ISSN 0103-2509

Conditional likelihood estimation of a simple regression model with measurement error via the orthogonal parametrization.

RODRIGUES, J. & CORDANI, L.K.

Nº 3

São Carlos (SP)
1989

Conditional likelihood estimation of a simple regression model with measurement error via the orthogonal parametrization

Josemar Rodrigues
ICMSC-USP CP-668 CEP-13.560
SÃO CARLOS-SP-BRAZIL
Lisbeth K. Cordani
IME-USP-Rua do Matão,1010
SÃO PAULO-SP-BRAZIL

1 Summary

In this paper, working with an orthogonal parametrization, conditional likelihood based methods for obtaining exact and approximate confidence intervals for the slope in a simple linear regression model with measurement error are discussed. Our setup is simple and different from the one considered by Wong (1989).

2 Introduction

The classical linear regression analysis assumes that the independent variable is defined by

$$\begin{aligned} Y_t &= \alpha + \beta x_t + e_t, \\ t &= 1, \ldots, n \end{aligned} \tag{1}$$

where $(x_1, \ldots, x_n)$ is fixed in repeated sampling and e_t are independent $N(0, \sigma^2_{ee})$ random variables. It is assumed that x_t is measured without error,however, in practice, particulary in social sciences and biological assay, this assumption is often violated. There are a lot of work on the problem of parameter estimation when x_t contain errors of measurement, see for example, Madansky (1959), Sprent (1966),Kendall & Stuart (1961) and recently Fuller (1987). In the present paper we try to solve the estimation problem by working with a conditional orthogonal likelihood model, when only the independent variable is measured with error. We shall study models of type (1), with $\alpha = 0$, where instead of observing x_t, one observes the sum

$$X_t = x_t + u_t \tag{2}$$

where u_t is a $N(0, \sigma^2_{uu})$–random variable.We make the customary assumption that

$$(x_t, u_t, e_t) \sim N((\mu_x, 0, 0), diag(\sigma^2_{xx}, \sigma^2_{uu}, \sigma^2_{ee})). \tag{3}$$

We can without loss of generality take $\mu_x = 0$. Because (X_t, Y_t) is distributed as a bivariate normal, the conditional distribution of Y_t given X_t is normal with mean θX_t and variance $\sigma^2_{uu} K_{xx} \beta^2 + \sigma^2_{ee}$, where

$$K_{xx} = \frac{\sigma^2_{xx}}{\sigma^2_{xx} + \sigma^2_{uu}} \tag{4}$$

$$\theta = K_{xx}\beta. \tag{5}$$

From (5) we see the effect of the measurement error, that is, the regression coefficient β has been attenuated by the factor K_{xx}. The ratio K_{xx} is called the reliability of X_t in the social science,so as suggested by Fuller (1987), we call K_{xx} the reliability ratio. Working with orthogonal parametrization (Cox & Reid,1987), we constructed exact and approximate confidence intervals for β when the reliability ratio is known and unknown.

3 Exact confidence interval with known reliability ratio

In this section, our main interest is to make inference about β under the assumption that the reliability ratio is known. Assuming that x is a particular trait of the individual,there are a number of situations,particulary in psychology and sociology, where the factor K_{xx} is known. The QI test is the best known example. The value of the trait x for a particular individual is impossible to observe directly, and the response X_t is obtained as a score constructed from the answers to a number of questions.When this battery of questions has beeen studied, it is possible to get an idea of K_{xx}. In general,the factor is published together with the supporting material. The procedure to obtain a confidence interval for β is simplified by working with orthogonal parametrization,

$$\lambda = \sigma_{uu}^2 K_{xx}\beta^2 + \sigma_{ee}^2. \tag{6}$$

If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independently and identically distributed according to the model (3), the conditional likelihood function in the new parametrization is

$$l(\beta, \lambda) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln\lambda - \frac{1}{2\lambda}\sum_{t=1}^{n}(Y_t - \beta K_{xx}X_t)^2. \tag{7}$$

It is well known that maximum likelihood estimators of β and λ are

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{t=1}^{n} K_{xx}X_tY_t}{\sum_{t=1}^{n} K_{xx}^2 X_t^2} \\
&= K_{xx}^{-1}\frac{\sum_{t=1}^{n} X_tY_t}{\sum_{t=1}^{n} X_t^2} \\
&= K_{xx}^{-1}\hat{\beta}_1 && (8) \\
\hat{\lambda} &= \frac{1}{n-1}\sum_{t=1}^{n}(Y_t - \hat{\beta}_1 X_t)^2. && (9)
\end{aligned}$$

The same result was obtained by Fuller (1987) under a different way and a more general model. It is important here to emphasize the simplicity in obtaining the results by working with the orthogonal parameters.Now, the exact $1-\alpha$ level confidence interval for β is formed via the standardized student deviate

$$\left[\hat{\beta} - t_{\frac{1}{2}\alpha}\sqrt{\frac{\hat{\lambda}}{K_{xx}^2\sum_{t=1}^{n}X_t^2}}, \hat{\beta} + t_{\frac{1}{2}\alpha}\sqrt{\frac{\hat{\lambda}}{K_{xx}^2\sum_{t=1}^{n}X_t^2}}\right]. \tag{10}$$

For testing the hypothesis $\beta = \beta_0$, we can use the t-statistic

$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{\frac{\hat{\lambda}}{K_{xx} \sum_{t=1}^{n} X_t^2}}} \sim t(n-1). \tag{11}$$

4 Approximate confidence interval with unknown reliability ratio

It is well known that, when both σ_{uu}^2 and σ_{ee}^2 are unknown, β is not identifiable. The model (1)-(2), with the known ratio $\delta = \frac{\sigma_{ee}^2}{\sigma_{uu}^2}$,is called error-in-variables model. We suggested to the reader to see Fuller's book (1987) p. 30 for interesting comments on that model. In this section, we consider the " overidentifiable" situation, that is, $\sigma_{ee}^2 = \sigma_{uu}^2 = \sigma_0^2$, where the common variance is known. The "overidentifiable case" has aroused considerable comments and the maximum likelihood estimates were obtained by Barnett (1967). In this special case, we consider the parameter of interest as $\varphi = \theta$ and the nuisance parameter as $\phi = \beta$. Using the same argument of section 3, we have from (6)that

$$\lambda = \sigma_0^2(\phi\varphi + 1). \tag{12}$$

It is easy to see that λ is orthogonal to φ and that its natural estimators are

$$\begin{aligned} \hat{\theta} &= \hat{\beta}_1 \\ \hat{\lambda} &= \frac{\sum_{t=1}^{n}(Y_t - \hat{\beta}_1 X_t)^2}{n-1}. \end{aligned} \tag{13}$$

$$\sqrt{n}\begin{pmatrix} \hat{\theta} - \theta \\ \hat{\lambda} - \lambda \end{pmatrix} \Longrightarrow N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\lambda}{\sigma_{XX}^2} & 0 \\ 0 & 2\lambda^2 \end{pmatrix}\right), \tag{14}$$

where $\sigma_{XX}^2 = \sigma_{xx}^2 + \sigma_{uu}^2$. Since we are interested to find an approximate confidence interval for β, let us consider,from (12), the following reparametrization:

$$\beta = \frac{\frac{\lambda}{\sigma_0^2} - 1}{\theta}. \tag{15}$$

From (14)-(15) we get that

$$\begin{aligned}
\sqrt{n}(\hat{\beta}-\beta) &\Longrightarrow N(0;\sigma^2), \text{ where} \\
\sigma^2 &= \left(\frac{\frac{\lambda}{\sigma_0^2}-1}{\theta^2}\right)^2 \frac{\lambda}{\sigma_{XX}^2} + 2\lambda^2 \frac{1}{\sigma_0^2\theta^2} \quad \text{and} \qquad (16) \\
\hat{\beta} &= \frac{\frac{\hat{\lambda}}{\hat{\sigma_0^2}}-1}{\hat{\theta}}. \qquad (17)
\end{aligned}$$

An approximate $(1-\alpha)-$ level confidence interval for β is formed via standardized deviate

$$\left(\hat{\beta} - z_{\frac{1}{2}\alpha}\sqrt{\frac{\hat{\sigma^2}}{n}}, \hat{\beta} + z_{\frac{1}{2}\alpha}\sqrt{\frac{\hat{\sigma^2}}{n}}\right) \qquad (18)$$

where $\hat{\sigma^2}$ and $\hat{\beta}$ are obtained by (13) from (15)and (16), with $\hat{\sigma_{XX}^2} = \frac{\sum_{i=1}^{n} X_i^2}{n-1}$. It is interesting to mention that if we take $\varphi = \beta$ and $\phi = \theta$, as pointed out by Cox&Rreid (1987), the differencial equation is hard to be handled and the solution is not feasible.

TABLE 1- Five hundred sets of simulated data for $\alpha = 0, \beta = 1, \sigma_0^2 = 1$ for different σ_{xx} and n.Comparison of confidence intervals of β for different coverages when K_{xx} is unknown.

		0.95	0.90
σ_{xx}	n	K_{xx}	K_{xx}
–	–	*unknown*	*unknown*
1.0	50	0.91	0.90
1.0	30	0.90	0.87
1.0	20	0.89	0.84
4.0	50	0.92	0.88
4.0	30	0.90	0.87
4.0	20	0.89	0.87
8.0	50	0.94	0.87
8.0	30	0.91	0.88
8.0	20	0.92	0.84

TABLE 2- Five hundred sets of simulated data for $\alpha = 0, \beta = 1$, $\sigma_0^2 = 1$ and different σ_{xx} and n. Comparison of average widths of confidence intervals with coverage 0.95 when K_{xx} is known and unknown.

	$\sigma_{xx} = 1.0$		$\sigma_{xx} = 4.0$		$\sigma_{xx} = 8.0$	
n	k_{xx}	K_{xx}	K_{xx}	K_{xx}	K_{xx}	K_{xx}
–	*unknown*	*known*	*unknown*	*known*	*unknown*	*known*
20	6.988	1.671	2.648	0.401	2.553	0.189
30	5.722	1.328	2.129	0.320	2.026	0.150
50	2.946	1.002	1.682	0.243	1.586	0.114

Table 1 shows the good performance of the coverage frequencies of the approximate intervals(K_{xx} unknown). Table 2 shows that the decrease in length of the exact intervals is considerable for $\sigma_{xx} = 8.0$. It is interesting to note that the increase in length of the confidence intervals of Wong (1989), for $\sigma_{xx} = 1$, is greater than our approximate intervals for small n. Also, Table 2 shows that the performance of the exact intervals is better than the approximate intervals, as expected.

References

[1] Barnett,V.D.(1967).A note on linear structural relationships with both residual variances known.Biometrika 54,670-2.

[2] Cox,D.R.& Reid, N.(1987).Parameter orthogonality and approximate conditional inference(with discussion).J.R.Statist.Soc. B 49,1-49

[3] Fuller,W.A.(1987). Measurement Error Models,Wiley.

[4] Kendall,M.G.& Stuart,A.(1961). The Advanced Theory of Statistics,2. London:Griffin

[5] Madansky,A.(1959).The fitting of straight line when both variables are subject to error.J.Am.Statist.Assoc. 54,173-205.

[6] Sprent,P.(1968). A generalized least-squares approach to linear functional relationships. J.R.Statist.Soc. Ser.B, 28,278-297

[7] Wong,M.Y.(1989). Likelihood estimation of a simple linear regression model when both variables have error. Biometrika,76,1,141-8.