628

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.01	-0.026	-0.44	-0.025	-0.006	-0.043***	-0.024	-0.048	-0.019	-0.007
	(0.012)	(0.028)	(0.278)	(0.027)	(0.008)	(0.013)	(0.024)	(0.038)	(0.02)	(0.012)
Homem	0.189**	0.379**	1.133*	0.361**	0.139**	0.301***	0.407**	0.352**	0.322**	0.259**
	(0.08)	(0.168)	(0.635)	(0.162)	(0.056)	(0.102)	(0.169)	(0.164)	(0.133)	(0.101)
N. filhos	0.036	0.084	0.664***	0.084	0.023	-0.007	0.063	0.069	0.06	0.03
	(0.038)	(0.08)	(0.255)	(0.076)	(0.027)	(0.047)	(0.078)	(0.074)	(0.06)	(0.048)
R. Investimentos	0.007	0.014	0.073	0.013	0.005	0.007	0.013	0.019	0.009	0.009
	(0.006)	(0.458)	(2.15)	(0.439)	(0.005)	(0.007)	(0.449)	(0.491)	(0.344)	(0.008)
G. Capital	0.013*	0.024*	0.065	0.022**	0.01*	0.001	0.017	0.017	0.013	0.011
	(0.007)	(0.014)	(0.043)	(0.011)	(0.006)	(0.01)	(0.017)	(0.02)	(0.011)	(0.012)
Escolaridade	-0.115	-0.239	-1.108***	-0.23	-0.084	-0.145*	-0.233*	-0.215*	-0.172*	-0.146
	(0.075)	(0.15)	(0.341)	(0.141)	(0.056)	(0.084)	(0.137)	(0.125)	(0.101)	(0.09)
Escolaridade^2	0.005*	0.01**	0.045***	0.01**	0.003	0.005*	0.01**	0.009*	0.007*	0.006**
	(0.003)	(0.005)	(0.013)	(0.005)	(0.002)	(0.003)	(0.005)	(0.005)	(0.004)	(0.003)
Idade	0.703***	1.515***	7.847***	1.466***	0.479***	0.539*	1.614***	1.589***	1.374***	0.905***
	(0.256)	(0.557)	(2.938)	(0.537)	(0.177)	(0.316)	(0.541)	(0.609)	(0.445)	(0.3)
Idade^2	-0.062**	-0.134**	-0.71**	-0.13**	-0.042*	-0.044	-0.146**	-0.144**	-0.126**	-0.081**
	(0.031)	(0.067)	(0.318)	(0.064)	(0.022)	(0.038)	(0.064)	(0.069)	(0.052)	(0.036)
Doenças	-0.039	-0.081	0.03	-0.077	-0.026	0.09	-0.088	-0.07	-0.079	-0.044
	(0.105)	(0.217)	(0.623)	(0.208)	(0.075)	(0.127)	(0.209)	(0.195)	(0.162)	(0.128)
Casado	0.079	0.167	0.815	0.16	0.056	0.067	0.177	0.168	0.128	0.101
	(0.096)	(0.207)	(0.951)	(0.2)	(0.067)	(0.116)	(0.194)	(0.194)	(0.157)	(0.113)
Operário	0.133	0.283	1.193*	0.272	0.092	0.208*	0.284	0.293	0.217	0.151
	(0.102)	(0.211)	(0.723)	(0.203)	(0.073)	(0.122)	(0.201)	(0.189)	(0.157)	(0.123)
Serviços	0.286***	0.584***	1.903***	0.558***	0.207***	0.357***	0.608***	0.551***	0.453***	0.374***
	(0.102)	(0.209)	(0.711)	(0.2)	(0.073)	(0.12)	(0.199)	(0.194)	(0.154)	(0.121)
Constante	-3.012***	-5.946***	-22.189***	-5.878***	-2***	-1.163	-4.115***	-4.053***	-3.916***	-1.917**
	(0.687)	(1.435)	(6.904)	(1.368)	(0.497)	(0.838)	(1.421)	(1.546)	(1.131)	(0.832)
Reset-pvalues	0.914	0.93	0.546	0.921	0.923	0.52	0.466	0.286	0.475	0.434
B. Score	0.06	0.06	0.063	0.06	0.06	0.14	0.14	0.138	0.14	0.141
AUC	0.627	0.628	0.631	0.628	0.626	0.629	0.628	0.631	0.627	0.626

Tabela C.5: Estimativas para $n = 10000$ com $H_S = 15\%$ e $H_S = 25\%$, respetivamente.

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.021**	-0.051**	-0.718*	-0.051**	-0.013**	-0.023***	-0.051***	-0.289**	-0.049***	-0.015***
	(0.009)	(0.021)	(0.367)	(0.021)	(0.006)	(0.008)	(0.018)	(0.13)	(0.017)	(0.005)
Homem	0.184***	0.382***	1.444***	0.366***	0.133***	0.209***	0.384***	0.641***	0.354***	0.168***
	(0.047)	(0.098)	(0.405)	(0.095)	(0.033)	(0.054)	(0.1)	(0.184)	(0.091)	(0.043)
N. filhos	0.044**	0.094**	0.368**	0.091**	0.031*	0.049*	0.093**	0.172**	0.086**	0.038*
	(0.022)	(0.045)	(0.151)	(0.043)	(0.016)	(0.025)	(0.046)	(0.075)	(0.041)	(0.02)
R. Investimentos	0.008***	0.015	0.041	0.013	0.006***	0.009***	0.015	0.023	0.013	0.008***
	(0.003)	(0.275)	(1.557)	(0.266)	(0.002)	(0.003)	(0.275)	(0.678)	(0.258)	(0.003)
G. Capital	0.004	0.007	0.014*	0.006	0.003	0.004	0.007	0.007	0.005	0.004
	(0.003)	(0.005)	(0.008)	(0.004)	(0.002)	(0.003)	(0.005)	(0.006)	(0.004)	(0.003)
Escolaridade	-0.104***	-0.2***	-0.505***	-0.187***	-0.082***	-0.12***	-0.206***	-0.243**	-0.18***	-0.109***
	(0.037)	(0.072)	(0.175)	(0.068)	(0.029)	(0.042)	(0.074)	(0.098)	(0.064)	(0.038)
Escolaridade^2	0.004***	0.008***	0.02***	0.007**	0.003***	0.005**	0.008***	0.01**	0.007***	0.004***
	(0.001)	(0.003)	(0.007)	(0.003)	(0.001)	(0.002)	(0.003)	(0.004)	(0.002)	(0.001)
Idade	0.576***	1.268***	6.634***	1.231***	0.393***	0.666***	1.296***	2.889***	1.193***	0.5***
	(0.147)	(0.317)	(1.682)	(0.307)	(0.104)	(0.168)	(0.324)	(0.766)	(0.3)	(0.13)
Idade^2	-0.048***	-0.106***	-0.572***	-0.103***	-0.032**	-0.056***	-0.11***	-0.249***	-0.1***	-0.041**
	(0.018)	(0.038)	(0.18)	(0.036)	(0.013)	(0.02)	(0.038)	(0.082)	(0.035)	(0.016)
Doenças	-0.007	-0.021	-0.155	-0.021	-0.001	-0.006	-0.02	-0.051	-0.016	0.002
	(0.06)	(0.124)	(0.433)	(0.119)	(0.044)	(0.067)	(0.123)	(0.2)	(0.111)	(0.055)
Casado	0.089	0.188	0.656	0.182	0.062	0.101	0.189	0.304	0.183*	0.078
	(0.055)	(0.12)	(0.568)	(0.116)	(0.039)	(0.062)	(0.117)	(0.24)	(0.109)	(0.049)
Operário	0.111*	0.226*	0.505	0.215*	0.08*	0.127*	0.234**	0.319	0.21*	0.102*
	(0.059)	(0.122)	(0.436)	(0.117)	(0.043)	(0.065)	(0.119)	(0.2)	(0.108)	(0.053)
Serviços	0.286***	0.576***	1.45***	0.548***	0.21***	0.328***	0.591***	0.763***	0.532***	0.271***
	(0.057)	(0.117)	(0.406)	(0.112)	(0.042)	(0.063)	(0.116)	(0.198)	(0.105)	(0.052)
Constante	-2.713***	-5.466***	-21.726***	-5.459***	-1.738***	-2.42***	-4.582***	-9.333***	-4.527***	-1.572***
	(0.376)	(0.783)	(3.877)	(0.748)	(0.277)	(0.437)	(0.821)	(1.839)	(0.745)	(0.358)
Reset-pvalues	0.86	0.838	0.5	0.893	0.8	0.113	0.131	0.118	0.141	0.166
B. Score	0.06	0.06	0.06	0.06	0.06	0.066	0.066	0.066	0.066	0.066
AUC	0.636	0.637	0.638	0.637	0.636	0.636	0.636	0.638	0.637	0.635

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.02**	-0.05**	-0.691*	-0.049**	-0.012**	-0.023***	-0.049***	-0.169**	-0.046***	-0.016***
	(0.009)	(0.021)	(0.366)	(0.021)	(0.006)	(0.009)	(0.018)	(0.076)	(0.017)	(0.006)
Homem	0.174***	0.361***	1.291***	0.346***	0.125***	0.211***	0.366***	0.425***	0.319***	0.184***
	(0.049)	(0.104)	(0.434)	(0.1)	(0.035)	(0.06)	(0.105)	(0.136)	(0.092)	(0.053)
N. filhos	0.047**	0.103**	0.42**	0.1**	0.033*	0.055*	0.098**	0.139**	0.09**	0.045*
	(0.023)	(0.047)	(0.165)	(0.046)	(0.016)	(0.028)	(0.048)	(0.057)	(0.041)	(0.025)
R. Investimentos	0.007***	0.013	0.035	0.012	0.005***	0.008***	0.014	0.017	0.01	0.008***
	(0.003)	(0.287)	(1.641)	(0.278)	(0.002)	(0.004)	(0.287)	(0.465)	(0.257)	(0.004)
G. Capital	0.003	0.005	0.011*	0.005	0.002	0.003	0.005	0.005	0.004	0.003
	(0.003)	(0.005)	(0.009)	(0.005)	(0.002)	(0.003)	(0.006)	(0.006)	(0.004)	(0.004)
Escolaridade	-0.1***	-0.192***	-0.497***	-0.181***	-0.078***	-0.124***	-0.202***	-0.189**	-0.165***	-0.124***
	(0.039)	(0.077)	(0.209)	(0.072)	(0.03)	(0.047)	(0.079)	(0.084)	(0.064)	(0.046)
Escolaridade^2	0.004***	0.007***	0.02***	0.007**	0.003***	0.005**	0.008***	0.007**	0.006***	0.005***
	(0.001)	(0.003)	(0.008)	(0.003)	(0.001)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
Idade	0.562***	1.23***	6.305***	1.194***	0.385***	0.71***	1.292***	1.883***	1.14***	0.58***
	(0.155)	(0.335)	(1.787)	(0.324)	(0.109)	(0.186)	(0.338)	(0.545)	(0.301)	(0.157)
Idade^2	-0.046***	-0.102***	-0.536***	-0.099***	-0.032**	-0.06***	-0.11***	-0.161***	-0.096***	-0.048**
	(0.019)	(0.04)	(0.192)	(0.039)	(0.013)	(0.022)	(0.04)	(0.059)	(0.035)	(0.019)
Doenças	-0.004	-0.014	-0.072	-0.014	0	-0.003	-0.013	-0.027	-0.013	0.006
	(0.064)	(0.131)	(0.469)	(0.126)	(0.047)	(0.075)	(0.129)	(0.153)	(0.111)	(0.067)
Casado	0.088	0.187	0.729	0.18	0.061	0.106	0.188	0.222	0.17*	0.088
	(0.058)	(0.125)	(0.611)	(0.122)	(0.04)	(0.069)	(0.122)	(0.172)	(0.109)	(0.059)
Operário	0.109*	0.221*	0.452	0.21*	0.079*	0.136*	0.237**	0.25	0.201*	0.119*
	(0.062)	(0.129)	(0.478)	(0.124)	(0.045)	(0.072)	(0.125)	(0.151)	(0.109)	(0.064)
Serviços	0.274***	0.552***	1.478***	0.526***	0.201***	0.338***	0.58***	0.583***	0.488***	0.304***
	(0.06)	(0.124)	(0.439)	(0.119)	(0.044)	(0.071)	(0.122)	(0.151)	(0.105)	(0.063)
Constante	-2.7***	-5.41***	-21.057***	-5.396***	-1.738***	-2.236***	-4.054***	-5.839***	-3.958***	-1.469***
	(0.395)	(0.825)	(4.173)	(0.789)	(0.289)	(0.485)	(0.864)	(1.325)	(0.747)	(0.433)
Reset-pvalues	0.91	0.903	0.624	0.939	0.86	0.1	0.114	0.111	0.136	0.1153
B. Score	0.06	0.06	0.06	0.06	0.06	0.083	0.083	0.082	0.083	0.083
AUC	0.635	0.636	0.637	0.636	0.635	0.635	0.635	0.637	0.635	0.634

Tabela C.6: Estimativas para $n = 10000$ com $H_S = 35\%$ e $H_S = 50\%$, respetivamente.

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.021**	-0.051**	-0.72*	-0.05**	-0.013**	-0.025***	-0.049***	-0.121**	-0.056***	-0.018***
	(0.009)	(0.021)	(0.377)	(0.021)	(0.006)	(0.009)	(0.018)	(0.055)	(0.017)	(0.007)
Homem	0.17***	0.35***	1.165***	0.335***	0.123***	0.214***	0.362***	0.352***	0.298***	0.198***
	(0.052)	(0.109)	(0.449)	(0.106)	(0.037)	(0.065)	(0.11)	(0.119)	(0.092)	(0.06)
N. filhos	0.051**	0.112**	0.506**	0.109**	0.035*	0.059*	0.103**	0.125**	0.011**	0.052*
	(0.024)	(0.05)	(0.176)	(0.048)	(0.017)	(0.03)	(0.051)	(0.052)	(0.042)	(0.029)
R. Investimentos	0.007***	0.013	0.041	0.013	0.005***	0.008***	0.014	0.014	0.015	0.009***
	(0.003)	(0.301)	(1.741)	(0.292)	(0.003)	(0.004)	(0.3)	(0.394)	(0.258)	(0.004)
G. Capital	0.003	0.005	0.013*	0.005	0.002	0.003	0.005	0.005	-0.007	0.003
	(0.003)	(0.006)	(0.011)	(0.005)	(0.003)	(0.004)	(0.006)	(0.006)	(0.004)	(0.004)
Escolaridade	-0.103***	-0.198***	-0.55***	-0.188***	-0.081***	-0.132***	-0.213***	-0.174**	-0.187***	-0.143***
	(0.042)	(0.083)	(0.228)	(0.077)	(0.033)	(0.052)	(0.086)	(0.081)	(0.065)	(0.054)
Escolaridade^2	0.004***	0.008***	0.022***	0.007**	0.003***	0.005**	0.008***	0.007**	0.007***	0.005***
	(0.002)	(0.003)	(0.009)	(0.003)	(0.001)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
Idade	0.547***	1.189***	5.629***	1.151***	0.377***	0.734***	1.285***	1.47***	0.784***	0.636***
	(0.163)	(0.352)	(1.829)	(0.34)	(0.114)	(0.2)	(0.352)	(0.458)	(0.302)	(0.179)
Idade^2	-0.044***	-0.097***	-0.452***	-0.094***	-0.03**	-0.062***	-0.109***	-0.123***	-0.063***	-0.053**
	(0.02)	(0.042)	(0.197)	(0.041)	(0.014)	(0.024)	(0.041)	(0.051)	(0.035)	(0.022)
Doenças	0.001	-0.005	-0.099	-0.006	0.004	0.005	-0.001	-0.018	-0.052	0.016
	(0.067)	(0.139)	(0.493)	(0.134)	(0.049)	(0.081)	(0.136)	(0.137)	(0.112)	(0.078)
Casado	0.082	0.174	0.7	0.168	0.057	0.102	0.175	0.176	0.133*	0.091
	(0.061)	(0.132)	(0.631)	(0.128)	(0.042)	(0.074)	(0.127)	(0.146)	(0.109)	(0.067)
Operário	0.121*	0.247*	0.602	0.236*	0.087*	0.155*	0.264**	0.265	0.232*	0.142*
	(0.065)	(0.136)	(0.496)	(0.13)	(0.047)	(0.078)	(0.131)	(0.135)	(0.109)	(0.074)
Serviços	0.282***	0.572***	1.507***	0.545***	0.207***	0.361***	0.606***	0.558***	0.39***	0.343***
	(0.063)	(0.13)	(0.465)	(0.125)	(0.046)	(0.076)	(0.129)	(0.136)	(0.106)	(0.072)
Constante	-2.655***	-5.303***	-19.652***	-5.283***	-1.709***	-2.063***	-3.649***	-4.369***	-2.482***	-1.341***
	(0.419)	(0.873)	(4.31)	(0.833)	(0.308)	(0.528)	(0.912)	(1.132)	(0.755)	(0.504)
Reset-pvalues	0.952	0.953	0.781	0.967	0.912	0.166	0.181	0.131	0.183	0.106
B. Score	0.06	0.06	0.06	0.06	0.06	0.104	0.104	0.104	0.101	0.104
AUC	0.636	0.636	0.639	0.636	0.635	0.635	0.635	0.637	0.633	0.634

	Com correção					Sem correção				
Variáveis	Probit	Logit	Cauchit	Cloglog	Loglog	Probit	Logit	Cauchit	Cloglog	Loglog
S. Desempregado	-0.021**	-0.052**	-0.744*	-0.051**	-0.013**	-0.025***	-0.048***	-0.093**	-0.044***	-0.019***
	(0.009)	(0.023)	(0.394)	(0.022)	(0.006)	(0.01)	(0.019)	(0.041)	(0.017)	(0.008)
Homem	0.182***	0.376***	1.341***	0.36***	0.131***	0.235***	0.391***	0.354***	0.298***	0.23***
	(0.056)	(0.118)	(0.493)	(0.114)	(0.04)	(0.071)	(0.118)	(0.113)	(0.093)	(0.071)
N. filhos	0.047**	0.104**	0.475**	0.102**	0.032*	0.055*	0.093**	0.098**	0.077**	0.05*
	(0.026)	(0.054)	(0.188)	(0.052)	(0.018)	(0.033)	(0.054)	(0.05)	(0.042)	(0.033)
R. Investimentos	0.007***	0.014	0.05	0.014	0.005***	0.008***	0.014	0.014	0.01	0.009***
	(0.004)	(0.322)	(1.78)	(0.312)	(0.003)	(0.005)	(0.319)	(0.344)	(0.256)	(0.005)
G. Capital	0.003	0.006	0.007*	0.005	0.003	0.003	0.005	0.005	0.003	0.004
	(0.004)	(0.007)	(0.016)	(0.006)	(0.003)	(0.004)	(0.007)	(0.007)	(0.004)	(0.005)
Escolaridade	-0.112***	-0.215***	-0.6***	-0.205***	-0.088***	-0.147***	-0.236***	-0.185**	-0.174***	-0.172***
	(0.047)	(0.091)	(0.231)	(0.083)	(0.036)	(0.058)	(0.095)	(0.086)	(0.067)	(0.065)
Escolaridade^2	0.004***	0.008***	0.024***	0.008**	0.003***	0.006**	0.009***	0.007**	0.007***	0.006***
	(0.002)	(0.003)	(0.009)	(0.003)	(0.001)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
Idade	0.59***	1.281***	5.684***	1.24***	0.407***	0.83***	1.411***	1.374***	1.171***	0.767***
	(0.172)	(0.374)	(1.911)	(0.361)	(0.121)	(0.216)	(0.37)	(0.411)	(0.303)	(0.206)
Idade^2	-0.049***	-0.108***	-0.466***	-0.104***	-0.034**	-0.072***	-0.123***	-0.119***	-0.103***	-0.066**
	(0.021)	(0.045)	(0.206)	(0.043)	(0.015)	(0.026)	(0.044)	(0.046)	(0.035)	(0.025)
Doenças	-0.005	-0.016	-0.085	-0.016	0	-0.001	-0.011	-0.024	-0.021	0.015
	(0.072)	(0.15)	(0.529)	(0.144)	(0.052)	(0.088)	(0.145)	(0.132)	(0.113)	(0.091)
Casado	0.086	0.184	0.886	0.179	0.059	0.11	0.185	0.169	0.144*	0.104
	(0.065)	(0.14)	(0.678)	(0.136)	(0.045)	(0.08)	(0.134)	(0.132)	(0.109)	(0.078)
Operário	0.135*	0.281*	0.804	0.269*	0.096*	0.176*	0.295**	0.277	0.227*	0.169*
	(0.07)	(0.147)	(0.53)	(0.141)	(0.051)	(0.085)	(0.14)	(0.128)	(0.11)	(0.087)
Serviços	0.303***	0.619***	1.616***	0.59***	0.22***	0.397***	0.66***	0.592***	0.498***	0.399***
	(0.069)	(0.142)	(0.52)	(0.136)	(0.05)	(0.084)	(0.138)	(0.132)	(0.107)	(0.085)
Constante	-2.696***	-5.407***	-19.579***	-5.384***	-1.731***	-1.983***	-3.421***	-3.559***	-3.265***	-1.28***
	(0.452)	(0.938)	(4.508)	(0.892)	(0.334)	(0.578)	(0.979)	(1.047)	(0.762)	(0.594)
Reset-pvalues	0.964	0.962	0.817	0.971	0.929	0.217	0.245	0.187	0.218	0.133
B. Score	0.06	0.06	0.06	0.06	0.06	0.139	0.139	0.14	0.139	0.139
AUC	0.636	0.636	0.64	0.636	0.635	0.634	0.634	0.636	0.634	0.633

C.3 Resultados dos Modelos Flexíveis

	Modelos Potência								
	5000			10000			Toda a amostra		
Variáveis	Probit	Logit	Cauchit	Probit	Logit	Cauchit	Probit	Logit	Cauchit
S. Desempregado	-0.015*	-0.016	-0.035*	-0.017***	-0.025***	-0.018***	-0.022***	-0.023***	-0.015***
	(0.009)	(0.01)	(0.02)	(0.006)	(0.008)	(0.006)	(0.002)	(0.002)	(0)
Homem	0.166**	0.18**	0.182***	0.13***	0.188***	0.168***	0.178***	0.186***	0.145***
	(0.065)	(0.07)	(0.067)	(0.036)	(0.051)	(0.042)	(0.013)	(0.014)	(0.002)
N. filhos	0.048	0.057*	0.04	0.036**	0.049**	0.031	0.033***	0.034***	0.024***
	(0.03)	(0.032)	(0.03)	(0.017)	(0.023)	(0.02)	(0.006)	(0.006)	(0.001)
R. Investimentos	0.003	0.004	0.003	0.002	0.003	0.004	0.001	0.002**	0.001***
	(0.003)	(0.004)	(0.005)	(0.002)	(0.003)	(0.003)	(0.001)	(0.001)	(0)
G. Capital	0.007*	0.009	0.009	0.007***	0.009	0.008	0.004***	0.005	0.004
	(0.004)	(0.179)	(0.16)	(0.002)	(0.135)	(0.108)	(0.001)	(0.037)	(0.005)
Escolaridade	-0.09*	-0.092	-0.102*	-0.079***	-0.112***	-0.113***	-0.097***	-0.103***	-0.096***
	(0.051)	(0.056)	(0.054)	(0.028)	(0.039)	(0.037)	(0.01)	(0.011)	(0.002)
Escolaridade^2	0.004**	0.004**	0.004**	0.003***	0.004***	0.004***	0.004***	0.004***	0.004***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0)	(0)	(0)
Idade	0.624***	0.73***	0.625***	0.497***	0.687***	0.526***	0.439***	0.456***	0.329***
	(0.207)	(0.229)	(0.216)	(0.113)	(0.161)	(0.13)	(0.041)	(0.043)	(0.006)
Idade^2	-0.053**	-0.063**	-0.054**	-0.042***	-0.058***	-0.045***	-0.034***	-0.036***	-0.025***
	(0.024)	(0.027)	(0.025)	(0.013)	(0.019)	(0.016)	(0.005)	(0.005)	(0.001)
Doenças	-0.033	-0.041	-0.022	-0.02	-0.02	0.004	-0.02	-0.021	-0.013***
	(0.081)	(0.089)	(0.081)	(0.045)	(0.063)	(0.053)	(0.017)	(0.017)	(0.003)
Casado	0.069	0.068	0.064	0.05	0.084	0.079*	0.128***	0.133***	0.101***
	(0.076)	(0.082)	(0.075)	(0.041)	(0.059)	(0.047)	(0.015)	(0.016)	(0.002)
Operário	0.126	0.131	0.112	0.081*	0.12**	0.093*	0.08***	0.084***	0.062***
	(0.078)	(0.085)	(0.078)	(0.043)	(0.061)	(0.051)	(0.016)	(0.017)	(0.003)
Serviços	0.269***	0.285***	0.262***	0.216***	0.31***	0.271***	0.22***	0.231***	0.187***
	(0.077)	(0.084)	(0.079)	(0.042)	(0.059)	(0.05)	(0.016)	(0.016)	(0.003)
Constante	-2.944***	-1.88***	-1.024*	-1.594***	-1.77***	-0.435	-2.392***	-1.103***	0.001
	(0.536)	(0.597)	(0.574)	(0.291)	(0.416)	(0.359)	(0.106)	(0.111)	(0.02)
ζ	1***	3.738***	5.055***	2.705***	3.159***	5.444***	1***	3.41***	6.34***
	(0.021)	(0.077)	(0.104)	(0.039)	(0.046)	(0.079)	(0.004)	(0.014)	(0.026)
Reset-pvalues	-	-	-	-	-	-	0.61	-	-
B. Score	0.06	0.060	0.06	0.06	0.063	0.06	0.06	0.06	0.06
AUC	0.631	0.629	0.634	0.635	0.634	0.635	0.639	0.64	0.64

	Aranda-ordaz			Weibull			Weibull Reflexa		
Variáveis	N=5000	N=10000	N=121930	N=5000	N=10000	N=121930	N=5000	N=10000	N=121930
S. Desempregado	-0.057*	-0.081***	-0.076***	-0.003	-0.004**	-0.005***	-0.003	-0.004***	-0.003***
	(0.031)	(0.027)	(0.007)	(0.002)	(0.002)	(0)	(0.002)	(0.001)	(0)
Homem	0.799***	0.982***	0.838***	0.043**	0.042***	0.049***	0.038**	0.037***	0.03***
	(0.275)	(0.224)	(0.059)	(0.017)	(0.011)	(0.003)	(0.015)	(0.01)	(0.002)
N. filhos	0.103	0.113	0.127***	0.012	0.01**	0.008***	0.011	0.01**	0.005***
	(0.128)	(0.105)	(0.028)	(0.008)	(0.005)	(0.002)	(0.007)	(0.005)	(0.001)
R. Investimentos	0.017	0.03	0.01**	0.001	0.001	0***	0.001	0.001	0***
	(0.024)	(0.024)	(0.005)	(0.001)	(0.001)	(0)	(0.001)	(0.001)	(0)
G. Capital	0.03	0.034	0.026***	0.002**	0.002	0.001	0.002**	0.002	0.001
	(0.653)	(0.546)	(0.005)	(0.001)	(0.03)	(0.009)	(0.001)	(0.026)	(0.006)
Escolaridade	-0.617**	-0.709***	-0.637***	-0.025*	-0.024***	-0.029***	-0.019	-0.023***	-0.017***
	(0.27)	(0.228)	(0.06)	(0.014)	(0.009)	(0.003)	(0.012)	(0.008)	(0.002)
Escolaridade^2	0.023**	0.027***	0.023***	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.01)	(0.008)	(0.002)	(0)	(0)	(0)	(0)	(0)	(0)
Idade	2.515***	2.661***	1.782***	0.154***	0.144***	0.111***	0.148***	0.134***	0.072***
	(0.843)	(0.67)	(0.178)	(0.051)	(0.035)	(0.011)	(0.047)	(0.031)	(0.007)
Idade^2	-0.219**	-0.229***	-0.137***	-0.013**	-0.012***	-0.009***	-0.013**	-0.011***	-0.006***
	(0.103)	(0.083)	(0.022)	(0.006)	(0.004)	(0.001)	(0.006)	(0.004)	(0.001)
Doenças	-0.037	0.185	-0.057	-0.007	-0.003	-0.005	-0.008	-0.005	-0.003
	(0.355)	(0.3)	(0.078)	(0.021)	(0.014)	(0.005)	(0.019)	(0.013)	(0.003)
Casado	0.33	0.483**	0.574***	0.017	0.019	0.034***	0.015	0.015	0.021***
	(0.303)	(0.241)	(0.064)	(0.019)	(0.013)	(0.004)	(0.017)	(0.012)	(0.003)
Operário	0.368	0.463*	0.346***	0.031	0.029**	0.021***	0.029	0.023*	0.013***
	(0.335)	(0.277)	(0.073)	(0.02)	(0.014)	(0.004)	(0.018)	(0.012)	(0.003)
Serviços	1.264***	1.566***	1.117***	0.071***	0.071***	0.061***	0.061***	0.062***	0.038***
	(0.338)	(0.278)	(0.073)	(0.02)	(0.013)	(0.004)	(0.018)	(0.012)	(0.003)
Constante	-5.398**	-5.106**	-3.592***	0.132	0.23**	0.217***	-1.632***	-1.543***	-1.353***
	(2.478)	(2.025)	(0.532)	(0.137)	(0.09)	(0.028)	(0.123)	(0.081)	(0.018)
ζ	23.975***	31.475***	28.65***	3.636***	4.32***	3.087***	4.21***	4.382***	5.146***
	(2.021)	(0.566)	(0.566)	(0.079)	(0.079)	(0.066)	(0.084)	(0.084)	(0.062)
Reset-pvalues	0.192	0.01	0	0.217	0.204	0	-	-	0
B.Score	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
AUC	0.634	0.632	0.641	0.631	0.634	0.639	0.629	0.634	0.64

C.4 Resultados dos AMPEs

	Modelos *standard* com toda a amostra					Modelos Potência			Weibull	W. Reflexo	A-O
	probit	logit	cauchit	cloglog	loglog	probit	logit	cauchit			
S. Desempregado	-0.00253	-0.00302	-0.01277	-0.00309	-0.00217	-0.00253	-0.00255	-0.00216	-0.00196	-0.00234	-0.00189
R. Investimentos	0.000512	0.00047	0.000204	0.000446	0.000557	0.000512	0.000525	0.000605	0.000544	0.000534	0.000637
G. Capital	0.000171	0.000156	9.75E-05	0.000149	0.000189	0.000171	0.000176	0.000211	0.000185	0.00018	0.000238
Doenças	-0.00237	-0.0026	-0.0028	-0.00264	-0.0021	-0.00237	-0.00235	-0.00189	-0.002	-0.00224	-0.0014
Homem	0.02083	0.020595	0.017623	0.02052	0.02091	0.02083	0.02085	0.020934	0.020848	0.020889	0.02073
Casado	0.014889	0.014997	0.015112	0.015045	0.014727	0.01489	0.01485	0.014517	0.014745	0.014816	0.014196
N. filhos	0.003849	0.004043	0.004138	0.004077	0.003634	0.003849	0.003836	0.003491	0.003587	0.003745	0.003137
Operário	0.009385	0.009597	0.009943	0.009656	0.009117	0.009385	0.009362	0.008949	0.009024	0.009259	0.00857
Serviços	0.025712	0.02485	0.017794	0.024607	0.026361	0.025712	0.025822	0.026977	0.026218	0.026057	0.027642

	Amostras com seleção: $n = 10000$ e $H_S = 0.5$									
	Corrigido					Não corrigido				
	probit	logit	cauchit	cloglog	loglog	probit	logit	cauchit	cloglog	loglog
S. Desempregado	-0.00242	-0.00293	-0.01016	-0.003	-0.0021	-0.00836	-0.00977	-0.01966	-0.01122	-0.00624
R. Investimentos	0.000821	0.000804	0.000682	0.0008	0.000883	0.002819	0.002826	0.002849	0.002654	0.003078
G. Capital	0.000368	0.000323	0.000102	0.00031	0.000428	0.001079	0.001072	0.001041	0.000657	0.001307
Doenças	-0.00052	-0.00091	-0.00116	-0.00096	3.54E-05	-0.00037	-0.00219	-0.00509	-0.00536	0.004927
Homem	0.021187	0.021203	0.018305	0.021148	0.02175	0.078723	0.079023	0.074548	0.075492	0.077373
Casado	0.009978	0.010354	0.012089	0.010479	0.009857	0.036877	0.037441	0.035676	0.036628	0.03479
N. filhos	0.005495	0.005863	0.006482	0.005958	0.005367	0.018326	0.018708	0.020686	0.019661	0.016847
Operário	0.015655	0.015826	0.010972	0.01578	0.01589	0.058905	0.059713	0.05834	0.057723	0.056752
Serviços	0.035225	0.034852	0.022051	0.034631	0.0366	0.133078	0.133406	0.12456	0.126278	0.133998