Revista da Estatística UFOP, Vol VI, 2017, ISSN 2237-8111

MODELAGEM DE EVENTOS RAROS: UMA APLICAÇÃO UTILIZANDO REGRESSÃO *PROBIT*

Patrick Franco Alves[1], Alan Ricardo da Silva[2]

Resumo: Em avaliações utilizando bases de dados observacionais existem situações onde variáveis dicotômicas possuem distribuição extremamente desbalanceada entre sucessos e fracassos. Entre tais situações se encontram as avaliações de políticas públicas, onde os beneficiados são uma pequena proporção da população. O trabalho de KING & ZENG (2001) mostra a possibilidade de obtenção de estimativas viciadas para a regressão logística quando da modelagem de eventos raros. Neste artigo o objetivo é demonstrar o efeito da modelagem de eventos raros no contexto da função de ligação *probit* e suas consequências sobre os efeitos marginais. Foi possível verificar que, sob a hipótese de auto-seleção do evento "sucesso", as probabilidades estimadas se afastam consideravelmente das probabilidades reais. Como solução para a baixa qualidade das probabilidades estimadas é proposta uma simulação computacional baseada em re-amostragem, onde são escolhidas sub-amostras aleatórias e o modelo *probit* é então estimado em cada uma delas.
Palavra-chave: Eventos Raros, Simulação Computacional, Modelo *Probit*.

Abstract: In evaluations over observational databases there are situations where dichotomous variables show extreme unbalance in the distribution between success and failure. Public policy evaluations are found in such situations, where the chosen units are just a small fraction from the entire population. The work of KING & ZENG (2001) shows biased parameters estimates when there is a rare event situation. The main issue in this paper is to demonstrate empirically the effect of rare event data in the context of a probit link function as well as some insights over the marginal effects. We reveal, under the auto-selection over the success event, the estimates probability dismisses from the actual probability. As a solution for this low quality probability estimates we propose a computational simulation where sets of random samples are chosen and the probit model is estimated for each of them.
Key-words: Rare Events, Computational Simulation, Probit Model.

1. Introdução

As funções de ligação *probit* e *logit* encontram-se entre as mais utilizadas dentro da classe de modelos lineares generalizados. Um fato aceito em trabalhos aplicados é que tais funções de ligação produzem resultados similares, exceto quando da presença de valores extremos no vetor de variáveis explicativas (CHAMBERS & COX, 1967). GREENE (2008) aponta que as funções de ligação *probit* e *logit* produzem resultados distintos em amostras tendo poucos "sucessos" (y_i=1) em relação ao número de "fracassos" (y_i=0), ou poucos "fracassos" em relação ao número de "sucessos".

No contexto dos modelos probabilísticos multivariados, estudos realizados por HAHN & SOYER (2005) demonstram que há dispersão de resultados devido à escolha de função de ligação *probit* ou *logit*. A disparidade dos resultados torna-se mais evidente com o aumento do tamanho da amostra. A disponibilidade de uma boa

[1]Programa de Pós-Graduação Departamento de Economia - Universidade de Brasília - patrickfrancoalves@yahoo.com.br

[2]Departamento de Estatística - Universidade de Brasília - alansilva@unb.br

estimativa das probabilidades individuais de sucesso é particularmente importante na aplicação da técnica *Propensity Score Matching* (PSM). Esta metodologia é utilizada em avaliações de políticas públicas por DEHEJIA & WAHBA (2002).

O algoritmo de PSM proposto por PARSONS (2004) pressupõe o ajuste de um modelo probabilístico anterior ao pareamento das observações dos grupos de controle e tratamento. No entanto a existência de viés de seleção e um desbalanceamento entre o percentual de sucessos e fracassos, pode implicar em probabilidades preditas com baixa qualidade e num baixo desempenho do algoritmo de PSM.

2. Modelos Probabilísticos em Eventos Raros

2.1 Modelo Logístico:

No modelo *logit* cada uma das ocorrências de y_i $(i = 1,\ldots n)$ possui distribuição de Bernoulli com probabilidade de sucesso π_i (y_i=1). Os valores de π_i possuem variabilidade explicada por $\mathbf{X}_i$.

$$\pi_i = \{1 + \exp(-\mathbf{X}_i \boldsymbol{\beta})\}^{-1} \quad \textbf{(1)}$$

$$P(Y_i = 1 \mid \pi_i) = \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \quad \textbf{(2)}$$

As estimativas dos parâmetros $\boldsymbol{\beta}$, indicadas por $\boldsymbol{\beta}$, são obtidos através da maximização da função de verossimilhança:

$$\ln(\boldsymbol{\beta} \mid \mathbf{X}) = \sum_{i=1}^{m} \ln(\pi_i) + \sum_{i=m+1}^{n} \ln(1-\pi_i) \quad \textbf{(3)}$$

onde existem m observações para y_i=1 e n = m+1 observações para y_i=0. As estimativas $\hat{\boldsymbol{\beta}}$ possuem matriz de variância-covariância dada por:

$$V(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} [\pi_i (1-\pi_i) \mathbf{X}_i' \mathbf{X}_i]^{-1} \quad \textbf{(4)}$$

KING & ZENG (2001) apontam que, se o modelo *logit* possui razoável poder de explicação, as probabilidades estimadas $\hat{\pi}_i$ serão relativamente próximas de 0,5 para $y_i = 1$ e mais próximas de zero para $y_i = 0$. A quantidade $\pi_i(1-\pi_i)$ será maior entre os eventos raros, consequentemente a quantidade $[\pi_i(1-\pi_i)\mathbf{x}_i'\mathbf{x}_i]^{-1}$ será menor quando $y_i = 1$. Tal característica indica que a inclusão de mais sucessos na amostra é mais informativa do que a inclusão de mais fracassos.

Quando o evento $y_i = 1$ é raro pode-se aplicar o método de estratificação endógena, o qual consiste em coletar as informações na população onde $y_i = 1$ e selecionar aleatoriamente informações onde $y_i = 0$ (KING & ZENG, 2001). Este método é complementado utilizando as frações populacionais de sucesso e fracasso.

Existe ainda o método denominado *case-cohort.* Este consiste na seleção aleatória de uma grande quantidade de observações, analisando-se todas as observações onde $y_i = 1$ e uma sub-amostra aleatória de $y_i = 0$. Quando existe um número razoável de observações para $y_i = 1$ e $y_i = 0$, então a amostra balanceada em 50% de sucessos e fracassos é uma solução razoável em diversas situações. Com exceção do intercepto, os demais parâmetros obtidos para o vetor $\hat{\boldsymbol{\beta}}$ são consistentes. Uma correção do intercepto é obtida através da expressão seguinte (Na epidemiologia e na estatística, tal correção tem

sido atribuída a PRENTICE & PYKE (1976, *apud* KING & ZENG, 2001), enquanto na econometria tem sido atribuída a MANSKI & LERMAN (1977)):

$$\hat{\beta}_0^* = \hat{\beta}_0 - \ln\left[\frac{1-\tau}{\tau} \times \frac{\overline{y}}{1-\overline{y}}\right] \quad \textbf{(5)}$$

onde, τ é a proporção populacional de sucessos e $\overline{y}$ é a proporção amostral. Em (5) $\hat{\beta}_0^* = \hat{\beta}_0$ se e somente se $\tau = \overline{y}$. A expressão acima é valida nos casos em que há uma única extração aleatória de $y_i = 1$ e $y_i = 0$. A correção (5) é relevante já que o interesse da análise se concentra também nas probabilidades preditas: $\hat{\pi}_i = \{1 + \exp(-\mathbf{x}_i \hat{\boldsymbol{\beta}})\}^{-1}$, tendo fundamental importância na aplicação da técnica PSM (PARSONS, 2004), onde as probabilidades preditas alimentam um algoritmo de casamento de observações.

Um procedimento conhecido como estimador de máxima verossimilhança com pesos exógenos (*weighted exogenous sampling maximum-likelihood estimator*) consiste em maximizar a função:

$$\ln_w(\boldsymbol{\beta} \mid \mathbf{X}) = w_1 \sum_{i=1}^{m} \ln(\pi_i) + w_0 \sum_{i=m+1}^{n} \ln(1-\pi_i) \quad \textbf{(6)}$$

onde: $w_1 = \tau / \overline{y}$ e $w_0 = (1-\tau)/(1-\overline{y})$. Entretanto, a aplicação dos pesos w_1 e w_0 pode apresentar desempenho ruim em amostras pequenas.

2.2. Modelo *Probit*:

A modelagem de eventos raros utilizando regressão *probit* (ou seja, aquela que considera a distribuição normal) não foi considerada nos estudos de KING & ZENG (2001). Para introdução desta situação considere uma variável latente ($-\infty < y_i^* < +\infty$) tendo distribuição normal ($y_i^* \sim N(\mu_y, \sigma^2)$) e sendo determinada pelo vetor de características observadas ($\mathbf{X}_i$). A variável latente (y_i^*) determina cada uma das ocorrências binárias segundo a relação (7) e (8):

$$y_i^* = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i \quad \textbf{(7)}$$

$$\begin{cases} y_i^* \le k \;\Rightarrow\; y_i = 0 \\ y_i^* > k \;\Rightarrow\; y_i = 1 \end{cases} \quad \textbf{(8)}$$

onde *k* é o valor limite para a variável latente. A formulação da regressão *probit* implica ainda nos pressupostos $E(\varepsilon_i \mid \mathbf{x}) = 0$ e $Var(\varepsilon_i \mid \mathbf{X}) = 1$. A função de ligação é dada por:

$$P(y_i = 1) = P(y_i^* > 0) = P(\mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i > 0) = P\left(\frac{\varepsilon_i}{\sigma} > -\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right) = \Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right) \quad \textbf{(9)}$$

Por sua vez, $P(y_i = 0) = 1 - \Phi(\mathbf{X}_i' \boldsymbol{\beta} / \sigma)$.

As estimativas dos parâmetros ($\hat{\boldsymbol{\beta}}$) são obtidas através da maximização de:

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} y_i \ln[\Phi(\mathbf{X}_i' \boldsymbol{\beta} / \sigma)] + (1 - y_i) \ln[1 - \Phi(\mathbf{X}_i' \boldsymbol{\beta} / \sigma)] \\ &= \sum_{\substack{i=1 \\ y_i = 1}}^{n-m} \ln[\Phi(\mathbf{X}_i' \boldsymbol{\beta} / \sigma)] + \sum_{\substack{i=n-m+1 \\ y_i = 0}}^{n} \ln[1 - \Phi(\mathbf{X}_i' \boldsymbol{\beta} / \sigma)] \end{aligned} \quad \textbf{(10)}$$

Na modelagem de eventos raros o termo $\sum \ln[1 - \Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma)]$ possui maior influência na função de verossimilhança que o termo $\sum \ln[\Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma)]$.

Consequentemente, as estimativas $\hat{\boldsymbol{\beta}}$ refletirão principalmente as características do vetor $\mathbf{X}_i$ quando $y_i = 0$.

A existência de um evento raro também apresenta consequências sobre os efeitos marginais do modelo *probit*. O efeito marginal é definido para cada observação:

$$\frac{\partial E_i(y_i)}{\partial x_{ik}} = \phi\,(\mathbf{X}_i'\boldsymbol{\beta})\,\beta_k \qquad \textbf{(11)}$$

Usualmente, os efeitos marginais individuais são agregados e interpretados através de uma média $\sum \phi\,(\mathbf{x}_i'\boldsymbol{\beta})\,\beta_k / n$. Em eventos raros as probabilidades individuais preditas são muito pequenas e os efeitos marginais médios serão governados principalmente pelas observações onde $y_i = 0$.

O comportamento da regressão *probit* em eventos raros pode também ser entendido segundo o comportamento da variável latente (y_i^*) e do valor de corte (*k*) que gera $y_i = [0;1]$. A baixa frequência de sucessos reflete a existência de uma baixa probabilidade $P(y_i^* > k)$. Nesta situação o valor limite (*k*) encontra-se muito distante de μ_y. Isto gera uma disparidade de resultados entre a regressão *logit* e *probit* (GREENE, 2008). O valor de corte para determinação da ocorrência $y_i = 1$ se encontra na cauda da distribuição normal, enquanto a estimativa mais verossímil encontra-se próxima da média μ_y, conforme mostra a Figura 1.

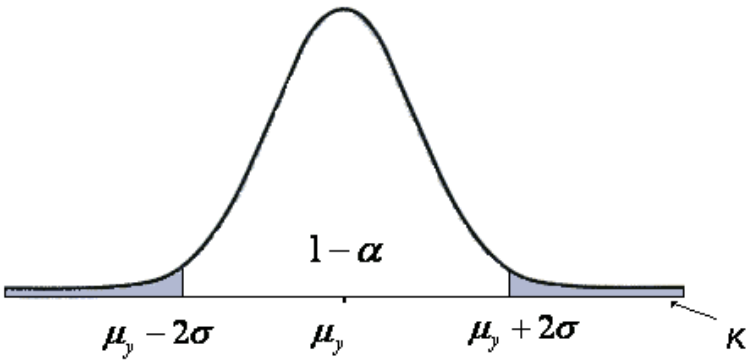

Figura 1: Distribuição da Variável Latente ($y_i^* \sim N(\mu_y, \sigma^2)$) e Valor de Corte *k*.

2.3. Escore de Propensão em Estudos Observacionais:

A mensuração de efeito do tratamento utilizando bases de dados observacionais envolve não-aleatoriedade e correlação entre a aplicação do tratamento e as características prévias das unidades amostrais. A literatura denomina tais características de viés de auto-seleção, ou seja, nem todas as unidades amostrais possuem chance de serem selecionadas. Em tais situações a metodologia do escore de propensão (RUBIN, 2006) é utilizada para mensuração de efeito de tratamento. Nesta subseção é demonstrada que a estimação de modelos probabilísticos (*probit e logit*) se relaciona intimamente com esta literatura. Na presença de auto-seleção o escore de propensão estimado não é estatisticamente igual ao escore de propensão real. As seguir demonstra-se tais condições.

Seja a *i*-ésima observação contendo as respostas W_{i1} e W_{i0}. A resposta W_{i1} ocorre quando a *i*-ésima observação é exposta ao tratamento e W_{i0} ocorre quando a *i*-ésima observação não é exposta ao tratamento. O efeito do tratamento é dado por:

$$E(\tau_i) = E(W_{i1}) - E(W_{i0}) \quad \textbf{(12)}$$

A quantidade $E(W_{i1})$ é facilmente calculada a partir de dados observacionais, mas em geral não se obtém $E(W_{i0})$, ou seja, o valor da *i*-ésima observação que não foi exposta ao tratamento.

Em experimentos aleatórios é conhecido o vetor de covariáveis $\mathbf{X}_i$ que determina a aplicação do tratamento nas unidades amostrais. Em geral se conhece também o escore de propensão $e(\mathbf{X}_i)$, ou seja, a probabilidade de cada uma das observações receberem o tratamento. Mais formalmente, em experimentos aleatórios as respostas (W_{i1}, W_{i0}) e a aplicação do tratamento (Y_i) são independentes, dado o vetor de covariáveis $\mathbf{X}_i$:

$$(W_{i1}, W_{i0}) \perp Y_i \mid \mathbf{x}_i \quad \textbf{(13)}$$

Um resultado conhecido é a invalidade do pressuposto (13) para experimentos não controlados. Outra distinção adicional entre experimentos controlados e dados observacionais é que, em experimentos controlados existe uma probabilidade não-nula e conhecida de todas as unidades amostrais serem exportas ao tratamento ($0 < p(Y_i = 1 \mid \mathbf{x}_i) < 1$).

RUBIN (2006) mostra que a aplicação do tratamento entre as unidades observacionais pode ser considerada aleatória dada uma função de balanceamento: $P(Y_i = 1 \mid b(\mathbf{x}_i))$. Por sua vez, um escore de balanceamento $b(\mathbf{x}_i)$ é uma função das covariáveis observadas $\mathbf{X}_i$ tal que a distribuição condicional: $f(\mathbf{x}_i \mid b(\mathbf{x}_i))$ é a mesma para $Y = [0,1]$, ou seja: $\mathbf{x}_i \perp Y \mid b(\mathbf{x}_i)$. Esta é uma forma de induzir a existência do pressuposto (13) em dados observacionais. Em exemplo simples de um escore de balanceamento, válido para experimentos controlados, é a igualdade: $b(\mathbf{x}_i) = \mathbf{x}_i$.

Seja a seguinte probabilidade condicional: $P(Y_i = 1 \mid \mathbf{x}_i) = e(\mathbf{x}_i)$, onde para toda a amostra temos:

$$p(Y_1 = 1, \ldots, Y_n = 1 \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} e(x_i)^{Y_i} \{1 - e(x_i)^{1-Y_i}\} \quad \textbf{(14)}$$

A função $e(\mathbf{x}_i)$ é também denominada escore de propensão, ou seja, representa a aptidão ao tratamento 1, dadas as covariáveis observáveis $\mathbf{X}_i$. Um resultado importante é que o escore de propensão funciona também como um escore de balanceamento. A solução para dados observacionais apresentada em RUBIN (2006) é a utilização do escore de propensão para indução da independência das respostas em relação à aplicação do tratamento.

$$W_{i1}, W_{i0} \perp Y_i \mid e(\mathbf{x}_i) \quad \textbf{(15)}$$

Se a aplicação do tratamento é aleatória dada um escore de balanceamento, também é aleatória dado o escore de propensão $e(\mathbf{X}_i)$ e o seguinte desenvolvimento é válido:

$$\begin{aligned} E[\tau_i \mid e(\mathbf{x}_i)] &= E[W_{i1} \mid Y_i = 1;\, e(\mathbf{x}_i)] - E[W_{i0} \mid Y_i = 0;\, e(\mathbf{x}_i)] = \\ &= E[W_{i1} \mid e(\mathbf{x}_i)] - E[W_{i0} \mid e(\mathbf{x}_i)] \end{aligned} \quad \textbf{(16)}$$

Sob as condições estabelecidas em RUBIN (2006), o estimador do efeito do tratamento proposto em (16) é consistente para a obtenção de inferências.

Usualmente se obtém $e(\mathbf{X}_i)$ através da estimação de um modelo probabilístico. Um tópico pouco abordado está relacionado à obtenção de $E[\tau_i \mid e(\mathbf{X}_i)]$ utilizando um escore de propensão estimado $\hat{p}(\mathbf{x}_i)$ ao invés de $e(\mathbf{X}_i)$. Em estudos observacionais a expressão exata $e(\mathbf{X}_i)$ é desconhecida, e o distanciamento entre o escore de propensão

real e o escore de propensão estimado invalida a aplicação desta teoria. Portanto é importante se certificar que o escore de propensão é estimado corretamente. Tais limitações ressaltam a necessidade de uma boa estimativa para $e(\mathbf{X}_i)$ através de uma modelagem probabilística. Isto assegura os pressupostos teóricos relacionados à equação (16). A próxima seção apresenta a metodologia aplicada para obtenção de um escore de propensão eficiente através da regressão *probit*.

2.4. Re-Amostragem em Eventos Raros:

A metodologia proposta baseia-se no ajuste de um modelo *probit* a partir da seleção de todas as *m* unidades amostrais pertencentes ao grupo de tratamento ($Y_i = 1$) e a seleção aleatória e sem reposição de *m* unidades do grupo de tratamento. A sub-amostra resultante será composta de 50% observações pertencentes ao grupo de controle e de 50% observações pertencentes ao grupo de tratamento. A seguir ajusta-se um modelo *probit*, armazenando, para cada uma das observações, as probabilidades preditas de sucesso $\hat{p}_{ib}(\mathbf{X}_b)$. Repete-se este procedimento *B* vezes até que todas as unidades observacionais do grupo controle sejam selecionadas ao menos uma vez. Após este procedimento, calculam-se, para cada uma das observações, as médias das probabilidades preditas.

$$b = 1,\ldots,B \quad \Rightarrow \begin{cases} P(Y_{bi} = 1) = \quad \Phi(\mathbf{X}'_b \boldsymbol{\beta}_b) \\ P(Y_{bi} = 0) = 1 - \Phi(\mathbf{X}'_b \boldsymbol{\beta}_b) \end{cases} \tag{17}$$

$$\hat{p}_{ib}(\mathbf{X}_b) = [Y_{bi} \times \Phi(\mathbf{X}'_b \hat{\boldsymbol{\beta}}_b)] + [(1 - Y_{bi})(1 - \Phi(\mathbf{X}'_b \hat{\boldsymbol{\beta}}_b))] \tag{18}$$

$$\bar{\hat{p}}_i(\mathbf{X}) = \sum_{b=1}^{B} \frac{\hat{p}_{ib}(\mathbf{X}_b)}{B} \tag{19}$$

Ao ajustar *B* modelos *probit* em sub-amostras de 50% de sucesso e 50% de fracasso, contorna-se o problema de excesso de zeros na amostra. A aleatoriedade na atribuição dos zeros, selecionadas por amostragem sem reposição, contorna a presença de viesamento do viés de seleção. A função de verossimilhança não será mais excessivamente influenciada pela grande quantidade de fracassos. Em cada uma das sub-amostras o modelo probabilístico consegue discernir sucessos dos fracassos no momento de construir as probabilidades preditas.

A metodologia proposta é relativamente simples, sendo uma de extensão do método *case-cohort* (subseção 2.1). Selecionam-se múltiplas amostras aleatórias de observações para a qual $Y_i = 0$.

O método é computacionalmente intensivo. A seleção de várias sub-amostras é necessária para que todas as unidades observacionais possuam ao menos uma probabilidade estimada. Por isto serão ajustados tantos modelos probabilísticos quanto forem necessários até que todas as unidades iniciais sejam selecionadas. Este método pode facilmente ser realizado através do procedimento SURVEYSELECT do software SAS.

3. Simulações Computacionais

3.1. Viés de Seleção:

Na literatura o conceito de auto-seleção é apresentado somente na forma de exemplos. A auto-seleção de empresas por um banco para concessão de empréstimo pode ocorrer devido à falta de auto-estima de algumas empresas, que nem realizam a tentativa de solicitação de tais empréstimos. Entretanto, para averiguação da validade da

metodologia proposta é necessária construir um esquema matemático que permita simular diferentes graus de auto-seleção.

Seja a variável y_i^*, tendo distribuição normal ($y_i^* \sim N(\mu_y, \sigma^2)$). Seja também a variável U_i^* tendo distribuição uniforme ($U_i^* \sim U(a,1)$) e considere que y_i^* determina cada uma das ocorrências binárias segundo o esquema:

$$\begin{cases} \Phi(y_i^*) \leq U_i^* & \Rightarrow \quad y_i = 0 \\ \Phi(y_i^*) > U_i^* & \Rightarrow \quad y_i = 1 \end{cases} \tag{20}$$

Se $a = 0$, então U_i^* tem distribuição $U_i^* \sim U(0,1)$ e todas as unidades observacionais possuem alguma chance de sucesso, inclusive aquelas que possuem probabilidade nula ($\Phi(y_i^*) = 0$); Se $a > 0$ nem aquelas unidades que possuem baixa propensão, não possuem chance de sucesso. As diferentes intensidades do processo de auto-seleção são efetuadas variando-se $0{,}00 < a < 0{,}99$. Por exemplo, para $a = 0{,}99$ temos um processo de auto-seleção, onde somente 1% das observações possui alguma chance de sucesso ($\Phi(y_i^*) \geq 0{,}99$).

Foi escolhida uma equação tendo duas variáveis explicativas (x_{i1} e x_{i2}), entretanto a simulação pode ser estendida para um modelo com mais variáveis. As variáveis explicativas possuem distribuição de probabilidade, respectivamente, normal e binomial: $x_{i1} \sim N(0;1)$ e $x_{i2} \sim BIN(15;0{,}5)$.

Foram escolhidos os seguintes tamanhos de amostra:

$$n = \{ 500 , 1000 , 5000 , 10000 \} \tag{21}$$

O intercepto possui valor fixo $\beta_0 = 3{,}6$ e os coeficientes angulares associados a x_{i1} e x_{i2} possuem valores também fixos, respectivamente, 2,7 e -1,17.

$$y_i^* = 3{,}6 + 2{,}7x_{i1} - 1{,}17x_{i2} \tag{22}$$

Para a equação (22) e para cada um dos tamanhos de amostras (21) foram simuladas 100 amostras contendo os seguintes processos de auto-seleção.

$$a = \{ 0 \ ; \ 0{,}01 \ ; \ 0{,}02 \ \ ;\ldots; \ \ 0{,}99 \ \} \tag{23}$$

Retomando o exemplo de concessão de empréstimo bancário, para $a = 0$ todas as firmas possuem chance de serem selecionadas pelo banco. Para $a = 0{,}99$ somente 1% das melhores firmas serão analisadas para concessão do empréstimo bancário. A simulação do processo de auto-seleção encontra-se representado pela expressão:

$$\begin{cases} \Phi(3{,}6 + 2{,}7x_{i1} - 1{,}17x_{i2}) \leq U_i^* \sim U(a,1) & \Rightarrow \quad y_i = 0 \\ \Phi(3{,}6 + 2{,}7x_{i1} - 1{,}17x_{i2}) > U_i^* \sim U(a,1) & \Rightarrow \quad y_i = 1 \end{cases} \tag{24}$$

Este procedimento torna possível avaliar o efeito da intensidade do processo de auto-seleção dado diferentes tamanhos de amostra. A probabilidade real de sucesso é conhecida sendo possível também se obter o erro quadrático médio, dado por:

$$EQM(p, a; n) = E(p_{an} - \hat{p}_{an})^2 \tag{25}$$

Espera-se que em processos de auto-seleção caracterizados por baixos valores de a, o ajuste de um único modelo probabilístico produza melhores resultados do que a simulação proposta. Para processos de auto-seleção caracterizados por altos valores de a, a metodologia proposta na Subseção 2.4 apresentará melhores resultados.

Após a simulação de diferentes intensidades de eventos raros nos diferentes tamanhos de amostra realizou-se a seleção das sub-amostras. A Figura mostra o percentual de sucessos em cada um dos tamanhos de amostra simulados, segundo as diferentes intensidades de auto-seleção. O percentual de sucessos encontra-se entre 9%

e 1%. Há uma diminuição do percentual de sucesso com o aumento da auto-seleção em todos os tamanhos de amostra simulados. A dinâmica apresentada na Figura 2 representa as características encontradas em bases de dados observacionais.

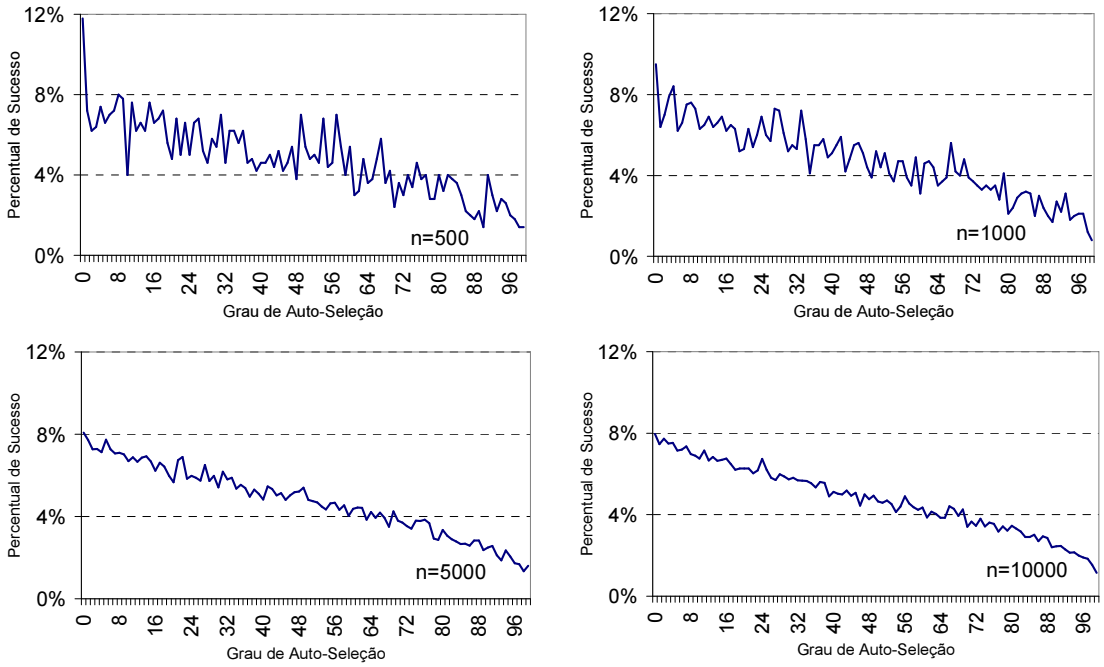

Figura 2 – Percentual de Sucesso segundo os Diferentes Tamanhos de Amostra e Graus de Auto-Seleção.

3.2. Resultados: Modelo *Probit* e Re-Amostragem

Nesta Subseção são apresentados os resultados da comparação da metodologia com o ajuste de um modelo *probit* sobre toda a população, dado os diferentes tamanhos de amostra e graus de auto-seleção.

A Figura apresenta a evolução do erro quadrático médio segundo as diferentes intensidades do processo de auto-seleção $(0 < a < 1)$. A linha vermelha pontilhada mostra a comparação do EQM da regressão *probit* ajustado para toda a população. A linha azul contínua mostra o EQM segundo a metodologia apresentada na Seção 2.4. O eixo horizontal mostra diferentes intensidades do processo de auto-seleção, enquanto o eixo vertical mostra o EQM.

Considerando a amostra de tamanho 500 (Figura), na presença de um processo de auto-seleção de baixa intensidade (<25%) o modelo *probit* se comporta melhor que o método de simulação proposto. Entretanto, para altas intensidades do processo de auto-seleção ($a > 0{,}5$), o EQM gerado pelo método de re-amostragem se comporta ligeiramente melhor. Para uma amostra de tamanho 1000 as linhas se distanciam, sendo ressaltada a vantagem da metodologia de re-amostragem em comparação a um único modelo *probit*. A partir de um processo de auto-seleção de 27% a metodologia proposta apresenta melhores probabilidades estimadas que uma regressão *probit* única.

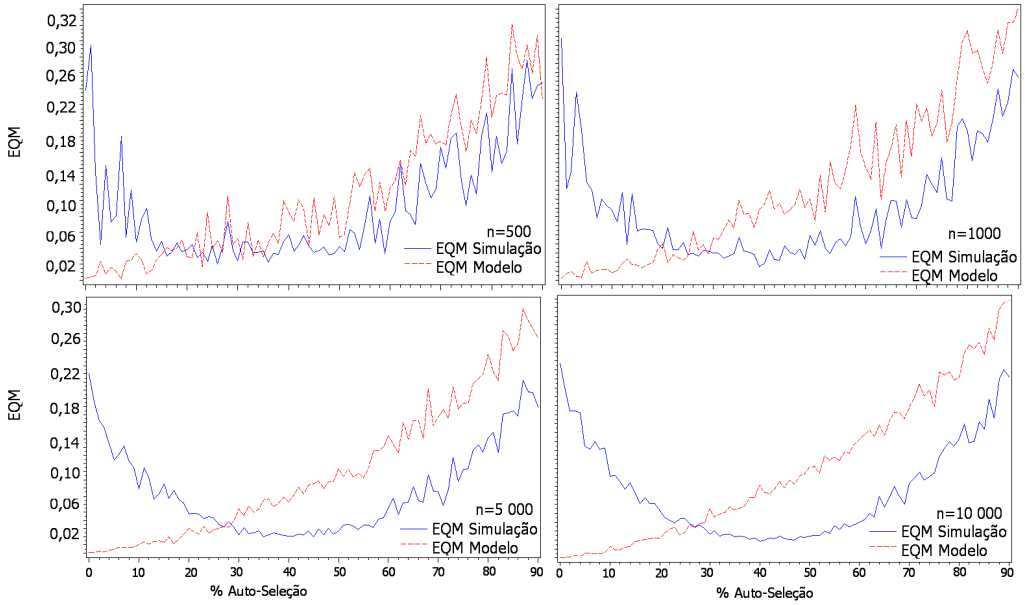

Figura 3 – Comparação do EQM segundo a Metodologia de Re-Amostragem e Modelo *Probit* Estimado para toda a população.

Para um tamanho de amostra igual a 5000 a linha azul contínua se posiciona acentuadamente abaixo da linha vermelha pontilhada quando o processo auto-seletivo é maior que 27% (Figura). Para processos auto-seletivos menores que 27% o desempenho da metodologia de re-amostragem é inferior. De fato, conforme o aumento do tamanho da amostra ocorre também o distanciamento entre as curvas de EQM associadas ao método de re amostragem e regressão *probit*. A Figura 3 indica que entre 30% e 60% de auto-seleção, o EQM do método de re-amostragem apresenta menores valores.

De particular interesse na modelagem probabilística é a predição das ocorrências binárias entre as observações, baseando-se para isto nas probabilidades preditas estimadas. Uma verificação simples pode ser realizada adotando-se a regra (26). Caso tenham sido produzidas boas estimativas para as probabilidades, se espera uma alta concentração relativa nos pontos ($y_i = 0$; $\hat{y}_i = 0$) e ($y_i = 1$; $\hat{y}_i = 1$). O Gráfico 3 mostra o percentual de acertos segundo os diferentes tamanhos de amostra.

$$\begin{cases} \Phi(\hat{\beta}_{0b} + \hat{\beta}_{1b} x_{i1} - \hat{\beta}_{2b} x_{i2}) < 0,5 \quad \Rightarrow \quad \hat{y}_i = 0 \\ \Phi(\hat{\beta}_{0b} + \hat{\beta}_{1b} x_{i1} - \hat{\beta}_{2b} x_{i2}) \geq 0,5 \quad \Rightarrow \quad \hat{y}_i = 1 \end{cases} \qquad \textbf{(26)}$$

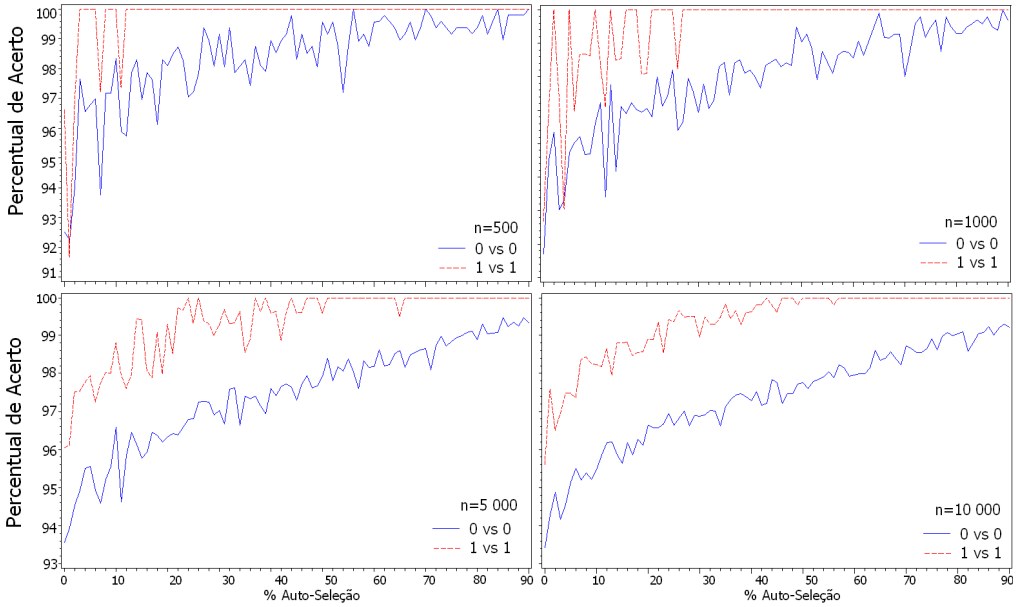

Figura 4 – Percentual de Acerto Segundo Método de Re-Amostragem.

As Figuras 4 e 5 mostram a evolução dos percentuais de acerto na classificação dos pontos $y_i = 0$; $\hat{y}_i = 0$ e $y_i = 1$; $\hat{y}_i = 1$. Encontra-se representado abaixo as células "A" e "D" indicadas na Tabela 1 ("Predição Correta").

Tabela 1: Regra de Classificação das Firmas pelo Modelo *Probit*.

Estimado / Real	$\hat{p} < 0,5 \Rightarrow \hat{y}_i = 0$	$\hat{p} \geq 0,5 \Rightarrow \hat{y}_i = 1$
$y_i = 0$	A - Predição Correta	B - Erro de Classificação
$y_i = 1$	C - Erro de Classificação	D - Predição Correta

Segundo os resultados apresentados na Figura 4 o método de re-amostragem possibilita grande percentual de acerto no ponto $y_i = 1$; $\hat{y}_i = 1$ para altos graus de auto-seleção. Para um tamanho de amostra igual a 500 e uma intensidade de auto-seleção acima de 30%, o percentual de acerto se situa acima de 90%. Para uma amostra de tamanho igual a 10000 e auto-seleção com intensidade acima 10%, o percentual de classificação correta encontra-se acima de 95%.

A Figura 5 mostra o percentual de acertos de classificação quando se realiza a estimação de um único modelo *probit* sobre toda a população. Observam-se, de forma geral, baixos percentuais no ponto $y_i = 1$; $\hat{y}_i = 1$ e altos percentuais no ponto $y_i = 0$; $\hat{y}_i = 0$. A baixa concentração no ponto $y_i = 1$; $\hat{y}_i = 1$ parece aumentar com o aumento do tamanho da amostra. Independente do tamanho da amostra, o padrão visualizado na Figura 4 apresenta uma maior concentração no ponto $y_i = 1$; $\hat{y}_i = 1$ em comparação com a Figura 5.

O comportamento visualizado na Figura 5 possui implicações relevantes para o algoritmo de Greedy de *Propensity Score Matching* (PARSONS, 2004). Este método consiste, entre outras etapas, na escolha de pares $y_i = 1$; $\hat{y}_i = 1$ para comparação de médias. O bom desempenho classificatório do modelo, segundo regra apresentada na Tabela 1 e equação (26) é fundamental para a consistência dessa metodologia.

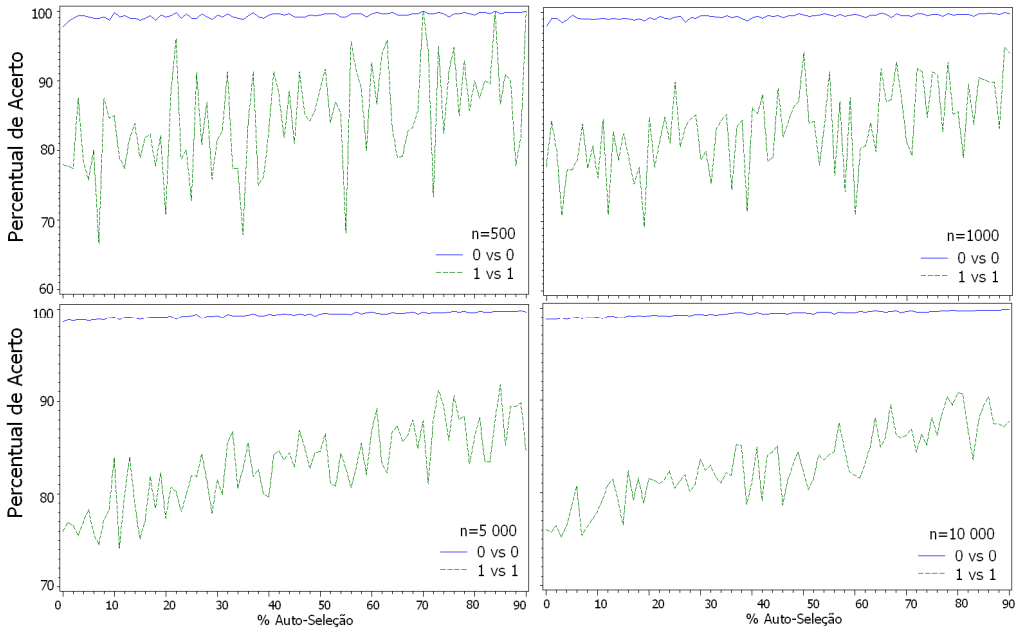

Figura 5 – Percentual de Acerto Segundo Modelo *Probit* Ajustado para toda a População.

4. Considerações Finais

Os resultados apresentados na Subseção 3.2 mostram que o erro quadrático médio do modelo *probit* ajustado para toda a população aumenta conforme aumenta a intensidade do processo auto-seletivo. Para processos auto-seletivos caracterizados por baixos valores de a ($a \leq 0{,}30$), a metodologia proposta não produz resultados melhores que o ajuste de um único modelo probabilístico (Figura). Entretanto para processos auto-seletivos mais intensos (acima de 30%) esta metodologia de re-amostragem apresenta superioridade, pois possibilita estimativas de valores preditos mais próximas das estimativas reais.

O método de re-amostragem mostrou bom desempenho na classificação das observações utilizando as probabilidades preditas (Figura 4) em relação ao modelo *probit* (Figura 5). Esta característica é particularmente relevante na aplicação técnica de *propensity score matching*, sendo indicada para trabalhos aplicados utilizando bases de dados observacionais.

Referências

[1]CHAMBERS, E. & COX, D. R. Discrinimation Betwenn Alternative Binary Response Models. Biometrika, Vol. 54, 573-578, 1967.

[2]DEHEJIA, R. & WAHBA, S. Propensity Score-Matching Methods for Nonexperimental Causal Studies. The Review of Economics and Statistics. Vol. 84(1), 151-161, 2002.

[3]GREENE, W. Econometric Analysis. Prentice Hall, 2008.

[4]HAHN, E. & SOYER, R. Probit and Logit Models: Difference in the Multivariate Realm. The Journal of the Royal Statistical Society, Series B. Forthcoming, 2005.

[5]KING, G. & ZENG, L. Logistic Regression in Rare Event Data. Political Analysis. Vol. 9(2), 137-163, 2001.

[6]MANSKI,C. F. & LERMAN,S. R. The Estimation of Choice Probabilities From Choice Based Samples. Econometrica, Vol. 45, No. 8, 1977-1988, 1977.

[7]PARSONS, S. Performing a 1:N Case-Control Match on Propensity Score. SAS Institute. 29° Annual SAS® User Group International Conference, 165-229, 2004.

[8]RUBIN, D. Matched Sampling for Casual Effects. Cambridge University Press, 2006.