ON THE ESTIMATION OF PRODUCTION FUNCTIONS INVOLVING EXPLANATORY VARIABLES WHICH HAVE ZERO VALUES

George E. Battese

No. 86 - May 1996

ISSN 0 157 0188
ISBN 1 86389 330 X
On the Estimation of Production Functions Involving
Explanatory Variables Which Have Zero Values

George E. Battese*

Abstract

The estimation of regression coefficients of explanatory variables, having zero-observations, is discussed in this note. By using a dummy variable, associated with the occurrence of zero observations, the appropriate slope coefficient is correctly estimated by least-squares regression.

* Department of Econometrics, University of New England, Armidale, NSW 2351.
1. Introduction

Frequently regression analyses are used to estimate production functions involving at least one explanatory variable, for which a significant number of observations have zero values. For example, input-output data on agricultural production in developing countries frequently have zero values for synthetic fertiliser, pesticides, etc. Given that a significant proportion of the sample farmers do apply fertiliser, there is interest in estimating a suitable production function from which the elasticity of output with respect to fertiliser can be estimated. Confining one's attention to only those farmers who apply a positive amount of fertiliser may not be the most appropriate method of estimation because the data on the farmers who apply no fertiliser may be useful in the estimation of parameters which are common to all farmers.

The "zero-observation" problem in production function analyses has been addressed in a number of different ways. As stated above, some empirical economists, seeking to estimate Cobb-Douglas or translog production functions, have confined their attention to only those farmers who have positive values of the key input or explanatory variables. However, frequently the zero-observation cases are included in the analyses by substituting the value of one for zero for the farmers concerned. This may be thought to be a procedure which should not affect the estimates of the basic parameters, because "one is close to zero", at least for input variables, such as fertiliser, pesticides, etc. However, the procedure is not independent of the units of measurement of the variable concerned. In addition, if the number of "zero cases" is a significant proportion of the total number of sample observations, then the procedure may result in seriously biased estimators of key parameters of the production function.
In this note we show that the problem can be solved by the use of a dummy variable such that efficient estimators are obtained using the full data set but no bias is introduced as involved in the substitution of zeros by ones to estimate production functions when logarithms of the variables are involved.

In the next section the problem is outlined in the context of a linear model involving a single explanatory variable for which a proportion of the sample observations have zero values. This illustrates the issue involved which can be extended to a general production function involving several explanatory variables as involved in Cobb-Douglas or translog production functions.

2. A Simple Linear Model

Consider the linear model

\[ Y_i = \alpha_i + \beta_1 x_i + V_i , \quad i = 1, 2, \ldots, n_1 \]  \hspace{1cm} (1)

\[ \begin{align*}
Y_i &= \alpha_0 + V_i , \quad i = n_1 + 1, n_1 + 2, \ldots, n_1 + n_2 = n, \\
\end{align*} \hspace{1cm} (2)
\]

where \( Y_i \) is the production variable (say, the logarithm of output) for the \( i \)-th sample unit;

\( x_i \) is the value of the explanatory variable (say, the logarithm of fertiliser) for the \( i \)-th sample unit for which the variable has a non-zero value;

\( n_1 \) is the number of the sample units for which the explanatory variable has non-zero values;

\( n_2 \) is the number of the sample units for which the explanatory variable involved has zero values;\(^1\)

---

\(^1\) Given that the explanatory variable is fertilizer and \( x_i = \ln \) (Fertiliser applied), then the \( x \)-variable is not defined if fertiliser applied is zero. Hence in this case, the model, defined by equation (2), implies that the production is simply a constant parameter plus a random error.
the $V_i$s are assumed to be uncorrelated and have mean zero and variance, $\sigma^2$; and

$\alpha_1, \beta_1, \alpha_0$ and $\sigma^2$ are unknown parameters to be estimated.

This model specifies that the relationship between the production variable and the $x$-variable is given by a simple linear model. The mean production associated with the cases of the $x$-variable having zero values (or is not defined) is not assumed to be related to the parameters for those cases where the $x$-variable has non-zero values. However, the variance of the production variable is assumed to be the same value, $\sigma^2$, irrespective of whether the $x$-variable has non-zero or zero values (or is undefined). This latter feature of the model implies that the production data should be pooled for analysis.\(^2\)

The model, defined by equations (1) and (2), is expressed in terms of one equation by

$$Y_i = \alpha_1 D_{1i} + \alpha_0 D_{0i} + \beta_1 x_{i^*} + V_i, \quad i = 1, 2, \ldots, n,$$

where $D_{1i} = 1$ if $i = 1, 2, \ldots, n_1$;

$= 0, \quad$ otherwise;

$D_{0i} = 0$ if $i = 1, 2, \ldots, n_1$;

$= 1, \quad$ otherwise; and

$x_{i^*} = x_i, \quad i = 1, 2, \ldots, n_1$;

$= 0, \quad$ otherwise.

It can be shown that the ordinary least-squares estimators for the parameters, $\beta_1$, $\alpha_0$ and $\alpha_1$, are

\(^2\) In practice, other explanatory variables are involved and there are other parameters which are common to the two cases.
\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}) (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]  

(4)

\[ \hat{\alpha}_1 = \bar{Y} - \hat{\beta}_1 \bar{x} \]  

(5)

and

\[ \hat{\alpha}_0 = \bar{Y}_2 \]  

(6)

where \( \bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i} \), \( \bar{Y}_i = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i \) and \( \bar{Y}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n} Y_i \).

It is evident that the estimator for \( \beta_1 \) is equivalent to that which would be obtained if the \( Y_i \)-values corresponding to the non-zero \( x \)-values are regressed on those \( x \)-values. Further, the estimator for \( \alpha_0 \) is simply the sample mean of the \( Y_i \)-value associated with the \( x \)-variable being zero.\(^3\)

If, however, the model was expressed simply as

\[ Y_i = \alpha_i + \beta_1 x_{1i} + V_i, \quad i = 1,2,...,n, \]  

(7)

then the least-squares estimator for \( \beta_1 \) is

\[ \bar{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_{1i} - \bar{x}^*)}{\sum_{i=1}^{n} (x_{1i} - \bar{x}^*)^2} \]  

(8)

where \( \bar{Y} \) is the sample mean of the \( n \) \( Y_i \)-values and \( \bar{x}^* \) is the sample mean of the \( x_{1i} \)-values (where \( \bar{x}^* = n_1 \bar{x}_1/n \)). The above estimator is the best-linear unbiased estimator for \( \beta_1 \) if and only if \( \bar{\alpha}_0 = \alpha_1 \).

It can be shown that \( \bar{\beta}_1 \) and \( \hat{\beta}_1 \) are related according to the expression

\[^3\text{The model, defined by equation (3), is equivalently expressed by } Y_i = \alpha_i + \delta_0 D_{\alpha} + \beta_1 x_{1i}^* + V_i, \quad i = 1,2,...,n, \text{ where } \delta_0 = (\alpha_0 - \alpha_1). \text{ The least-squares estimators for the original parameters, } \alpha_1, \alpha_0 \text{ and } \beta_1, \text{ are equivalently obtained by use of this parameterisation.} \]
Whether $\tilde{\beta}_1$ is greater or less than $\hat{\beta}_1$ depends on the relative values of $Y_1$, $\bar{Y}$, $\bar{x}_1$, $n_1$, and $n_2$. The possible bias in estimation of the slope parameter is illustrated in Figure 1. In this illustration, the seven observations associated with the non-zero x-values lie around the solid line of slope, $\tilde{\beta}_1$. However, since the three Y-values associated with the "zero x-values" are above the estimated line of slope, $\hat{\beta}_1$, then the slope of the broken line of slope, $\tilde{\beta}_1$, is less than that of the solid line. The reverse situation would apply if the Y-values associated with the "zero x-values" tend to be less than the solid line of slope $\hat{\beta}_1$. These two cases clearly indicate that not using the dummy-variable approach yields biased estimation of the coefficient of the x-variable in the model.

\[
\tilde{\beta}_1 = \frac{\hat{\beta}_1 + \left\{ n_1 \bar{x}_1 (\bar{Y} - \bar{Y}) / \left( \sum_{i=1}^{2t} (x_i - \bar{x}_1)^2 \right) \right\} }{1 + \left\{ n_1 n_2 \bar{x}_1^2 / \left[ n \sum_{i=1}^{2t} (x_i - \bar{x}_1)^2 \right] \right\}}.
\]
Conclusions

This note highlights the importance of dealing with cases of zero-observations for explanatory variables in production function analyses. This approach is used in Battese, Malik and Broca (1993) and Battese, Malik and Gill (1996) in the estimation of stochastic frontier production functions for wheat farmers in selected districts in Pakistan. In some of the districts involved the proportion of sample farmers who applied no fertiliser was as high as 20 per cent. The stochastic frontier production functions estimated were modified Cobb-Douglas models involving explanatory variables such as land, labour, seed, etc., in addition to fertiliser. It is assumed that the elasticity of output with respect to land, labour, seed, etc. were the same for those farmers who do not fertilise as for those who do fertilise.

References


A Note on A Bayesian Estimator in an Autocorrelated Error Model. William Griffiths and Dan Dao, No. 3 - April 1979.


A General Equilibrium Approach to the Construction of Multilateral Index Numbers.


Bayesian Econometrics and How to Get Rid of Those Wrong Signs. William E. Griffiths, No. 31 - November 1987.


