Journal of Econometrics 32 (1986) 291–292. North-Holland

REVERSE REGRESSIONS FOR LATENT-VARIABLE MODELS

David K. LEVINE*

University of California, Los Angeles, CA 90024, USA

Received June 1985, final version received April 1986

Under joint normality of all regressors, the errors-in-variables bounds for linear regression may be extended to probit and related models of censorship and truncation.

Klepper and Leamer (1984) show how to find bounds and other diagnostics for the errors-in-variables model. In this paper I show how to extend their results to models with a latent dependent variable, such as probit, and models of censorship and truncation. Klepper and Leamer's results, as well as similar results of Koopmans (1937), Kalman (1981) and others, are based on the idea of reverse regressions, that is, regressing each of the explanatory variables measured with error on the remaining explanatory variables and the endogenous variable. When the endogenous variable is a latent variable so that it is observed only via a proxy the direct analogue of a reverse regression is not useful. The problem is that the covariance between the latent variable and explanatory variables cannot be estimated by the sample moments, since the latent variable is unobservable. However, the covariance between the latent and explanatory variables can be estimated from knowledge of the regression coefficients of the latent on the explanatory variables, and these coefficients can be estimated by maximum likelihood.

Adopting the notation and assumptions of Klepper and Leamer the latent dependent variable y_t, the unobservables χ_t and the observables x_t are jointly normally distributed with $x_t - \chi_t$ independent of χ_t and y_t. Let N be the covariance matrix of x_t and D that of $x_t - \chi_t$. Let σ_y^2 be the variance of y and σ_b^2 be the variance of y conditional on x_t. Because of joint normality $\mathrm{E}[y_t|x_t] = b_0 + b'x_t$, while $\mathrm{E}[y_t|\chi_t] = \beta_0 + \beta'\chi_t$. Since $x_t - \chi_t$ is independent of y_t, it follows that $\operatorname{cov}(y_t, x_t) = \operatorname{cov}(y_t, \chi_t)$ or that $Nb = (N - D)\beta$. Consequently,

$$\beta = (N - D)^{-1} Nb. \tag{1}$$

*I am grateful to Tim Erickson and Ed Leamer for stimulating my interest in this problem, and for helpful discussions. I am grateful to an anonymous referee for helpful criticism.

Moreover, joint normality implies

$$\sigma_y^2 = \sigma_b^2 + b'Nb. \tag{2}$$

Klepper and Leamer assume y_t is directly observable, so that σ_y^2 and $\mathrm{cov}(y_t, x_t)$ may be estimated directly from sample moments. Suppose instead that y_t is not observed, but instead only $z(y_t)$ is observable. For example, in probit

$$\begin{aligned} z(y_t) &= 1 \quad \text{for} \quad y_t > 0, \\ &= 0 \quad \text{for} \quad y_t \le 0, \end{aligned}$$

and models of censorship and truncation can be similarly represented. Obviously σ_y^2 and $\mathrm{cov}(y_t, x_t)$ cannot be consistently estimated by sample moments using z_t in place of y_t. However, once b is estimated by maximum likelihood, (1) and (2) above show the correct estimation procedure for σ_y^2 and $\mathrm{cov}(y_t, x_t)$ and the methods of Klepper and Leamer may then be applied.

In the case of probit, there is a slight complication in that b and β can only be estimated up to a normalization. Let σ_β^2 be the variance of y_t conditional on χ_t. Estimating b using probit we would ordinarily impose the normalization $\sigma_b^2 = 1$, while in estimating β we would ordinarily impose the normalization $\sigma_\beta^2 = 1$. The most straightforward solution, since only the ratio of components of β can be estimated anyway, is to impose $\sigma_b^2 = 1$, implicitly using the normalization

$$\sigma_\beta^2 = 1 + b'Nb - \beta'(N - D)\beta. \tag{3}$$

Relative to this normalization, the Klepper–Leamer results apply directly.

The upshot is that b and σ_b^2 should be estimated ignoring measurement error and using x_t in place of χ_t. The matrix N is estimated by the empirical moment matrix for x_t. These estimates are then combined using

$$\mathrm{var}(y_t, x_t') = \begin{bmatrix} \sigma_b^2 + b'Nb & b'N \\ Nb & N \end{bmatrix}, \tag{4}$$

to which the Klepper and Leamer methods apply.

References

Kalman, R.G., 1981, System identification from noisy data, Paper presented at the International Symposium of Dynamical Systems, Gainesville, FL.

Klepper, S. and E. Leamer, 1984, Consistent sets of estimates for regressions with errors in all variables, Econometrica 52, 163–183.

Koopmans, T., 1937, Linear regression analysis of economic time series (Netherlands Econometric Institute, Haarlem/de Erven F. Bohn, New York).