ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY
Lönnrotinkatu 4 B, 00120 Helsinki 12, Finland, tel. 601322

Keskusteluaiheita
Discussion papers

Timo Teräsvirta

USEFULNESS OF PROXY VARIABLES IN LINEAR MODELS WITH STOCHASTIC REGRESSORS

No. 159 20 August 1984

Usefulness of proxy variables in linear models with stochastic regressors

Abstract. This paper considers a linear model with two stochastic regressors where one regressor is not observed but may be observed with error (proxy variable). The interest is focussed upon the estimation accuracy of the regression coefficient of the observable regressor. Three situations are compared with each other: the estimation of the parameters of the original model, the estimation of parameters when a proxy variable is substituted for the non-observable and the omission of the non-observable variable. Conditions for one of these three alternatives to be superior to the two others are given. Omission of the non-observable variable can generally be superior to the use of a proxy variable only in small samples.

1. Introduction

The choice between omitting a non-observable stochastic variable and substituting a proxy variable for it has been a recent topic for discussion, see Aigner (1974), Kinal and Lahiri (1983) and references therein. The proxy variable is defined as the non-observable variable measured with error. The starting-point of discussion has been a linear model with two stochastic regressors, and the interest has been focussed upon the estimation of the regression coefficient of the observable regressor. Both the omission of the non-observable variable and the proxy variable approach generally yield biased estimates for this coefficient.

These two biased estimators may be compared with each other using their mean square errors. Aigner (1974) derived a sufficient condition for the least squares estimator using a proxy variable to have a smaller MSE than the omitted variable (OV) estimator. He completed this result with a contour chart, showing the situations where the condition was satisfied. Kinal and Lahiri (1983), among other things, derived a necessary and sufficient condition for the proxy variable estimator to have smaller MSE of the two. They also briefly mentioned the possibility that the use of a proxy variable yields a lower MSE than would be the case if the non-observable variable were observed. The broad conclusion of the aforementioned authors was that the use of the proxy variable is a better alternative than the omission of the non-observable variable.

This paper elaborates the issue by including the case where the non-observable variable can be observed and investigating the situation

for all possible proxy variables simultaneously. (The necessary and sufficient condition of Kinal and Lahiri (1983) refers to a single proxy variable.) The three possibilities: the estimation of the original model (OLS estimator), the use of a proxy variable and the OV estimator are compared with each other. Four different cases seem to emerge. It is possible that the OLS estimator yields the lowest MSE and the MSE of the proxy estimator increases monotonically with the error variance of the proxy variable. The OV estimator is the worst alternative with the largest MSE. This is a common situation in practice. Another possibility is that the OV estimator is still the worst alternative but there exists a proxy variable with a smaller MSE than the OLS estimator. The third alternative is that while there exists a proxy variable with a smaller MSE than that of the OV or the OLS estimator, the OLS estimator has a larger MSE than the OV estimator. Finally, it is possible that the OLS estimator has the largest MSE, and the MSE of the proxy variable is a decreasing function of the error variance of this variable. The OV estimator has about the smallest MSE of all alternatives. The last two cases are only likely to appear in small samples.

All four possibilities can be uniquely described by the value of the first derivative of the MSE of the proxy estimator with respect to the measurement error variance of the proxy variable in the origin. While the value of the derivative in the origin does depend on the number of observations in the model, its sign does not. The sign distinguishes the first case from the three others.

In Section 2 the model is introduced and a basis for MSE comparisons established. Section 3 contains the main results and Section 4 final remarks.

2. The model

Consider the regression model

$$y = x\beta + z\gamma + u \tag{2.1}$$

where y, x, z and u are stochastic n x 1 vectors with zero expectations. The error u is independent of x and z. If z is a non-observable stochastic proxy variable $z^* = z + e$ where $Ee = 0$, $\text{cov}(e,u) = 0$. Let $\text{cov}(x,z) = \sigma_{xz}$ and $\text{Var}(a) = \sigma_{aa}$, a = e, u, x, z. Now, β may be estimated from

$$y = x\beta + z^*\gamma + u^* \tag{2.2}$$

where $u^* = u - \gamma e$, by ordinary least squares. The MSE of the least squares estimator $\hat{\beta}_E$ is (Kinal and Lahiri, 1983)

$$\text{MSE}(\hat{\beta}_E) = \gamma^2\left(\frac{\sigma_{xz}\sigma_{ee}}{\phi}\right)^2 + \frac{\sigma_{uu}}{n-3}\,\frac{\sigma_{zz}+\sigma_{ee}}{\phi}$$

$$+ \frac{\gamma^2}{n-3}\,\frac{\sigma_{ee}\sigma_{xx}\sigma_{zz.x}}{\phi}\,\frac{\sigma_{zz}+\sigma_{ee}}{\phi} \tag{2.3}$$

where $\phi = \sigma_{xx}(\sigma_{zz.x}+\sigma_{ee})$ and $\sigma_{zz.x} = \sigma_{zz} - \sigma_{xz}^2/\sigma_{xx} = \sigma_{zz}(1 - \rho_{xz}^2)$ with $\rho_{xz}^2 = \sigma_{xz}^2/\sigma_{xx}\sigma_{zz}$.

Another way of estimating β is by omitting the non-observable z altogether and estimating β from

$$y = x\beta + u^{**} \tag{2.4}$$

where $u^{**} = z\gamma + u$, by ordinary least squares. The MSE of the OV estimator $\hat{\beta}_{OV}$ equals (Kinal and Lahiri, 1983)

$$\hat{\beta}_{OV} = \frac{\sigma_{uu}}{(n-2)\sigma_{xx}} + \frac{\gamma^2\sigma_{zz.x}}{(n-2)\sigma_{xx}} + \frac{\gamma^2\sigma_{xz}^2}{\sigma_{xx}^2} . \tag{2.5}$$

Let $\sigma_{ee} \to \infty$ in (2.3). Then

$$L(\hat{\beta}_E) = \lim_{\sigma_{ee}\to\infty} MSE(\hat{\beta}) = \frac{\sigma_{uu}}{(n-3)\sigma_{xx}} + \frac{\gamma^2\sigma_{zz.x}}{(n-3)\sigma_{xx}}$$

$$+ \frac{\gamma^2\sigma_{xz}^2}{\sigma_{xx}^2} . \tag{2.6}$$

By substituting n-2 for n-3 in (2.6) we obtain (2.5). Asymptotically, as $n \to \infty$, $MSE(\hat{\beta}_{OV}) = L(\hat{\beta}_E)$ and we may regard $\hat{\beta}$, the OLS estimator of β from (2.1), and $\hat{\beta}_{OV}$ as two extreme special cases of estimator $\hat{\beta}_E$. In finite samples, this is not exactly true, as $MSE(\hat{\beta}_{OV}) - L(\hat{\beta}_E) > 0$. However, as $n \to \infty$, the difference goes to zero at the same rate as n^{-2} and is therefore often negligible already in moderate size samples.

3. Main results

Since $\hat{\beta}$ and $\hat{\beta}_{OV}$ may approximately be regarded as two extremes of $\hat{\beta}_E$, it seems useful to chart the behaviour of $MSE(\hat{\beta}_E)$ at different values of σ_{ee}. In order to do that we need

$$d_E(\sigma_{ee}) = \frac{\partial}{\partial\sigma_{ee}} MSE(\hat{\beta}_E) = \frac{2\gamma^2\sigma_{xz}^2\sigma_{ee}\sigma_{zz\cdot x}\sigma_{xx}}{\phi^3}$$

$$+ \frac{\gamma^2\sigma_{zz\cdot x}\sigma_{xx}\sigma_{zz}}{(n-3)\phi^2} - \left(\frac{\sigma_{uu}}{n-3}\frac{\sigma_{xz}^2}{\phi^2} + \frac{2\gamma^2\sigma_{xx}\sigma_{zz\cdot x}\sigma_{xz}^2\sigma_{ee}}{(n-3)\phi^3}\right). \quad (3.1)$$

In particular, note that

$$d_E(0) = \frac{\gamma^2\sigma_{zz}}{(n-3)\sigma_{xx}\sigma_{zz\cdot x}} - \frac{\sigma_{uu}\sigma_{xz}^2}{(n-3)\sigma_{xx}^2\sigma_{zz\cdot x}^2} \quad (3.2)$$

and that $d_E(\sigma_{ee}) \to 0$ as $\sigma_{ee} \to \infty$. By setting (3.1) equal to zero and solving for σ_{ee} we obtain

$$\sigma_{ee} = \frac{-\sigma_{zz\cdot x}(\gamma^2\sigma_{zz\cdot x}\sigma_{xx}\sigma_{zz} - \sigma_{uu}\sigma_{xz}^2)}{2(n-4)\gamma^2\sigma_{xz}^2\sigma_{zz\cdot x} + (\gamma^2\sigma_{zz\cdot x}\sigma_{xx}\sigma_{zz} - \sigma_{uu}\sigma_{xz}^2)}$$

$$= \frac{-\sigma_{zz\cdot x}\sigma_{xx}d_E(0)}{\frac{2(n-4)\gamma^2\rho_{xz}^2}{(n-3)(1-\rho_{xz}^2)} + \sigma_{xx}d_E(0)}. \quad (3.3)$$

Thus, as a function of σ_{ee}, $MSE(\hat{\beta}_E)$ has at most one extreme value in the interval $[0,\infty)$ cf. also Kinal and Lahiri (1983, footnote 12). A sufficient condition for no extreme value in that interval is $\sigma_{ee} < 0$. That it is necessary as well can be seen from the second derivative of the MSE which has at most one zero in $[0,\infty)$. For $n > 4$, this condition is equivalent to

$$d_E(0) > 0 \text{ or } d_E(0) < -\frac{2(n-4)\gamma^2\rho_{xz}^2}{(n-3)\sigma_{xx}(1-\rho_{xz}^2)} = -2a \quad (3.4)$$

where

$$a = \frac{(n-4)\gamma^2\rho_{xz}^2}{(n-3)\sigma_{xx}(1-\rho_{xz}^2)} .$$

If $n = 4$, (3.3) is always negative.

Let us first consider (3.4a). If $MSE(\hat{\beta}_E)$ begins to increase when σ_{ee} turns positive, then $MSE(\hat{\beta}) < MSE(\hat{\beta}_E) < MSE(\hat{\beta}_{OV})$ for any positive value of σ_{ee}. Omitting z altogether instead of using a proxy variable z* is thus the worst possible alternative, and not observing z always implies a loss in the estimation accuracy of β.

Following Kinal and Lahiri (1983), set $t_\gamma^2 = (n-3)\gamma^2\sigma_{zz.x}/\sigma_{uu}$. Then (3.4a) is equivalent to

$$t_\gamma^2 > (n-3)\rho_{xz}^2 . \tag{3.5}$$

It is seen that (3.5) is always satisfied when $\rho_{xz} = 0$ and is less likely to be valid if x and z are heavily correlated. If γ is important, i.e., the "signal-to-noise ratio" γ^2/σ_{uu} is large, (3.5) is more likely to hold than if the ratio is small. Furthermore, (3.5) is independent of the number of observations in the model.

Now, suppose (3.4b) holds. Then $MSE(\hat{\beta}) > MSE(\hat{\beta}_E) > L(\hat{\beta}_E)$ for any $\sigma_{ee} > 0$. In this case, any z* is a better alternative than z. Condition (3.4b) can also be written as

$$t_\gamma^2 < (n-3)\rho_{xz}^2 \left[1 + 2(n-4)\rho_{xz}^2\right]^{-1} . \tag{3.6}$$

For (3.6) to be valid, the sample cannot be large, the signal-to-noise ratio γ^2/σ_{uu} must be low and ρ_{xz}^2 not close to zero.

There are two intermediate situations between (3.4a) and (3.4b). First, it is possible that although (3.3) is positive, $\hat{\beta}$ may still be a worse alternative than any $\hat{\beta}_E$. This happens when $L(\hat{\beta}_E) < MSE(\hat{\beta})$. A necessary and sufficient condition for this inequality to hold is

$$d_E(0) < -a \tag{3.7}$$

or

$$t_\gamma^2 < (n-3)\rho_{xz}^2 [1 + (n-4)\rho_{xz}^2]^{-1} . \tag{3.8}$$

The r.h.s. of (3.8) is always less than one but approaches unity as $n \to \infty$. Note that $t_\gamma^2 < 1$ is a necessary and sufficient condition for $MSE(\hat{\beta}_{OV}) < MSE(\hat{\beta})$, see Kinal and Lahiri (1983). The authors (footnote 12) also give the inverse of inequality (3.6) as a necessary condition for $MSE(\hat{\beta}_E) < MSE(\hat{\beta})$ but (3.8) shows that not to be a necessary condition.

If (3.3) is positive and (3.7) holds then

$$-2a \leq d_E(0) < -a$$

and there exists an optimal proxy variable $\hat{\beta}_E^{opt}$ with $0 < \sigma_{ee} < \infty$ such that $MSE(\hat{\beta}_E^{opt}) \leq MSE(\hat{\beta}_E)$.

The second intermediate case is

$$-a \leq d_E(0) < 0 . \tag{3.9}$$

When (3.9) holds, $\hat{\beta}_E^{opt}$ again exists but at least some z* lead to a larger MSE than $\hat{\beta}$. In fact, $d_E(0) > -a$ is a sufficient condition for $\hat{\beta}_{OV}$ to be a worse alternative than the inclusion of any z*.

From (3.5) and (3.6) we may conclude that the necessary and sufficient condition for the existence of $\hat{\beta}_E^{opt}$ is

$$(n-3)\rho_{xz}^2 \left[2(n-4)\rho_{xz}^2+1\right]^{-1} \leq t_\gamma^2 \leq (n-3)\rho_{xz}^2 . \qquad (3.10)$$

Suppose $d_E(0) < 0$. It is seen from (3.10) that there always exists n_0 such that for $n \geq n_0$, $\hat{\beta}_E^{opt}$ also exists. Thus for all models such that $d_E(0) < 0$ it is always possible to improve on the OLS estimator by introducing measurement error into z.

4. Final remarks

Our conclusion is that the use of a proxy variable in multiple regression is generally advisable in large samples if the alternative is the omission of the non-observable variable. There are even situations where introducing some measurement error may be a superior strategy to the use of correctly measured variables. However, $d_E(0) \to 0$ as $n \to \infty$. Thus in large samples the possible gain from introducing measurement error (when $d_E(0) < 0$) may remain minor at best, and often only a very small measurement error can improve the estimation accuracy at all. Strong correlation between the predictors together with a small sample remains the case where it may be profitable to delete the non-observable variable altogether and not try to substitute a proxy variable for it.

References

Aigner, D.J. (1974). MSE dominance of least squares with errors-of-observation. *Journal of Econometrics* 2, 365-372.

Kinal, T. and K. Lahiri (1983). Specification error analysis with stochastic regressors. *Econometrica* 51, 1209-1218.