Calibração da regressão em modelos de regressão beta com erro de medida

Jalmar M. F. Carrasco
Universidade de São Paulo

Silvia L. P. Ferrari
Universidade de São Paulo

Reinaldo Arellano-Valle.
Pontifícia Universidad Católica de Chile

Resumo

Este trabalho aborda uma metodologia simples de estimação, chamada de calibração da regressão, em modelos de regressão beta em que uma covariada é medida com erro. A implementação computacional da metodologia é discutida. Um estudo de simulação e uma aplicação a dados reais são apresentados.

Palavras chaves: Modelo de regressão beta, erro de medida, calibração da regressão, Monte Carlo, gás natural.

1 Modelo

Modelos de regressão beta têm se tornado uma ferramenta comum na análise de dados medidos em escala contínua e observados no intervalo unitário, (0,1). Dados dessa natureza compreendem frações que correspodem a proporções medidas de forma contínua, como, por exemplo, a proporção da renda familiar gasta com alimentação. O modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004), com a posterior extensão considerada por Espinheira (2007), pode ser definido como segue.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes e identicamente distribuídas tais que, para cada $i = 1, \ldots, n$, y_i tem distribuição beta com função densidade da forma

$$f(y_i; \mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma(1-\mu_i)\phi_i)} y_i^{\mu_i\phi_i}(1-y_i)^{(1-\mu_i)\phi_i}, \quad 0 < y_i < 1, \tag{1}$$

onde $0 < \mu_i < 1$ e $\phi_i > 0$. Pode-se mostrar que $\mathrm{E}(y_i) = \mu_i$ e $\mathrm{Var}(y_i) = \mu_i(1-\mu_i)/(1+\phi_i)$. Nota-se que, uma vez fixada a média da variável resposta, μ_i, quanto maior for o valor que ϕ_i assume, menor é a variância de y_i. Assim, ϕ_i pode ser interpretado como um parâmetro de precisão. O modelo de regressão beta considerado aqui é definido pela função densidade (1) com

$$g(\mu_i) = \boldsymbol{z}_i^\top \boldsymbol{\beta}, \tag{2}$$

$$h(\phi_i) = \boldsymbol{v}_i^\top \boldsymbol{\gamma}, \tag{3}$$

onde vetores de parâmetros desconhecidos ($\beta \in \Re^p$, $\gamma \in \Re^q$, $p + q < n$), $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2}, \cdots, \boldsymbol{z}_{ip})^\top$ e $\boldsymbol{v}_i = (\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, \cdots, \boldsymbol{v}_{iq})^\top$ são vetores de covariáveis, assumidas fixas e conhecidas, $g(\cdot)$ e $h(\cdot)$ são funções estritamente monótonas e duplamente diferenciáveis.

É comum na prática que algumas variáveis explicativas não sejam observadas diretamente, mas sejam obtidas com possíveis erros de medição. O foco desse trabalho está em inferência estatística para modelos de regressão beta em que uma covariada é medida com erro. Nesse sentido, o modelo a ser considerado é dado por (1) com as equações (2) e (3) substituídas por

$$g(\mu_i) = \boldsymbol{z}_i^\top \boldsymbol{\beta} + x_i\beta_x,$$

$$h(\phi_i) = \boldsymbol{v}_i^\top \boldsymbol{\gamma} + x_i\gamma_x,$$

em que β_x e γ_x são parâmetros desconhecidos.

Existem duas abordagens para lidar com o problema de erros de medida em modelos de regressão. A primeira é conhecida como modelo funcional e trata as covariáveis não observadas como parâmetros incidentais. Por outro lado, se as covariáveis são tratadas como variáveis aleatórias independentes, identicamente distribuídas e independentes dos erros de medição, o modelo é conhecido como um modelo estrutural. Consideraremos nesta ocasião o caso do modelo estrutural. Admitimos aqui que o erro de medida é aditivo, ou seja, definimos

$$w_i = x_i + e_i,$$

onde $x_i \sim N(\mu_x, \sigma_x^2)$, x_i e e_i são independentes e $e_i \sim N(0, \sigma_e^2)$, para $i = 1, \ldots, n$. Aqui, w_i corresponde à variável explicativa observada com erro de medição associada à i-ésima observação e x_i é o verdadeiro valor da variável explicativa, não observável, associada à i-ésima observação.

Para se obter a função densidade conjunta de (y_i, w_i), que é a observação para o i-ésimo indivíduo, integra-se a função densidade conjunta dos dados completos (y_i, w_i, x_i), denotada por $f(y_i, w_i, x_i; \boldsymbol{\theta})$ com respeito a x_i, em que $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \beta_x, \boldsymbol{\gamma}^\top, \gamma_x, \sigma_e^2, \mu_x, \sigma_x^2)^\top$ é o vetor de parâmetros desconhecidos a serem estimados, isto é,

$$\begin{aligned} f(y_i, w_i; \boldsymbol{\theta}) &= \int_{-\infty}^{\infty} f(y_i, w_i, x_i; \boldsymbol{\theta}) dx_i, \\ &= \int_{-\infty}^{\infty} f(y_i|x_i; \boldsymbol{\beta}^\top, \beta_x, \boldsymbol{\gamma}^\top, \gamma_x) f(w_i|x_i, \sigma_e^2) f(x_i; \mu_x, \sigma_x^2) dx_i, \end{aligned} \tag{4}$$

em que $f(y_i|x_i; \boldsymbol{\beta}^\top, \beta_x, \boldsymbol{\gamma}^\top, \gamma_x)$ é a função densidade da distribuição beta dada em (1), $f(w_i|x_i; \sigma_e^2)$ é a função densidade condicional de w_i dado x_i, ou seja, a função densidade da distribuição $N(x_i, \sigma_e^2)$, e $f(x_i; \mu_x, \sigma_x^2)$ é a função densidade da variável explicativa x_i, cuja distribuição é $N(\mu_x, \sigma_x^2)$. Note que, para a obtenção de (4) assume-se que o erro de medida é não diferenciável (Bolfarine e Arellano-Valle, 1998), ou seja, que a distribuição de y_i dado (w_i, x_i) depende apenas de x_i.

O logaritmo da função de verossimilhança correspondente à distribuição conjunta da amostra observada tem a forma

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log f(y_i, w_i; \boldsymbol{\theta}), \\ &= \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} f(y_i|x_i; \boldsymbol{\beta}^\top, \beta_x, \boldsymbol{\gamma}^\top, \gamma_x) f(w_i|x_i; \sigma_e^2) f(x_i; \mu_x, \sigma_x^2) dx_i. \end{aligned}$$

O cálculo do logaritmo da função de verossimilhança é bastante complexo. Nesse sentido, abordaremos a seguir uma metodologia de estimação numérica dos parâmetros para o caso do modelo de regressão beta estrutural com erro de medida.

2 Estimação dos parâmetros mediante o método de calibração da regressão

A ideia central do método de calibração da regressão consiste em substituir a variável não observável, x_i, por uma estimativa da esperança condicional de x_i dado w_i, $\mathrm{E}(x_i|w_i)$. Após a substituição, as estimativas dos parâmetros são obtidas por algum método confiável, por exemplo, máxima verossimilhança (Carroll, Ruppert e Stefanski, 1995). Considere que $\mathrm{E}(x_i|w_i) = r(w_i, \boldsymbol{\psi})$, chamada *função de calibração*, que depende de um vetor de parâmetros $\boldsymbol{\psi}$. A troca da variável não observada x_i por $r(w_i, \widehat{\boldsymbol{\psi}})$ estabelece um modelo modificado para os dados. Aqui, $\widehat{\boldsymbol{\psi}}$ é uma estimativa de $\boldsymbol{\psi}$.

Para o caso do modelo de regressão beta estrutural com erro de medida temos que a função de calibração é

$$r(w_i, \boldsymbol{\psi}) = \mu_x + k_x(w_i - \mu_x),$$

em que $\boldsymbol{\psi} = (\mu_x, \sigma_x^2, \sigma_e^2)^\top$ e $k_x = \sigma_x^2/(\sigma_x^2 + \sigma_e^2)$ é a chamada razão de confiabilidade. Como $w_i \sim N(\mu_x, \sigma_x^2 + \sigma_e^2)$, $\overline{w} = \sum_{i=1}^n w_i/n$ e $s_w^2 = \sum_{i=1}^n (w_i - \overline{w})^2/(n-1)$ são estimadores ótimos de μ_x e $\sigma_x^2 + \sigma_e^2$ respectivamente. Entretanto, não é possível estimar k_x através dos dados observados $w_1, \ldots, w_n$. Aqui vamos assumir que k_x é conhecido. Alternativamente, k_x pode ser estimado desde que seja possível observar réplicas de x, sem erro de medida.

Substituindo a função de calibração estimada na função densidade de probabilidade de y_i dado x_i, ou seja, em $f(y_i|x_i; \boldsymbol{\beta}^\top, \beta_x, \boldsymbol{\gamma}^\top, \gamma_x)$, obtemos o logaritmo da função de verossimilhança modificada $\ell^*(\boldsymbol{\theta})$, dado por

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}),$$

em que

$$\ell_i(\boldsymbol{\theta}) = \log\Gamma(\phi_i) - \log\Gamma(\mu_i\phi_i) - \log\Gamma((1-\mu_i)\phi_i) + (\mu_i\phi_i)\log(y_i) + [(1-\mu_i)\phi_i - 1]\log(1-y_i),$$

com

$$\begin{aligned} \mu_i &= g^{-1}(\boldsymbol{z}_i^\top \boldsymbol{\beta} + x_i^* \beta_x), \\ \phi_i &= h^{-1}(\boldsymbol{v}_i^\top \boldsymbol{\gamma} + x_i^* \gamma_x), \end{aligned}$$

sendo $x_i^* = \overline{w} + k_x(w_i - \overline{w})$.

Note que a log-verossimilhança modificada coincide com a do modelo de regressão beta usual, definido na Seção 1, sendo que x_i^* atua como uma variável explicativa, medida sem erro. Esse fato é relevante pois a estimação dos parâmetros por máxima verossimilhança pode ser realizada, sem dificuldade, utilizando por exemplo, o pacote `betareg` (Cribari-Neto e Zeiles, 2010) disponível na plataforma `R`. Evidentemente, os erros padrão das estimativas devem ser calculados separadamente. Sugere-se utilizar o método bootstrap.

3 Estudo de simulação

O estudo de simulação foi feito considerando que o parâmetro de precisão é o mesmo para todas as observações, isto é , $\phi_i = \phi$ e o modelo de regressão beta estrutural com erro de medida, com a seguinte estrutura

$$g(\mu_i) = \beta_0 + \beta_1 z_i + \beta_x x_i, \qquad i = 1, \ldots, n, \tag{5}$$

com $g(\cdot)$ sendo a função de ligação e z_i covariável fixa gerada a partir de uma distribuição $U(0,1)$. Definimos os seguintes valores para os parâmetros que permanecem fixos para todas as simulações $\beta_0 = 0.5$, $\beta_1 = 0.0$, $\beta_x = 1.0$, $\mu_x = 3.5$ e $\sigma_e^2 = 0,03$. Para o parâmetro de precisão são atribuídos dois valores $\phi = 50$ e $\phi = 200$. Consideramos os tamanhos de amostra $n = 25$, $n = 50$ e $n = 100$, e o número de réplicas Monte Carlo foi fixado em 10000. Geramos a variável não observada x_i com distribuição normal com média μ_x e variância σ_x^2, vários valores de σ_x^2, que levam a $k_x = 0,95$, $k_x = 0,90$, $k_x = 0,85$ e $k_x = 0,80$, foram considerados. Escolhemos a ligação logito e obtemos μ_i a partir de (5). Geramos as variáveis dependentes y_i a partir de uma distribuição beta com média μ_i. Ademais, geramos a covariável w_i a partir de sua distribuição condicional dado x_i, ou seja, de uma distribuição normal com média x_i e variância σ_e^2. Para cada amostra simulada os parâmetros do modelo foram estimados sob dois enfoques: primeiro, assume-se (erroneamente) que não há erro de medida na covariada x_i; segundo , assume-se que x_i é medida com erro.

A Tabela 1 mostra parte dos resultados obtidos, para $n = 50$ e $\phi = 50$. São reportadas estimativas (obtidas por simulação) da média, do viés e da raiz quadrada do erro quadrático médio (REQM) dos estimadores dos parâmetros do modelo.

Nota-se que conforme o valor de k_x decresce o viés do estimador de β_x aumenta quando este é estimado sem que o erro de medida seja levado em conta. Por outro lado, β_x é estimado praticamente sem viés quando a estimação é feita sob o modelo com erro de medida. Nota-se também que a REQM do estimador é maior sob o modelo incorreto, ou seja, sem considerar erro de medida. A estimação dos demais parâmetros é pouco afetada quando o erro na covariável x_i é desconsiderado.

Tabela 1: Resultados da simulação para $n = 50$ e $\phi = 50$

		Sem erro			Com erro		
	$\boldsymbol{\theta}$	Média	Viés	REQM	Média	Viés	REQM
$k = 0.95$	ϕ	53,902	3,9019	13,952	53,902	3,9019	13,952
	β_0	0,6527	0,1527	0,6038	0,4764	-0,0235	0,5902
	β_1	0,0131	0,0131	0,4705	0,0131	0,0131	0,4705
	β_x	0,9574	-0,0425	0,1682	1,0079	0,0078	0,1657
$k = 0.90$	ϕ	54,337	4,3367	15,206	54,337	4,3367	15,206
	β_0	0,8314	0,3314	0,9548	0,4768	-0,0231	0,91933
	β_1	-0,0021	-0,0022	0,4887	-0,0021	-0,0021	0,4886
	β_x	0,9111	-0,0888	0,2661	1,0123	0,0123	0,2611
$k = 0,85$	ϕ	54,220	4,2201	14,555	54,220	4,2201	14,555
	β_0	0,9922	0,4921	1,2054	0,4599	-0,0400	1,1515
	β_1	0,0155	0,0155	0,4881	0,0155	0,0155	0,48814
	β_x	0,8616	-0,1383	0,3414	1,0137	0,0137	0,32975
$k = 0,80$	ϕ	54,425	4,4245	15,069	54,425	4,4245	15,069
	β_0	1,2119	0,7119	1,4937	0,5126	0,0126	1,3689
	β_1	0,0167	0,0167	0,4755	0,0167	0,0167	0,4755
	β_x	0,7992	-0,2007	0,4219	0,9990	-0,0009	0,3901

4 Aplicação

O Instituto de Pesquisa Tecnológica (IPT) e a Companhia de Gás de São Paulo (COMGÁS) realizaram um estudo no qual determinou-se o fator de simultaneidade (FS), que representa a relação percentual entre a potência verificada com que trabalha simultaneamente um grupo de aparelhos (fogão de 6 bocas e aquecedor) e a soma das capacidades máximas de consumo desses mesmos aparelhos (C), em um conjunto de 42 edificações residenciais. Foram utilizadas duas ferramentas para a determinação do fator de simultaneidade: questionários e aparelho de registro de dados de vazão, ou *data-logger*. Os questionários foram úteis para determinar a capacidade máxima de consumo de gás natural pelos aparelhos em cada edificação e os *data-logger* permitiram determinar a potência verificada com que trabalham simultaneamente os aparelhos de cada edificação em estudo; para maiores detalhes ver Molinari (2008).

Analisamos esse conjunto de dados mediante o modelo de regressão beta estrutural com erro de medida com função de ligação logito. Para isso assumiremos que a variável resposta y_i (FS) tem uma distribuição beta com média μ_i e C é a covariável medida com erro. Consideramos que o parâmetro de precisão ϕ é o mesmo para todas as observações.

A Tabela 2 mostra as estimativas e os erros padrão (entre parênteses) dos parâmetros do modelo de regressão beta sem e com erro de medida. Observa-se nessa tabela que o erro padrão do estimador do parâmetro de precisão ϕ, considerando erro na covariável, diminui à medida que a razão de confiabilidade k_x decresce. No entanto, os erros padrão dos estimadores dos parâmetros β_0 e β_1 do modelo com erro de medida cresce à medida em que a razão de confiabilidade k_x decresce. Adicionalmente, estes erros padrão são maiores relativamente aos obtidos para o modelo que desconsideranda erro de medida.

Tabela 2: Estimativas e erros padrão (entre parênteses)

		β_0	β_1	ϕ
	sem erro	-1,3036(0,1408)	-0,3647(0,0536)	45,3909(10,027)
k_x	0,95	-1,2496(0,1513)	-0,3840(0,0559)	45,391(10,617)
	0,90	-1.1894(0.1630)	-0.4053(0.0601)	45,391(10.164)
	0,85	-1.1223(0.1789)	-0.4292(0.0657)	45,3910(9.4072)
	0,80	-1.0467(0.1967)	-0.4560(0.0715)	45,3910(8.6517)

5 Conclusões

Neste trabalho consideramos um modelo de regressão beta com dispersão variável em que uma das variáveis explicativas é medida com erro. Mostramos que o método de calibração da regressão é uma ferramenta simples para estimação dos parâmetros do modelo e que produz estimativas confiáveis. Mostramos ainda que ignorar o erro de medição pode provocar um viés considerável na estimação do parâmetro correspondente à variável medida com erro.

Extensões da metodologia para modelos mais gerais, por exemplo, permitindo que mais do que uma covariada seja medida com erro, ou que os preditores sejam não-lineares, são factíveis e estão sendo implementadas. Estamos também trabalhando na implementação de outros métodos numéricos de estimação por máxima verossimilhança, a saber, o algoritmo EM e MCEM e da metodologia Bayesiana. Adicionalmente, estamos desenvolvendo técnicas de diagnóstico.

Referências

Bolfarine, H., Arellano−Valle, R.B. (1998). Weak nondifferential measurement error models. *Statistics and Probability Letter*, **40**, 279-287.

Carroll, R.J., Ruppert, D., Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. New York: Chapman & Hill.

Cribari-Neto, F., Zeiles, A. (2010). Beta Regression in R. *Journal of Statistical Software*, **34**, 1-24.

Espinheira, P.L. (2007). *Regressão Beta*. São Paulo. Tese (Doutorado) - IME, Universidade de São Paulo.

Ferrari, S.L.P., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799-815.

Molinari, L.F.Z. (2008). *Predição de Fator de Simultaneidade através de Modelos de Regressão para Proporções Contínuas*. São Paulo, 2008. Dissertação (Mestrado) - IME, Universidade de São Paulo.