Modelos com erros nas variáveis: teoria e aplicações

Alexandre Galvão Patriota

Agradecimentos à Fapesp, Cnpq e Capes pelo apoio financeiro

Roteiro

Modelo de regressão linear simples

Modelo de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + q_i$$

Algumas suposições:

- Os erros $q_1, \ldots, q_n$ são considerados independentes e identicamente distribuídos (e.g., distribuição normal).

- As variáveis X e Y são relacionadas linearmente e são observadas sem erro de medida.

Modelo de regressão linear simples com erro nas variáveis

Seja o modelo:

$$Y_i = \beta_0 + \beta_1 x_i + q_i$$

No modelo com erro nas variáveis, não observamos a verdadeira covariável x_i. No lugar observamos X_i com a seguinte relação aditiva:

$$X_i = x_i + u_i.$$

em que u_i é uma variável aleatória que representa os erros de medição da variável x.

Quando $x_1, \ldots, x_n$ são uma variáveis aleatórias, dizemos que o modelo é estrutural. Quando $x_1, \ldots, x_n$ são parâmetros incidentais, dizemos que o modelo é funcional.

Modelo linear heterocedástico com erros nas variáveis e na equação

O modelo foi proposto por Kulathinal et al. (2002) e tem a seguinte relação:

$$y_i = \beta_0 + \beta_1 x_i + q_i$$
$$Y_i = y_i + e_i$$
$$X_i = x_i + u_i$$

em que y_i e x_i são variáveis latentes

$u_i \overset{ind}{\sim} \mathcal{N}(0, \tau_{ui})$

$e_i \overset{ind}{\sim} \mathcal{N}(0, \tau_{ei})$,

$q_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$,

$x_i \overset{iid}{\sim} \mathcal{N}(\mu_x, \sigma_x^2)$ (todas as quantidades mutuamente independentes).

Aplicações

- (Projeto WHO MONICA) Relacionar um índice que representa doença coronariana (y) com outro índice que representa os fatores de risco (x).

- (Observatório Chandra) Relacionar a luminosidade emitida (y) pelo disco de acresção de buracos negros com sua massa (x).

Nestes exemplos, é comum que uma medida de variabilidade venha acompanhada com cada observação estimada. Ou seja, $\hat{y}_i = Y_i$, com $\tau_{ei} = \widehat{\text{Var}}(Y_i - y_i)$ e $\hat{x}_i = X_i$, com $\tau_{ui} = \widehat{\text{Var}}(X_i - x_i)$ para todo $i = 1, \ldots, n$

Resultados obtidos

- Encontramos a matriz de covariâncias assintótica dos estimadores de momentos e de máxima verossimilhança para o vetor $\sqrt{n}(\hat{\beta} - \beta)$, em que $\beta = (\beta_0, \beta_1)$. A distribuição limite é normal de média zero e matriz de covariâncias Ψ (que depende da abordagem adotada);

- Conduzimos vários estudos de simulação de Monte Carlo para verificar a robustez dos estimadores contra desvios da distribuição suposta para os dados.

Modelo polinomial heterocedástico com erros nas variáveis e na equação

Generalização do modelo anterior para o caso polinomial.

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + q_i, \\
Y_i &= y_i + e_i, \\
X_i &= x_i + u_i,
\end{aligned}
$$

em que

$u_i \overset{ind}{\sim} \mathcal{N}(0, \tau_{ui})$

$e_i \overset{ind}{\sim} \mathcal{N}(0, \tau_{ei})$,

$q_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ e x_i parâmetros incidentais (todas as quantidades mutuamente independentes).

O modelo sem erro na equação ($\sigma^2 = 0$) foi proposto por Zavala et al. (2007).

Resultados obtidos

- Propusemos estimadores consistentes para os parâmetros do modelo baseado na abordagem do escore corrigido.
- Encontramos a distribuição assintótica dos estimadores.
- Conduzimos alguns estudos de simulação de Monte Carlo para verificar o comportamento da estatística do tipo Wald em relações lineares e quadráticas.

Modelo com a distribuição flexível para a covariável substituta

Neste modelo assumimos que $Y_i = y_i$, ou seja, obervamos diretamente a variável y_i.

$$\begin{aligned} Y_i &= \boldsymbol{\gamma}^\top \boldsymbol{W}_i + \boldsymbol{\beta}^\top \boldsymbol{x}_i + q_i \\ \boldsymbol{X}_i | \boldsymbol{x}_i &\sim \mathcal{G}_i \in \mathcal{C}(\boldsymbol{x}_i, g_1, g_2) \end{aligned} \quad \text{para} \quad i = 1, \ldots, n$$

em que $\boldsymbol{W}_i$ é um vetor de covariáveis medidas sem erro, $\boldsymbol{x}_i$ é o vetor de covariaveis latentes, $q_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$,

$$\mathrm{E}(g_1(\boldsymbol{X}_i)|\boldsymbol{x}_i) = \boldsymbol{x}_i \quad \text{e} \quad \mathrm{E}(g_2(\boldsymbol{X}_i)|\boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

As funções g_1 e g_2 tem relações com o primeiro e segundo momento do vetor $\boldsymbol{x}_i$.

Exemplos

(Liang et al., 2004) Se $X_i|x_i \sim$ Poisson(x_i), então

$$g_1(X_i) = X_i, \qquad g_2(X_i) = X_i^2 - X_i$$

Se $X_i|x_i \sim$ Uniforme$\{0, \ldots, x_i\}$, então

$$g_1(X_i) = 2X_i, \quad g_2(X_i) = 3X_i^2 - X_i$$

Se $X_i|x_i \sim$ Gamma(x_i, ϕ_i) ou $X_i|x_i \sim \mathcal{N}(x_i, x_i^2\phi_i)$, com ϕ_i conhecido para todo $i = 1, \ldots, n$, então

$$g_1(X_i) = X_i, \qquad g_2(X_i) = X_i^2/(\phi_i + 1)$$

Resultados obtidos

- Propusemos estimadores consistentes para os parâmetros do modelo baseado na abordagem do escore corrigido.
- Encontramos a distribuição assintótica dos estimadores.
- Conduzimos alguns estudos de simulação de Monte Carlo para verificar o comportamento da estatística do tipo Wald.

Referências bibliográficas

- Kulathinal SB, Kuulasmaa K, Gasbarra D. (2002). Estimation of an errors-in-variables regression model when the variances of the measurement error vary between the observations. *Statistics in Medicine*. **21**:1089–1101.

- Patriota AG, Bolfarine H. (2009). Measurement error models with a general class of error distribution. *Statistics (Berlin)*, v. 44, p. 119–127, 2010.

- Patriota AG, Bolfarine H, de Castro M. (2009). A heteroscedastic structural errors-in-variables model with equation error. *Statistical Methodology*, v. 6, p. 408–423, 2009.

Referências bibliográficas

- Patriota AG, Bolfarine H. (2008). A heteroscedastic polynomial regression with measurement error in both axes. *Sankhya. Series B*, v. 70, p. 267–282, 2008.

- Zavala AAZ, Bolfarine H, de Castro M. (2007). Consistent estimation and testing in heteroscedastic polynomial errors-in-variables models, *Annals of the Institute of Statistical Mathematics* **59**, 515–530.

- Liang L, Palta M, Shao J. (2004). A measurement error model with a Poisson distributed surrogate, *Statistics in Medicine*, **23**, 2527–2536.