Regresja liniowa z zakłóconymi zmiennymi objaśniającymi

Krystyna Maciąg i Czesław Stępniak

UMCS i Uniwersytet Rzeszowski

XXXV Konferencja

Statystyka Matematyczna Wisła 2009

7-11.12.2009

1 Motywacja

Dane $(x_1, y_1), ..., (x_n, y_n)$

I grupa

- Wyznaczyć regresję (II rodzaju) zmiennej y względem x

II grupa

- Wyznaczyć regresję (II rodzaju) zmiennej x względem y

Student z grupy II:

$y = \beta x + \alpha$

Po przekształceniu:

$x = \frac{1}{\beta} y - \frac{\alpha}{\beta}$

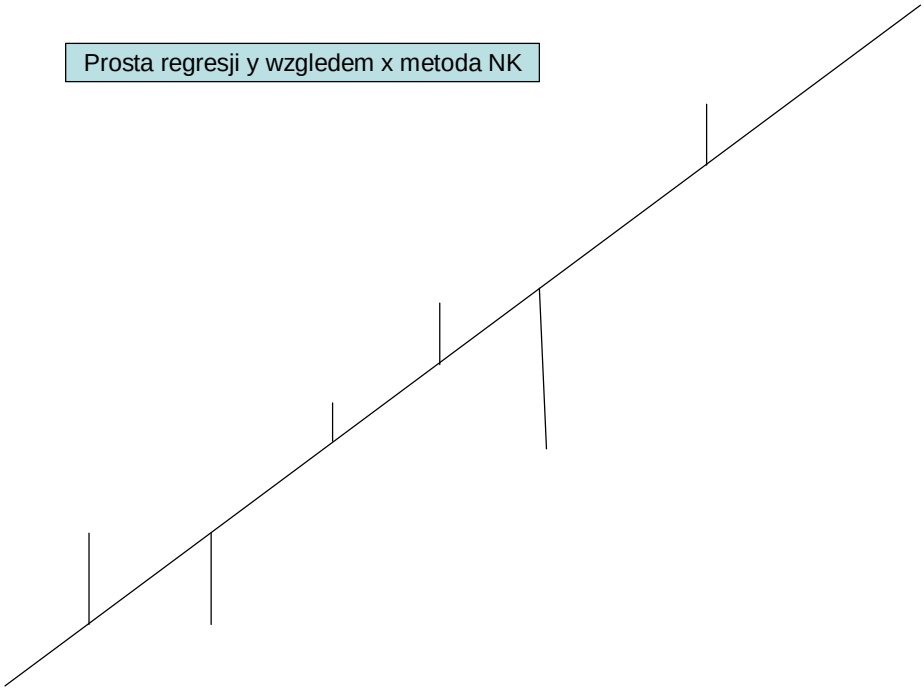
potraktował jako regresję zmiennej x względem y.

Pytanie: Kiedy takie postępowanie jest uzasadnione?

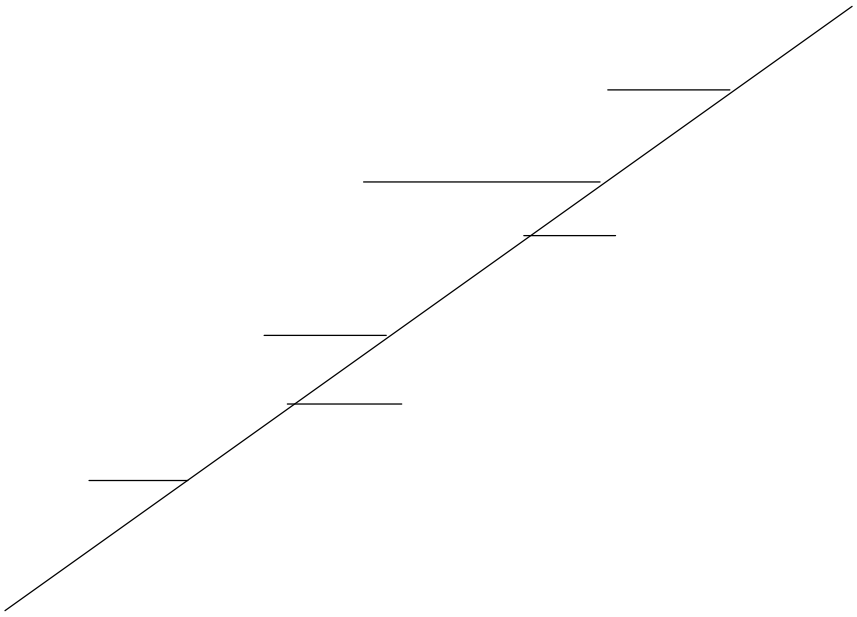
Odpowiedź: Wtedy i tylko wtedy gdy punkty

$(x_1, y_1), ..., (x_n, y_n)$ leżą na prostej.

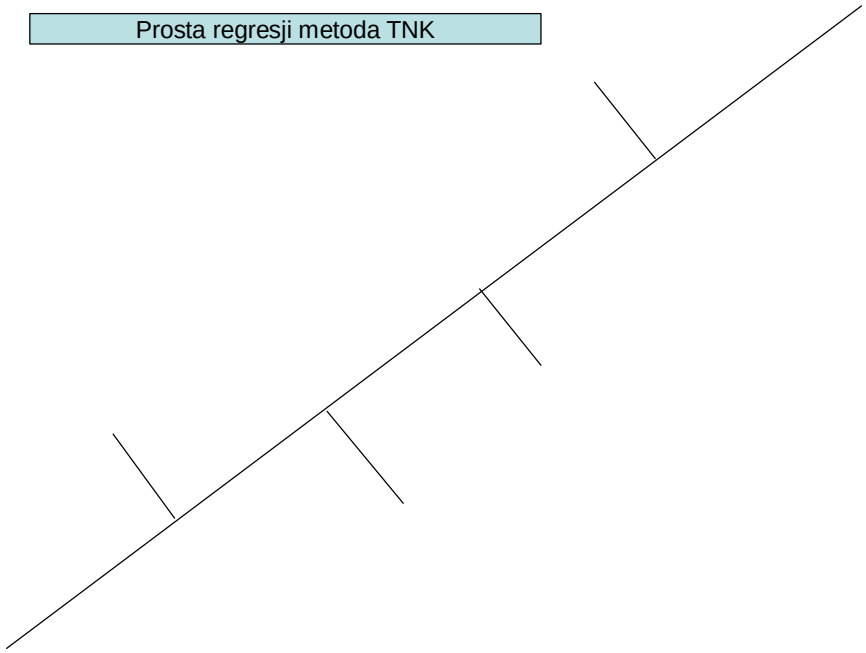
Prosta regresji y wzgledem x metoda NK



Prosta regresji x wzgledem y metoda NK



Prosta regresji metoda TNK



2 Tło

y - zmienna objaśniana (response variable)

$x_1, ..., x_k$ - zmienne objaśniające, regresyjne (explanatory variables)

$y = f(x_1, ..., x_k)$ $\qquad$ $f =$?

Wiedząc, że $f \in C$

- Wyznaczyć f w sposób eksperymentalny.

- W modelu eksperymentalnym uwzględnić błędy pomiaru wartości zmiennych.

2.1 Model I

$f \in \mathcal{L}$ (klasa funkcji liniowych)

n zestawów wartości zmiennych $x_1, ..., x_k$:

$\mathbf{x}^{(1)} = (x_1^{(1)}, ..., x_k^{(1)})$

..... (bez błędów pomiarowych)

$\mathbf{x}^{(n)} = (x_1^{(n)}, ..., x_k^{(n)})$

$\mathbf{y} = (y_1, ..., y_n)^T$

$$\mathbf{X} = \begin{bmatrix} x^{(1)} \\ . \\ . \\ .. \\ x^{(n)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ . \\ . \\ .. \\ y_n \end{bmatrix}$$

Układ równań liniowych

$$\mathbf{X}\boldsymbol{\alpha} = \mathbf{y} \tag{1}$$

- $\mathbf{y} \in R(\mathbf{X})$ i $r(\mathbf{X}) = k$

$$\Downarrow$$

układ (1) zgodny i oznaczony

- Równanie (1) prowadzi do rozwiązania

  $f(x_1, ..., x_k) = \sum_{i=1}^{k} \alpha_i \mathbf{x}_i$

  ($\mathbf{x}_i$- i-ta kolumna macierzy $\mathbf{X}$)

2.2 Model II

$f \in \mathcal{L}$

wartości $x_1, ..., x_k$ mierzone bez błędu

wartości y obarczone błędem (niesystematycznym, jednorodnym)

- Zwykle $n > k$

Na ogół układ $\mathbf{X}\boldsymbol{\alpha} = \mathbf{y}$ nie jest zgodny

Zamiast tego rozwiązujemy

$$\mathbf{X}\boldsymbol{\alpha} = \widetilde{\mathbf{y}} \tag{2}$$

gdzie $\widetilde{\mathbf{y}}$ taki aby

1° układ (2) zgodny

2° $||\widetilde{\mathbf{y}} - \mathbf{y}||_2$ minimalna

Zasada najmniejszych kwadratów

[ordinary least squares (OLS)]

Problem sprowadza się do układu

$$\mathbf{X}\boldsymbol{\alpha} = \mathbf{P}_X\mathbf{y} \tag{3}$$

$\mathbf{P}_X : R^n \mapsto R^n$

$\uparrow$

Operator rzutu ortogonalnego na $R(\mathbf{X})$

Jeśli układ (3) oznaczony to jego rozwiązanie

jest JNLN estymatorem wektora $\boldsymbol{\alpha}$.

2.3 Model III

$x_1, ..., x_k$ - mierzone z błędem (niesystem. jednor.)

y - mierzona z błędem (niesystem. jednor.)

np. x - temp. $^\circ C$

y - poziom opadów atm. $\frac{l}{m^2}$

Żeby wyznaczyć zależność y od $x_1, ..., x_k$

rozwiązujemy układ

$$\widetilde{\mathbf{X}}\boldsymbol{\alpha} = \widetilde{\mathbf{y}} \tag{4}$$

gdzie $\widetilde{\mathbf{X}}$ i $\widetilde{\mathbf{y}}$ dobieramy tak aby

1° Układ (4) zgodny

2° $\left|\left|\widetilde{\mathbf{X}}-\mathbf{X},\widetilde{\mathbf{y}}-\mathbf{y}\right|\right|_F$ minimalna

$||\mathbf{A}||_F := \sqrt{tr(\mathbf{A}^T\mathbf{A})}$ - norma Frobeniusa

Zasada totalnie najmniejszych kwadratów

[total least squares (TLS)]

Pytanie: Po co deformować macierz $\mathbf{X}$ skoro wystarczy zdeformować wektor $\mathbf{y}$ żeby otrzymać układ zgodny?

$$\min_{\widetilde{X},\widetilde{y}:\widetilde{y}\in R(\widetilde{X})} \left|\left|\widetilde{\mathbf{X}}-\mathbf{X},\widetilde{\mathbf{y}}-\mathbf{y}\right|\right|_F \leq \min_{\widetilde{y}:\widetilde{y}\in R(X)} ||\widetilde{\mathbf{y}}-\mathbf{y}||_2$$

(wystarczy położyć $\widetilde{\mathbf{X}}=\mathbf{X}$).

Uwaga: Problem TLS może nie mieć rozwiązania

Przykład

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\widetilde{\mathbf{X}}_\varepsilon = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix} : \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix} \quad \boldsymbol{\alpha} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\left\| \widetilde{\mathbf{X}}_\varepsilon - \mathbf{X}, \widetilde{\mathbf{y}} - \mathbf{y} \right\|_F = \left\| \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\|_F = \varepsilon$$

$$\min_{\widetilde{X}, \widetilde{y} : \widetilde{y} \in R(\widetilde{X})} \left\| \widetilde{\mathbf{X}} - \mathbf{X}, \widetilde{\mathbf{y}} - \mathbf{y} \right\|_F = \mathbf{0}$$

Ale

$$||\mathbf{A}, \mathbf{b}||_F = 0 \iff \mathbf{A} = \mathbf{0} \text{ i } \mathbf{b} = \mathbf{0}$$

czyli $\widetilde{\mathbf{X}} = \mathbf{X}$ i $\widetilde{\mathbf{y}} = \mathbf{y}$. Jednak

$\mathbf{X}\boldsymbol{\alpha} = \mathbf{y}$ sprzeczny.

Dotąd:

Numeryczna algebra liniowa:

Sabine Van Huffel

Jaki model statystyczny dla problemu III?

$$\mathbf{X}\boldsymbol{\alpha} = \mathbf{y}: \ \mathbf{X} \ \text{ i } \ \mathbf{y} \ \text{ nieobserwowalne} \qquad (5)$$

obserwowalne:

$$\begin{aligned} \widetilde{\mathbf{X}} &= \mathbf{X} + \mathbf{E}: \qquad \mathcal{E}(\mathbf{E}) = \mathbf{0}, \ Cox(\mathbf{D}\,1) = \sigma \mathbf{I}_{nk} \\ \widetilde{\mathbf{y}} &= \mathbf{y} + \mathbf{e}: \qquad \mathcal{E}(\mathbf{e}) = \mathbf{0}, \ \ Cov(\mathbf{e}) = \sigma \mathbf{I}_{n} \end{aligned}$$

Interesuje nas estymacja wektora parametrycznego $\boldsymbol{\alpha}$ w modelu (5).

3 Narzędzia algebraiczne

Rozkład spektralny macierzy $\mathbf{C} \in R^{m \times n}$ [spectral decomposition]

$\mathbf{U}^T \mathbf{C} \mathbf{V} = \mathbf{\Sigma} = diag(\sigma_1, ..., \sigma_p)$

$\sigma_1 \geq ... \geq \sigma_p \geq 0, \quad p = \min(m, n)$

$\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_m]$ ortogonalna

$\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_n]$ ortogonalna

σ_i - i-ta wartość spektralna, $i = 1, ..., p$

$\mathbf{v}_i$ - prawostronny wektor własny: $\mathbf{C}\mathbf{v}_i = \sigma_i \mathbf{u}_i$

$\mathbf{u}_i$ - lewostronny wektor własny: $\mathbf{C}^T \mathbf{u}_i = \sigma_i \mathbf{v}_i$

$(\sigma_i, \mathbf{u}_i, \mathbf{v}_i)$ - trójka spektralna,

zawiera całą informację o $\mathbf{C}$ (w dogodnej formie)

Dekompozycja dwójkowa [dyadic decomposition]

Wprowadzamy r :

$\sigma_1 \geq ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_p = 0$

$\mathbf{C} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$

gdzie

$$\begin{aligned} \mathbf{U}_r &= [\mathbf{u}_1, ..., \mathbf{u}_r] \\ \mathbf{\Sigma}_r &= diag(\sigma_1, ..., \sigma_r) \\ \mathbf{V}_r &= [\mathbf{v}_1, ..., \mathbf{v}_r] \end{aligned}$$

$$||\mathbf{C}||_F^2 = \sum_{i=1}^m \sum_{j=1}^n c_{ij}^2 = \sigma_1^2 + ... + \sigma_r^2$$

$$||\mathbf{C}||_2 = \sup_{\mathbf{y} \neq \mathbf{0}} \frac{||\mathbf{C}\mathbf{y}||_2}{||\mathbf{y}||_2} = \sigma_1.$$

Przeformułowanie problemu (4) TLS:

$$\mathbf{X}\boldsymbol{\alpha} \approx \mathbf{y}$$

$$\mathbf{X}\boldsymbol{\alpha} - \mathbf{y} \approx \mathbf{0}$$

$$[\mathbf{X}, \mathbf{y}][\boldsymbol{\alpha}^T, -1]^T \approx \mathbf{0}$$

Twierdzenie (Eckart-Young-Mirsky). Niech $\mathbf{C} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ będzie dekompozycją dwójkową macierzy $\mathbf{C}$ a $\mathbf{C}_k := \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ dla $k < r$. Wówczas

$$\min_{\{\mathbf{D}: r(\mathbf{D})=k\}} ||\mathbf{C} - \mathbf{D}||_2 = ||\mathbf{C} - \mathbf{C}_k||_2 = \sigma_{k+1},$$

$$\min_{\{\mathbf{D}: r(\mathbf{D})=k\}} ||\mathbf{C} - \mathbf{D}||_F = ||\mathbf{C} - \mathbf{C}_k||_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2}.$$

Eckart&Young (1936) dla $||\cdot||_F$

Mirsky (1960) dla normy $||\cdot||_2$.

4 Model regresji stochastycznej

ξ i η - zmienne losowe związane zależnością

$$\eta = \alpha + \beta\xi \text{ dla pewnego } \alpha, \beta. \tag{6}$$

Założenie: ξ i η podlegają zakłóceniom. Obserwujemy

$X = \xi + U$

$Y = \eta + V$, gdzie U i V pełnią rolę błędów losowych.

$(X_1, Y_1), ..., (X_n, Y_n)$ - próba dwuwymiarowa prosta

$$\begin{aligned} X_i &= \xi + U_i \\ Y_i &= \alpha + \beta\xi + V_i \end{aligned} \tag{7}$$

(U_i, V_i) - niezależne wektory losowe z w. ocz. zero

$$\mathbf{\Sigma}_{U,V} = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}$$

Parametry będą identyfikowalne jeśli $\sigma_{uv} = 0$ a $\frac{\sigma_u^2}{\sigma_v^2}$ znane, np. $= 1$.

Warunki (6)-(7) prowadzą do modelu

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_\xi \\ \alpha + \beta\mu_\xi \end{bmatrix}, \begin{bmatrix} \sigma_\xi^2 + \sigma_u^2 & \beta\sigma_\xi^2 \\ \beta\sigma_\xi^2 & \beta^2\sigma_\xi^2 + \sigma_u^2 \end{bmatrix} \right) \tag{8}$$

$i = 1, ..., n.$

Uwaga: Wnioskowanie w modelu (8), nie jest łatwe.

Interesuje nas β (inne parametry zakłócające)

Elementarne podejście $\longmapsto$Fuller (1987). Metoda NW.

Ortogonalna transformacja parametrów w celu eliminacji parametrów zakłócających:

Barnett (1967), Wong (1989).

Rozwiązanie: $\widehat{\beta} = \frac{s_Y^2 - s_X^2 + \sqrt{(s_X^2 - s_Y^2)^2 + 4s_{XY}}}{2s_{XY}}$, o ile $s_{XY} \neq 0$.

5 Regresja liniowa ze zmiennymi deterministycznymi przy zakłóceniach gaussowskich

x i y - deterministyczne, związane relacją

$$y = \alpha + \beta x, \text{ dla pewnych } \alpha, \beta \in R. \tag{9}$$

Obserwujemy

$$\begin{aligned} X &= x + U, \\ Y &= y + V. \end{aligned} \tag{10}$$

(U, V) - jak wyżej (niezależne, o standaryzowanych rozkładach normalnych)

$(X_1, Y_1), ..., (X_n, Y_n)$ - próba prosta z rozkładu (10), zależnego od nieobserwowalnych (x_i, y_i), $i = 1, ..., n$

Zadanie: Estymacja $(x_i, y_i), i = 1, ..., n$.

Rozwiązanie:

$$\begin{aligned}
\widehat{\beta} &= \frac{s_Y^2 - s_X^2 + \sqrt{(s_X^2 - s_Y^2)^2 + 4s_{XY}}}{2s_{XY}}, \text{ o ile } s_{XY} \neq 0. \\
\widehat{\alpha} &= \overline{Y} - \widehat{\beta}\overline{X}, \\
\widehat{x}_i &= \frac{X_i + \widehat{\beta}^2\overline{X} + \widehat{\beta}(Y_i - \overline{Y})}{1 + \widehat{\beta}^2}, \quad i = 1, ..., n. \\
\widehat{y}_i &= \widehat{\beta}\widehat{x}_i, \quad i = 1, ..., n.
\end{aligned}$$

$$\widehat{y} = \overline{Y} + \widehat{\beta}\left(\frac{X + \widehat{\beta}^2\overline{X} + \widehat{\beta}(Y - \overline{Y})}{1 + \widehat{\beta}^2} - \overline{X}\right).$$

Uwaga: Rozwiązania otrzymane dla regresji stochastycznej i deterministycznej pokrywają się prostą $y = \alpha + \beta x$ dobraną do punktów obserwacji metodą totalnie najmniejszych kwadratów (TLS).

Literatura:

Adcock, R.J. (1987). Note on the method of least squares, *Analyst* **4**, 183-184.

Adcock, R.J. (1878). A problem of least squares, *Analyst* **5**, 53-54.

Anderson, T.W. (1951). Estimating linear restrictions on regression coefficient for multivariate normal distribution, *nn. Math. Statist.* **22**, 327-351.

Barnett, V.D. (1967). A note on linear structural relationships when both residual variances are known, *Biometrika* **54**, 670-672.

Bolfarine, H., Cordani, L.K. (1993). Estimation of structural regression model with known reliability ratio, *Ann. Inst. Math. Statist.* **45**, 531-540.

Caroll, J.R., Gallo, P., Gleser, L.J. (1985). Comparison of least squares and error-in-variable regression, with special

reference to randomized analysis of covariance, *J. Amer. Statist. Assoc.* **80**, 929-932.

Eckart, G., Yung, G. (1936). The approximation of one matrix by another of lower rank, *Psychometrics* **1**, 211-218.

Fuller, M.Y. (1987). *Measurement Error Models*, J. Wiley, New York.

Golub, G.H., Van Loan, C.F. (1996). *Matrix Computations,* 3th ed. Johns Hopkins Uni. Press, baltimore, MD.

Hocking, R.R. (1983). developments in linear regression methodology 1959-1982, *Technometrics* **25**, 219=230.

Kendall, M.G., Stuart, A. (1979). *The Advanced Theory of Statistics,* Vol. 2, 4th ed. , Hafner, New York.

Ketellapper, R.H. (1983). On estimating parameters in a simple linear error-in-variable model, *Technometrics* **25**, 43-47.

Koopmans, T.C. (1937). *Linear Regression Analysis of Economic Time Series,* DeErven F. Bohr, Haarlem, The Netherlands.

Kummel, C.H. (1879). Reduction of observed equations which contain more than one observed quantity, *Analyst* **6**, 97-105.

Lindley, D.V. (1947). Regression lines and the linear functional relationship, *J. Roy. Statist. Soc. Suppl.* **9**, 218-244.

Markovsky, I., Van Huffel, S. (20070. Overview of total least-squares methods, *Signal Processing* **87**, 2283-2302.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error, *J. Amer. Statist. Assoc.* **54**, 173-205.

Mirsky, L. 91960). Symmetric gauge functions and unitarily invariant norms, *Quart. J. Math. Oxford* **11**, 50-59.

Moran, P.A.P. (1971). Estimatind structural and functional relationships, *J. Multivariate Anal.* **1**, 232-255.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philos. Mag.* **2**, 559=572.

Tintner, G. (1945). A note on rank, multicollinearity, and multiple regression, *Ann. Math. Statist.* **16**, 304-308.

Van Huffel, S., vandewalle, j. (1991). *The Total Least Squres Problem: Computational Aspects and Analysis,* SIAM, Philadelphia.

Van Huffel, S., Lammerling, P. (Eds). (2002). *Total Least squares and Error-in-variables modeling; Analysis, algorithms and Applications,* Kluwer Academic publishers, Dordrecht.

Vidal, I., Iglesias, P. (2008). Comparison between a measurement error model and a linear model wthout measurament error, *Comput. Statist. data Anal.* **55**, 92-202.

Wellman, M.J., Gunst, R.F. (1991). Inference diagnostics for linear measurement error models, *Biometrika* **78**, 373-380.

Wong, M.Y. (1989). Likelihood estimation of a simple linear regression model when both variables have error, *Biometrika* 76, 141-148.