ELSEVIER

Statistics & Probability Letters 58 (2002) 427–433

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

A simple adjustment for measurement errors in some limited dependent variable models ☆

Liqun Wang

Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2

Received November 2000; received in revised form May 2001

Abstract

This paper proposes a simple method to adjust for measurement errors in estimations of many popular limited dependent variable models, e.g., the binary response model, the censored and the truncated regression models. The procedure is based on a simple correction of the estimators for the corresponding "error-free" models and is easy to be incorporated into the existing statistical computer packages. The extra computing cost is minimal.

Keywords: Errors-in-variables; Binary response model; Censored data; Truncated data

1. Introduction

The binary response, the censored and the truncated regression models are examples of limited dependent variable (LDV) models which are widely used in biology, epidemiology, medicine, economics, engineering and many other fields. In practical applications it is often the case where some or all predictor variables are not or cannot be measured exactly, rather they are subject to measurement errors (ME). It is well known that the statistical inference of model parameters are biased and inconsistent in the presence of measurement errors if they are ignored (Fuller, 1987; Carroll et al., 1995). The statistical theories and methods for the "measurement-error-free" LDV models have been well developed, see, e.g., Amemiya (1985), Maddala (1985) and Greene (1993). The problem of ME has been paid more and more attention in recent years. See, e.g., Fuller (1987) for statistical theories and methods for linear ME models, and Carroll et al. (1995) for nonlinear models. For nonlinear models, most estimation procedures are based on corrections of certain estimating equations, such as the likelihood or quasi-likelihood equations. Recently, more sophisticated semiparametric

☆ This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.
E-mail address: liqun_wang@umanitoba.ca (L. Wang).

0167-7152/02/$ - see front matter
PII: S0167-7152(02)00165-7

and simulation estimation methods have been developed for various nonlinear models (Sepanski and Carroll, 1993; Wang and Hsiao, 1996).

Despite its importance and seriousness, the problem of ME has still been often ignored in practical applications, especially in econometrics. This might be because of two reasons. The first reason is that, many practitioners believe that, although the data are subject to ME, the magnitude of the ME is so small that their effects are negligible. The second reason is that in almost all statistical computer packages only the procedures for the "error-free" models are available. The first reasoning is clearly not justified, because in many situations the magnitude of the ME cannot be exactly determined from the out-of-sample sources. For the second point, it is helpful to develop procedures which takes the ME into account and which are easy enough to be incorporated into existing computer programs.

The aim of this paper is to derive such simple procedures for a class of LDV models under normal distributions. This class contains many popular models, such as the binary response, the censored regression and the truncated regression models. The idea is, first, to reduce the model into an "error-free" form; and then, to use the existing estimation procedures for the later to obtain the estimates for the original model through the parameter transformation. The only extra job is to calculate new asymptotic covariance matrix, which is also easy and straightforward.

Section 2 introduces the general LDV models with measurement errors and shows that this model can be reduced to the traditional "error-free" form. Section 3 proposes a two-step procedure to estimate the model. It is also shown that, using this approach it is possible to assess the effects of the ME and the consequence of misspecified a priori information. Section 4 shows how this estimation procedure is adapted to some popular LDV models. Conclusions are contained in Section 5.

2. The model and its reduction

We start with the underlying linear relationship

$$\eta_i = \beta_1 + \beta_2' \xi_i + u_i, \tag{1}$$

where $\eta_i \in \mathbb{R}$, $\xi_i \in \mathbb{R}^k$ are the response and explanatory variables, β_1, β_2 the regression parameters and u_i the random error with $E(u_i) = 0$ and $\mathrm{Var}(u_i) = \sigma_u$. In the LDV framework it is assumed that the dependent variable η_i is not fully observed. Rather, we observe

$$y_i = g(\eta_i), \tag{2}$$

where $g(\cdot)$ is a known real function. For example, a binary response model has $g(\eta) = I(\eta \geqslant 0)$, where $I(\cdot)$ is the indicator function, and a censored regression model has $g(\eta) = \eta I\ (\eta \geqslant 0)$. In addition, we assume that the explanatory variable ξ_i is not exactly observed, rather we observe

$$x_i = \xi_i + v_i, \tag{3}$$

where $x_i \in \mathbb{R}^k$ and $v_i \in \mathbb{R}^k$ represents the additive measurement error (ME). In the ME literature it is often assumed that x_i is a surrogate for ξ_i, i.e., the conditional distribution $f(y \mid \xi, x) = f(y \mid \xi)$ (Carroll et al., 1995). Furthermore, we assume that u_i, v_i and ξ_i are independently and normally distributed with means $0, 0, \mu_\xi$ and variances $\sigma_u, \Sigma_v, \Sigma_\xi$, respectively. If some components of ξ are exactly observed, then the corresponding components of v_i are constant zero and the corresponding

rows and columns of Σ_v are zeros. Thus, specification (1)–(3) includes the usual "error-free" LDV models.

As the ME bring much more uncertainty into the inference process, usually more a priori information is needed to identify the model (Fuller, 1987; Hsiao, 1983). This identifying information may take different forms, depending on the feature of the problem and the data at hand. In many cases the repeated sampling or validation data are available, which may be used to determine or estimate covariance Σ_v of the ME. This is equivalent to that the so-called reliability ratio $\Sigma_x^{-1}\Sigma_\xi$ is known in the sense that $\Sigma_x^{-1}\Sigma_\xi = I - \Sigma_x^{-1}\Sigma_v$ and Σ_x may be easily estimated from observed data. Throughout this paper we suppose that the covariance Σ_v is known or a consistent estimator of Σ_v is available.

Now we show that models (1)–(3) may be reduced to an "error-free" form. Indeed, writing

$$\xi_i = \Sigma_v\Sigma_x^{-1}\mu_x + \Sigma_\xi\Sigma_x^{-1}x_i + \Sigma_v\Sigma_x^{-1}(x_i - \mu_x) - v_i$$

and substituting this equation into (1) result in

$$\eta_i = \gamma_1 + \gamma_2' x_i + w_i, \tag{4}$$

where $\gamma_1 = \beta_1 + \beta_2'\Sigma_v\Sigma_x^{-1}\mu_x$, $\gamma_2 = \Sigma_x^{-1}\Sigma_\xi\beta_2$ and

$$w_i = u_i - \beta_2' v_i + \beta_2'\Sigma_v\Sigma_x^{-1}(x_i - \mu_x). \tag{5}$$

The new error w_i has a normal distribution $\mathrm{N}(0,\sigma_w)$ with $\sigma_w = \sigma_u + \gamma_2'\Sigma_v\Sigma_\xi^{-1}\Sigma_x\gamma_2$ and is independent of x_i because $E(w_i x_i) = 0$. Eqs. (4) and (2) combined represent a usual LDV model for (y_i, x_i'). The transformation between the parameters $(\gamma_1, \gamma_2, \sigma_w, \mu_x, \Sigma_x)$ in models (4) and (2) and the parameters $(\beta_1, \beta_2, \sigma_u, \mu_\xi, \Sigma_\xi)$ in the original model (1)–(3) are given by

$$\beta_1 = \gamma_1 - \mu_x'\Sigma_\xi^{-1}\Sigma_v\gamma_2, \tag{6}$$

$$\beta_2 = \Sigma_\xi^{-1}\Sigma_x\gamma_2, \tag{7}$$

$$\sigma_u = \sigma_w - \gamma_2'\Sigma_v\Sigma_\xi^{-1}\Sigma_x\gamma_2, \tag{8}$$

$$\mu_\xi = \mu_x, \tag{9}$$

$$\Sigma_\xi = \Sigma_x - \Sigma_v. \tag{10}$$

Clearly transformation (6)–(10) is one-to-one for the given Σ_v. Consequently, any estimator for model (4) and (2) implies a corresponding estimator for model (1)–(3) through (6)–(10).

3. Model estimation

3.1. A two-step estimator

Let (y_i, x_i'), $i = 1, 2, \ldots, n$ be the independent and identically distributed observations and denote the sample moments as $\hat{\mu}_x = (1/n)\sum_{i=1}^n x_i$ and $\hat{\Sigma}_x = (1/n)\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})'$. As assumed before, the covariance Σ_v is known or an estimate of Σ_v is available. Without loss of generality we assume

further that the estimate $\hat{\Sigma}_\xi = \hat{\Sigma}_x - \Sigma_v$ is nonsingular with probability one. Then estimates of model (1)–(3) may be obtained by the following two steps:

Step 1: Use the given data to compute the sample moments $\hat{\mu}_x$, $\hat{\Sigma}_x$ and the estimates $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\sigma}_w$ for models (4) and (2).

Step 2: Use (6)–(10) to compute the estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}_u$, $\hat{\mu}_\xi$ and $\hat{\Sigma}_\xi$ for the ME model (1)–(3).

Note that in practical applications, the moment estimate $\hat{\Sigma}_\xi = \hat{\Sigma}_x - \Sigma_v$ may not always be positive-definite, as in variance component problems. The problem, however, becomes less serious when the sample size is large.

Now, we derive the asymptotic properties of the estimators obtained through this two-stage procedure. Since the asymptotic properties for estimators $\hat{\mu}_\xi = \hat{\mu}_x$ and $\hat{\Sigma}_\xi = \hat{\Sigma}_x - \Sigma_v$ are easy to derive from the sample moments, in the following we concentrate on the estimators of $\theta = (\beta_1, \beta_2', \sigma_u)'$. In particular, we denote the first-step estimators as $\hat{\psi} = (\hat{\gamma}_1, \hat{\gamma}_2', \hat{\sigma}_w)'$ and the second-step estimators as $\hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2', \hat{\sigma}_u)'$. Let $\theta(\psi) : \Psi \mapsto \Theta$ denote the transformation (6)–(10) and $\hat{\theta} = \theta(\hat{\psi})$. Then the consistency of $\hat{\theta}$ follows immediately from the continuity of $\theta(\psi)$. To show the asymptotic normality of $\hat{\theta}$, note that $\theta(\psi)$ is continuously differentiable and hence we have the first-order Taylor expansion

$$\hat{\theta} - \theta = \frac{\partial \theta(\tilde{\psi})}{\partial \psi'}(\hat{\psi} - \psi), \tag{11}$$

where ψ and θ correspond to the true parameters of model (1)–(3) and $\tilde{\psi}$ lies between $\hat{\psi}$ and ψ. The derivative $\partial\theta/\partial\psi'$ is obviously a continuous function of ψ and, therefore, converges in probability to the matrix

$$A = \begin{pmatrix} 1 & -\mu_x' \Sigma_\xi^{-1} \Sigma_v & 0 \\ 0 & \Sigma_\xi^{-1} \Sigma_x & 0 \\ 0 & -2\beta_2' \Sigma_v & 1 \end{pmatrix}. \tag{12}$$

The following results follow then by Slutsky Theorem (Amemiya, 1985).

Theorem 1. *Suppose* Σ_v *is known,* $\Sigma_\xi > 0$, $\hat{\psi} \xrightarrow{\mathrm{P}} \psi$ *and* $\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{L} \mathrm{N}(0, \Omega)$. *Then, as* $n \to \infty$,

(1) $\hat{\theta} \xrightarrow{\mathrm{P}} \theta$,

(2) $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{L} \mathrm{N}(0, A\Omega A')$, *where* A *is given in* (12).

The efficiency of the second-step estimator depends on the efficiency of the first-step estimator. If the maximum likelihood estimator is used in the first step, then as a function of a maximum likelihood estimator, the second-step estimator is also efficient.

3.2. The effects of ME

If one ignores the ME and estimates the model parameters with x in the place of ξ, then the resulting estimators converge to $\psi = (\gamma_1, \gamma_2', \sigma_w)'$ instead of $\theta = (\beta_1, \beta_2', \sigma_u)'$. From (6)–(10) it is easy

to see that the asymptotic biases are given by

$$\gamma_1 - \beta_1 = \mu_x' \Sigma_x^{-1} \Sigma_v \beta_2,$$

$$\gamma_2 - \beta_2 = -\Sigma_x^{-1} \Sigma_v \beta_2,$$

$$\sigma_w - \sigma_u = \beta_2' \Sigma_\xi \Sigma_x^{-1} \Sigma_v \beta_2.$$

These equations show that the asymptotic biases can be significant unless $\Sigma_x^{-1}\Sigma_v\beta_2 \approx 0$. In the special case where β_2 is scalar, this means that the asymptotic biases can be significant, if the ME variance Σ_v is not very small relative to the variance of the observed covariates Σ_x. Furthermore, the slope parameter β_2 tends to be underestimated by positive β_2 and overestimated by negative β_2, whereas the opposite is true for β_1 and σ_u.

3.3. The effects of misspecified Σ_v

Using this two-step approach it is also possible to derive asymptotic biases of the estimators of model (1)–(3) when the identifying information Σ_v is misspecified. For instance, if $(\hat{\gamma}_1, \hat{\gamma}_2', \hat{\sigma}_w)'$ are consistent estimators of models (4) and (2) and $(\tilde{\beta}_1, \tilde{\beta}_2', \tilde{\sigma}_u)'$ are obtained through (6)–(10) where, instead of Σ_v, a wrong $\tilde{\Sigma}_v$ is used. Then, the asymptotic biases are given by

$$\operatorname{plim} \tilde{\beta}_1 - \beta_1 = \mu_x'(\Sigma_x - \tilde{\Sigma}_v)^{-1}(\Sigma_v - \tilde{\Sigma}_v)\beta_2,$$

$$\operatorname{plim} \tilde{\beta}_2 - \beta_2 = -(\Sigma_x - \tilde{\Sigma}_v)^{-1}(\Sigma_v - \tilde{\Sigma}_v)\beta_2,$$

$$\operatorname{plim} \tilde{\sigma}_u - \sigma_u = \beta_2' \Sigma_\xi (\Sigma_x - \tilde{\Sigma}_v)^{-1}(\Sigma_v - \tilde{\Sigma}_v)\beta_2.$$

From these equations we see that the estimation biases are of the same order as $\Sigma_v - \tilde{\Sigma}_v$ and, hence, can be significant unless $(\Sigma_v - \tilde{\Sigma}_v)\beta_2 \approx 0$. Again, in the case where β_2 is scalar, this means that the asymptotic biases can be significant, if the amount of misspecification $\Sigma_v - \tilde{\Sigma}_v$ is not very small relative to $\Sigma_x - \tilde{\Sigma}_v$. Furthermore, the slope parameter β_2 tends to be underestimated by underspecified Σ_v and overestimated by overspecified Σ_v, whereas the converse is true for β_1 and σ_u.

4. Some special models

The general framework of Sections 2 and 3 covers many popular LDV models, e.g., the binary response, the censored and the truncated regression models. In this section we show how to adapt the two-step estimation procedure to these special models.

4.1. The censored regression model

A censored regression model has the form (1)–(3) with $g(\eta_i) = \eta_i I\ (\eta_i \geqslant 0)$, i.e., one observes $y_i = \eta_i$ when $\eta_i > 0$ and $y_i = 0$ otherwise. This model was first applied to economic problem by Tobin (1958) and therefore is known as the Tobit model in econometrics. It is well known that the reduced models (4) and (2) is identified in general and, hence, all parameters can be consistently

estimated (Amemiya, 1985). Therefore, the two-step approach of Section 3 applies immediately to the censored regression model.

4.2. The truncated regression model

The truncated linear regression model is applied to data sets where only positive outcomes of the dependent variable is observed, i.e., one observes $y_i = \eta_i$ when $\eta_i > 0$ and neither η_i nor x_i are observed when $\eta_i \leqslant 0$. Similar to the censored regression model, for the error-free truncated regression model, the traditional moments estimator and more efficient maximum likelihood estimator for all parameters in models (4) and (2) are available (Greene, 1993). Thus, the two-step estimation procedure of Section 3 applies immediately.

4.3. The binary response model

The binary response model is defined as (1)–(3) with $g(\eta_i) = I\ (\eta_i \geqslant 0)$, i.e., one observes only the signs of the dependent variable η_i and not its values. In this case the observed dependent variable y_i takes only binary values 0 or 1. It is because of this very limited information, additional restriction is needed for the model to be identified. As for the Probit model, it is usually assumed that $\sigma_u = 1$, which we do in this paper. Under this assumption the error variance in the reduced model becomes $\sigma_w = 1 + \gamma_2' \Sigma_v \Sigma_\xi^{-1} \Sigma_x \gamma_2$.

For reduced models (4) and (2), standard procedures yield consistent estimates of the ratios $\alpha_j = \gamma_j / \sqrt{\sigma_w},\ j = 1, 2$. Let $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2')'$ be this first-step estimator. Then the second-step estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2')'$ is computed through (6)–(10) as

$$\beta_1 = \sqrt{\sigma_w}(\alpha_1 - \mu_x' \Sigma_\xi^{-1} \Sigma_v \alpha_2),$$

$$\beta_2 = \sqrt{\sigma_w} \Sigma_\xi^{-1} \Sigma_x \alpha_2,$$

where $\sigma_w = (1 - \alpha_2' \Sigma_v \Sigma_\xi^{-1} \Sigma_x \alpha_2)^{-1}$. Correspondingly, Eq. (11) is modified as

$$\hat{\beta} - \beta = \hat{B}(\hat{\alpha} - \alpha),$$

where $\alpha = \gamma / \sqrt{\sigma_w}$ and $\hat{B}$ converges in probability to

$$B = \sqrt{\sigma_w} \begin{pmatrix} 1 & -\mu_x' \Sigma_\xi^{-1} \Sigma_v \\ 0 & \Sigma_\xi^{-1} \Sigma_x \end{pmatrix}.$$

It follows that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{L} \mathrm{N}(0, B \tilde{\Omega} B'),$$

where $\tilde{\Omega}$ is the asymptotic covariance matrix of $\hat{\alpha}$.

Remark 1. For some special LDV models with measurement errors, direct estimation procedures have been used in the literature. For the Probit model, Carroll et al. (1984) proposed maximum likelihood estimator. For the censored regression model under a slightly different set-up, Wang (1998)

proposed moment and maximum likelihood estimators. A brief survey of estimation methods for the censored regression models both with and without measurement errors can be found in Wang (1999).

In general, the two-step procedure of this paper should give the same results as the existing maximum likelihood estimators, if the MLE are used in the first step. However, from the practical applications point of view, either explicit formula for covariance of the existing estimators is not available, or the estimation procedure involves sophisticated numerical computation and heavy computer programing. In contrast, the two-step estimator of this paper is very easy to implement and to be incorporated into the existing computer packages.

5. Conclusions

An easy to implement approach for correcting the effects of measurement errors in certain limited dependent variable models under normality is proposed. We have shown that such models may be transformed into an "error-free" form. The parameters in the model can be expressed as simple transformations of the reduced-form parameters, which may be estimated using usual procedures available in many statistical computer packages. This procedure is easy to implement and the marginal computation cost is minimal. The models covered by the framework include the binary response model, the censored and truncated regression models, as well as the linear model.

References

Amemiya, T., 1985. Advanced Econometrics. Basil Blackwell, Oxford.

Carroll, R.J., Spiegelman, C.H., Gordon, K.K., Bailey, K.K., Abbott, R.D., 1984. On errors-in-variables for binary regression models. Biometrika 71, 19–25.

Carroll, R.J., Ruppert, D., Stefanski, L.A., 1995. Measurement Error in Nonlinear Models. Chapman & Hall, London.

Fuller, W.A., 1987. Measurement Error Models. Wiley, New York.

Greene, W.H., 1993. Econometric Analysis, 2nd Edition. Macmillan, New York.

Hsiao, C., 1983. Identification. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, Vol. I. North-Holland, Amsterdam.

Maddala, G.S., 1985. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

Sepanski, J.H., Carroll, R.J., 1993. Semiparametric quasilikelihood and variance function estimation in measurement error models. J. Econometrics 58, 223–256.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica 26, 24–36.

Wang, L., 1998. Estimation of censored linear errors-in-variables models. J. Econometrics 84, 383–400.

Wang, L., 1999. Censored linear regression models. In: Kotz, S., Read, C., Banks, D. (Eds.), Encyclopedia of Statistical Sciences. Wiley, New York, pp. 71–75.

Wang, L., Hsiao, C., 1996. A semiparametric estimation of nonlinear errors-in-variables models. Proceedings of the Business and Economic Statistics Section, American Statistical Association, Alexandria, VA, pp. 231–236.