Semiparametric Regression with Errors in Variables

Secil YALAZ TOPRAK[1], Mujgan TEZ[2], H.Ilhan TUTALAR[3]

[1] Dicle University Department of Mathematics, Turkey, secilyalaz@gmail.com
[2] Marmara University Department of Statistics, Turkey, mujgantez@gmail.com
[3] Dicle University Department of Mathematics, Turkey, tutalarhi@gmail.com

International Conference on Trends and Perspectives
in Linear Statistical Inference (LinStat2014)
24-28 August 2014
Linköping, SWEDEN

Overview

Introduction

Measurement error in predictors causes loss of information and biases and even misleading conclusions for inference.

Introduction

Measurement error in predictors causes loss of information and biases and even misleading conclusions for inference.

Three main effects of measurement error are:

Introduction

Measurement error in predictors causes loss of information and biases and even misleading conclusions for inference.

Three main effects of measurement error are:

- It causes bias in parameter estimation for statistical models.

Introduction

Measurement error in predictors causes loss of information and biases and even misleading conclusions for inference.

Three main effects of measurement error are:

- It causes bias in parameter estimation for statistical models.
- It leads to a loss of power, sometimes profound, for detecting interesting relationship among variables.

Introduction

Measurement error in predictors causes loss of information and biases and even misleading conclusions for inference.

Three main effects of measurement error are:

- It causes bias in parameter estimation for statistical models.
- It leads to a loss of power, sometimes profound, for detecting interesting relationship among variables.
- It masks the features of the data, making graphical model analysis difficult.

Introduction

The bias resulting from the presence of measurement error in the explanatory variables is a common problem in regression analysis.

Although numerous solutions to this problem have been derived for parametric and nonparametric regression models, the corresponding problem in semiparametric specifications has remained relatively unexplored.

Motivation

In literature, semiparametric partially linear model has been mostly studied in case of the measurement error has a known distribution [1, 2].

This study presents more detailed answer to the question that how the predictions of regression functions and densities can be obtained if the measurement error has an unknown distribution in a semiparametric regression model.

The identification of the density of an unobserved random variable is possible when the joint density of two error-contaminated measurements of that variable is known [5].

Backround

$$Y_i = X_i^T \beta + g(x_i^*) + \Delta y_i$$
$$\chi = x^* + \Delta\chi$$

Backround

$$Y_i = X_i^T \beta + g(x_i^*) + \Delta y_i$$
$$\chi = x^* + \Delta \chi$$

Denote the densities of χ and x^* by $f_\chi(.)$ and $f_{x^*}(.)$, respectively. Then estimation of $f_{x^*}(.)$ is

$$\hat{f}_n(x^*) = \frac{1}{nh_n} \sum_{j=1}^{n} K_n\left(\frac{x^* - \chi_j}{h_n}\right)$$

Backround

$$Y_i = X_i^T \beta + g(x_i^*) + \Delta y_i$$
$$\chi = x^* + \Delta \chi$$

Denote the densities of χ and x^* by $f_\chi(.)$ and $f_{x^*}(.)$, respectively. Then estimation of $f_{x^*}(.)$ is

$$\hat{f}_n(x^*) = \frac{1}{nh_n} \sum_{j=1}^{n} K_n\left(\frac{x^* - \chi_j}{h_n}\right)$$

with $K_n(x^*) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} exp(-ist) \frac{\phi_K(s)}{\phi_{\Delta\chi}(s/h_n)} ds.$

Backround

Denote $\omega_{ni}(.) = K_n(\frac{\cdot - \chi_i}{h_n}) / \sum_j K_n(\frac{\cdot - \chi_j}{h_n}) \stackrel{\text{def}}{=} \frac{1}{nh_n} K_n(\frac{\cdot - \chi_i}{h_n}) / \hat{f}_n(.)$.

Backround

Denote $\omega_{ni}(.) = K_n(\frac{.-\chi_i}{h_n})/\sum_j K_n(\frac{.-\chi_j}{h_n}) \stackrel{\text{def}}{=} \frac{1}{nh_n}K_n(\frac{.-\chi_i}{h_n})/\hat{f}_n(.)$.

If β is known then the estimator of $g(.)$ is

$$g_n(x^*) = \sum_{i=1}^{n} \omega_{ni}(x^*)(Y_i - X_i^T\beta)[1].$$

Backround

Denote $\omega_{ni}(.) = K_n(\frac{.-\chi_i}{h_n})/\sum_j K_n(\frac{.-\chi_j}{h_n}) \overset{\text{def}}{=} \frac{1}{nh_n}K_n(\frac{.-\chi_i}{h_n})/\hat{f}_n(.)$.

If β is known then the estimator of $g(.)$ is

$$g_n(x^*) = \sum_{i=1}^{n} \omega_{ni}(x^*)(Y_i - X_i^T\beta)[1].$$

The generalized least squares estimator $\hat{\beta}_n$ of β can be indicated as

$$\hat{\beta}_n = (\tilde{X}^T\tilde{X})^{-1}(\tilde{X}^T\tilde{Y})$$

Backround

Denote $\omega_{ni}(.) = K_n(\frac{.-\chi_i}{h_n}) / \sum_j K_n(\frac{.-\chi_j}{h_n}) \overset{\text{def}}{=} \frac{1}{nh_n} K_n(\frac{.-\chi_i}{h_n}) / \hat{f}_n(.)$.

If β is known then the estimator of $g(.)$ is

$$g_n(x^*) = \sum_{i=1}^{n} \omega_{ni}(x^*)(Y_i - X_i^T \beta)[1].$$

The generalized least squares estimator $\hat{\beta}_n$ of β can be indicated as

$$\hat{\beta}_n = (\tilde{X}^T \tilde{X})^{-1} (\tilde{X}^T \tilde{Y})$$

where $\tilde{Y}$ denotes $(\tilde{Y}_1, ..., \tilde{Y}_n)$ with $\tilde{Y}_i = Y_i - \sum_{j=1}^{n} \omega_{nj}(\chi_i) Y_j$ and

$\tilde{X}$ denotes $(\tilde{X}_1, ..., \tilde{X}_n)$ with $\tilde{X}_i = X_i - \sum_{j=1}^{n} \omega_{nj}(\chi_i) X_j$ [2].

Backround

How can the predictions of regression functions and densities be obtained if the measurement error has an unknown distribution in a semiparametric regression model?

Estimation

The availability of two repeated measurements of x^*

$$\chi = x^* + \Delta\chi$$
$$z = x^* + \Delta z$$

provides enough information to identify any moment of the form $E[u(y^*, x^*)]$ for any function $u(y^*, x^*)$ [5].

Estimation

$$\underbrace{Y - X^T\hat{\beta}_n}_{y^*} \quad = \quad g(x^*) + \Delta y$$

where $E[\Delta y \,|x^*] = 0$

Estimation

$$\underbrace{Y - X^T \hat{\beta}_n}_{y^*} = g(x^*) + \Delta y$$

where $E[\Delta y \,|x^*] = 0$

$$\hat{g}(\tilde{x}^*, h) = \frac{n^{-1} \sum_{l=1}^{n} y_l^* K_h(x_l^* - \tilde{x}^*)}{n^{-1} \sum_{l=1}^{n} K_h(x_l^* - \tilde{x}^*)} = \frac{E[y^* K_h(x^* - \tilde{x}^*)]}{E[K_h(x^* - \tilde{x}^*)]}$$

Estimation

$$\underbrace{Y - X^T\hat{\beta}_n}_{y^*} = g(x^*) + \Delta y$$

where $E[\Delta y \, | x^*] = 0$

$$\hat{g}(\tilde{x}^*, h) = \frac{n^{-1}\sum_{l=1}^{n} y_l^* K_h(x_l^* - \tilde{x}^*)}{n^{-1}\sum_{l=1}^{n} K_h(x_l^* - \tilde{x}^*)} = \frac{E[y^* K_h(x^* - \tilde{x}^*)]}{E[K_h(x^* - \tilde{x}^*)]}$$

$K_h(x^*) = \frac{1}{h}K(\frac{x^*}{h})$

Estimation

$$\underbrace{Y - X^T\hat{\beta}_n}_{y^*} = g(x^*) + \Delta y$$

where $E[\Delta y \,|x^*] = 0$

$$\hat{g}(\tilde{x}^*, h) = \frac{n^{-1}\sum_{l=1}^{n} y_l^* K_h(x_l^* - \tilde{x}^*)}{n^{-1}\sum_{l=1}^{n} K_h(x_l^* - \tilde{x}^*)} = \frac{E[y^* K_h(x^* - \tilde{x}^*)]}{E[K_h(x^* - \tilde{x}^*)]}$$

$K_h(x^*) = \frac{1}{h}K(\frac{x^*}{h})$

Then a similar technique can be applied here, setting $u(y^*, x^*) = {y^*}^k K_h(x^* - \tilde{x}^*)$, for $k = 0, 1$.

Assumptions

1. $E[\Delta y \mid x^*, \Delta z] = 0$

   $E[\Delta \chi \mid x^*, \Delta z] = 0$

   Δz and $x^*$are mutually independent.

Assumptions

1. $E[\Delta y \mid x^*, \Delta z] = 0$

   $E[\Delta \chi \mid x^*, \Delta z] = 0$

   Δz and $x^*$are mutually independent.

2. $E[|x^*|], E[|\Delta \chi|]$ and $E[|y^*|]$ are finite.

Assumptions

1. $E[\Delta y \mid x^*, \Delta z] = 0$

   $E[\Delta\chi \mid x^*, \Delta z] = 0$

   Δz and $x^*$are mutually independent.

2. $E[|x^*|], E[|\Delta\chi|]$ and $E[|y^*|]$ are finite.

3. $E[y^{*k} h^{-1} K(h^{-1}(x^* - \tilde{x}^*))] < \infty$ for all $\tilde{x}^*$, any $h > 0$, and $k = 0, 1$.

Theorem

Under Assumptions $1-3$, and provided $\left|E[e^i\xi z]\right| > 0$ for any finite ξ, the function

$$\hat{g}(\tilde{x}^*, h) = \frac{E[y^* K_h(x^* - \tilde{x}^*)]}{E[K_h(x^* - \tilde{x}^*)]}$$

for $\tilde{x}^* \in \mathbb{R}$ and $h \geq 0$, can be expressed solely in terms of moments that involve the observable variables y^*, χ and z:

Theorem (Fourier representation of the numerator and the denominator of the Nadaraya-Watson estimator)

$$\hat{g}(\tilde{x}^*, h) = \frac{M_1(\tilde{x}^*, h)}{M_0(\tilde{x}^*, h)}$$

where, for $k = 0, 1$,

$M_k(\tilde{x}^*, h) = \frac{1}{2\pi} \int \kappa(h\xi)\phi_k(\xi) exp(-i\xi x^*) d\xi$

Theorem

and where $\phi_k(\xi) \equiv E[{y^*}^k exp(i\xi x^*)]$ is given by

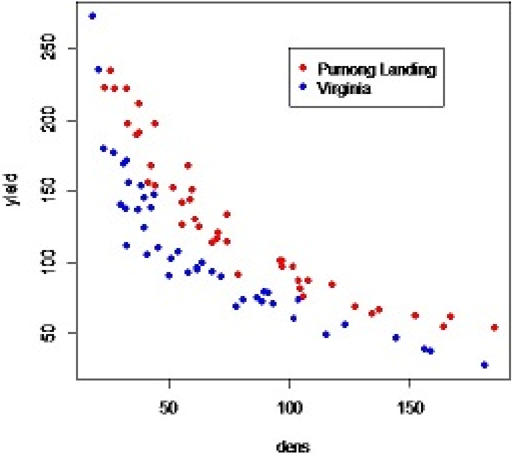
$$\phi_0(\xi) = exp(\int_0^\xi \frac{im_\chi(\zeta)}{m_1(\zeta)}),$$

$$\phi_1(\xi) = \phi_0(\xi)\frac{m_y^*(\xi)}{m_1(\xi)},$$

where $i = \sqrt{-1}$ and $\kappa(\xi)$ is the Fourier transform of the kernel $K(x^*)$ and $m_a(\xi) = E[aexp(i\xi z)]$ for $a = 1, \chi, y^*$.

Example

Scatterplot of the density and log.yield for the onions data. The plotting symbols indicate the two locations where the onions were cultivated [3].



Example

The semiparametric binary offset model for these data

$$log(yield_i) = \beta_1 PL_i + f(density_i) + \varepsilon_i$$

$$PL_i = \begin{cases} 0 & \text{if } i\text{th measurement is from Virginia,} \\ 1 & \text{if } i\text{th measurement is from Purnong Landing.} \end{cases}$$

Example

	Fourier	N-W	No M.Error
Bias Squared	1.1740	4.0640	3.4985
Variance	0.7475	0.0372	0.0413
Mean Square Error	1.3243	0.3042	0.3495

Tablo: Onions Data Results

Table compares the bias squared, the variance, and the mean square error of the three estimators considered. We choose bandwidth as $h = 1$.

Example

In comparison with the Nadaraya-Watson estimator, our estimator is clearly very effective at reducing the bias.
Of course, because the variance of our estimator is larger than the Nadaraya-Watson estimator, and the resulting bias, is slightly larger than in the error-free case.

The bias reduction made possible by the proposed estimator comes at the expense of an increased variance relative to the Nadaraya-Watson estimator. However, the decrease in the bias more than offsets the increase in the variance, so that the mean square error we obtain is still better than for the Nadaraya-Watson estimator.

Conclusion

This study presents a new kernel-based semiparametric estimator that extends the conventional Nadaraya-Watson kernel estimator to cover the case of an error ridden regressor. Identification is achievable when one repeated measurement of the error-contaminated regressor is available.

One remarkable property of our estimator is that it requires no knowledge of the distribution of the measurement error, contrary to the popular kernel deconvolution estimator.

Further Studies

Asymptotic Properties

We are going to compose the analysis of the asymptotic properties of the proposed estimator $\hat{g}(\tilde{x}^*, h)$. With this approach we will try to enable the derivation of the convergence rate and to establish the asymptotic normality of the estimator.

Simulation Study

We are going to add a simulation study to investigate the finite-sample properties of the proposed estimator through various Monte Carlo simulations.

References

- Fan, J.and Truong, Y.K. (1993). Nonparametric regression with errors in variables. *Annals of Statistics*, 21: 1900â1925.
- Liang, H. (2000). Asymptotic normality of parametric part in partially linear model with measurement error in the non-parametric part. *Journal of Statistical Planning and Inference*, 86: 51â62.
- Ratkowsky, D.A. (1983). Nonlinear Regression Modeling: A Unified Practical Approach. *New York: Marcel Dekker.*
- Ruppert, D.,Wand M.P. and Carroll, R.J. (2003). Semiparametric Regression. *Cambridge Series in Statistical and Probabilistic Mathematics*.
- Schennach, S.M. (2004). Nonparametric regression in the presence of measurement error. *Econometric Theory*, 20: 1046â1093.

THANK YOU