REGRESSION WITH ONE-SIDED ERRORS IN THE DEPENDENT VARIABLE

Gregory C. Chow

Econometric Research Program
Research Memorandum No. 214

August 1977

Econometric Research Program
PRINCETON UNIVERSITY
207 Dickinson Hall
Princeton, New Jersey

Regression With One-Sided Errors in the Dependent Variable

Gregory C. Chow[1]

1. Introduction

The model studied in this paper is

$$(1.1) \qquad y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i + \eta_i \qquad (i = 1, \ldots, n)$$

where y_i and x_{ij} are respectively the i-th observations on the dependent and the explanatory variables, and ε_i are independent, identically distributed normal random variables as usual. The special feature is the introduction of the random variable η_i. It is assumed that the observations are known to belong to one of three groups. In group I, $\eta_i \geq 0$; in group II, $\eta_i \leq 0$; and in group III, $\eta_i = 0$ and we are back to the standard normal regression situation.

This model appears to have important economic applications. Generally speaking, it offers an alternative to the use of a dummy variable to represent the effect of an unusual circumstance on the dependent variable. A dummy variable for the unusual observations implies a constant effect whereas our variable η_i implies a one-sided random effect. To the extent that one can question the assumption of a constant effect in a particular application, our model may be more appropriate. For example, in a time-series analysis, the use of a constant dummy variable to represent the effect of being in a war period, of strikes, of an oil crisis and the like may be replaced by the use of a positive or negative random variable η_i. In a cross-section analysis, one may

know that certain observations of the dependent variable are either over or underestimated, but does not know the magnitudes of the errors. To use an example which has prompted this paper, a study by John Ham of Princeton University attempts to explain the supply of labor by the wage rate and other explanatory variables using data on individual households. For many observations, the respondents stated that they had wished, if given the opportunity, to work less (or more) hours than they actually did, making the observed dependent variable y_i larger (or smaller) than the desired supply $\sum_j \beta_j x_{ij} + \varepsilon_i$ by the amount η_i, according to our model. Thus in many disequilibrium situations when the desired value of an economic variable differs from the observed one because of some form of interference, our model is applicable provided that the effect of the interference is a one-sided random effect.

In section 2, we assume that the distribution of η_i is exponential and apply the method of maximum likelihood to estimate the unknown parameters. In section 3, we assume that the distribution of η_i is one-sided truncated normal and provide the solution by maximum likelihood. Section 4 deals with rectangular, triangular, and other distributions for η_i.

2. Exponential Distribution for the One-Sided Error

We assume that for group I the additional random residual η_i has an exponential density $\alpha e^{-\alpha\eta}$ for $\eta \geq 0$, that the residual ε_i has a normal distribution with mean zero and variance σ^2, and that η_i and ε_i are independent. The sum $u_i = \varepsilon_i + \eta_i$ has the cummulative distribution function

$$(2.1) \qquad F(u) = P(u_i<u) = \int_0^{\infty} \alpha e^{-\alpha\eta} \int_{-\infty}^{u-\eta} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x^2/2\sigma^2)} dx\, d\eta$$

Differentiating $F(u)$ and simplifying we obtain the density function for u_i

$$(2.2) \qquad \frac{dF}{du} = \alpha e^{\frac{1}{2}[\sigma^2\alpha^2 - 2\alpha u]} G(\sigma\alpha - \sigma^{-1}u)$$

where the function G is defined by

$$(2.3) \qquad G(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Writing (1.1) as $u_i = y_i - x_i\beta$, we find from (2.2) the log-likelihood of the i-th observation in group I to be

$$(2.4) \qquad L_1(y_i;\beta,\sigma^2,\alpha) = \ln\alpha + \frac{1}{2}\sigma^2\alpha^2 - \alpha(y_i - x_i\beta) + \ln G(\sigma\alpha - \sigma^{-1}[y_i - x_i\beta])$$

Similarly, if for group II $-\eta_i$ has an exponential distribution with parameter α_2, the log-likelihood of the i-th observation is

$$(2.5) \qquad L_2(y_i;\beta,\sigma^2,\alpha_2) = \ln\alpha_2 + \frac{1}{2}\sigma^2\alpha_2^2 + \alpha(y_i - x_i\beta) + \ln G(\sigma\alpha_2 + \sigma^{-1}[y_i - x_i\beta])$$

For an observation in group III, $\eta_i = 0$, and the log-likelihood is standard.

Let the 3 groups consist of n_1, n_2 and n_3 observations respectively, with $n = n_1 + n_2 + n_3$. The log-likelihood function for the entire sample is

$$(2.6) \qquad L = n_1\ln\alpha + \frac{1}{2}n_1\sigma^2\alpha^2 - \alpha \sum_{i\in I}(y_i - x_i\beta) + \sum_{i\in I}\ln G(\sigma\alpha - \sigma^{-1}[y_i - x_i\beta])$$

$$+ n_2\ln\alpha_2 + \frac{1}{2}n_2\sigma^2\alpha_2^2 + \alpha_2 \sum_{i\in II}(y_i - x_i\beta) + \sum_{i\in II}\ln G(\sigma\alpha_2 + \sigma^{-1}[y_i - x_i\beta])$$

$$- \frac{1}{2}n_3\ln(2\pi) - \frac{1}{2}n_3\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i\in III}(y_i - x_i\beta)^2$$

To compute maximum likelihood estimates of β, σ, α and α_2 we recommend Newton's method since it has been found successful in many similar applications and since the matrix of the second partial derivatives of L is required in any case to approximate the covariance matrix of our estimates for the purpose of statistical inference. To do so, we obtain the first derivatives of L. Using the abbreviations

(2.7)
$$u_i = y_i - x_i\beta$$

$$G_i = G(\sigma\alpha - \sigma^{-1}u_i), \qquad f_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sigma\alpha - \sigma^{-1}u_i)^2}$$

$$G_{2i} = G(\sigma\alpha_2 + \sigma^{-1}u_i), \qquad f_{2i} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sigma\alpha_2 + \sigma^{-1}u_i)^2}$$

we can write the first derivatives as

(2.8)
$$\frac{\partial L}{\partial \beta} = \alpha \sum_{I} x_i' - \sigma^{-1} \sum_{I} G_i^{-1} f_i x_i' - \alpha_2 \sum_{II} x_i' + \sigma^{-1} \sum_{II} G_{2i}^{-1} f_{2i} x_i' + \sigma^{-2} \sum_{III} u_i x_i'$$

(2.9)
$$\frac{\partial L}{\partial \sigma} = n_1\sigma\alpha^2 - \sum_{I} G_i^{-1} f_i(\alpha + \sigma^{-2}u_i) + n_2\sigma\alpha_2^2 - \sum_{II} G_{2i}^{-1} f_{2i}(\alpha_2 - \sigma^{-2}u_i) - n_3\sigma^{-1} + \sigma^{-3} \sum_{III} u_i^2$$

(2.10)
$$\frac{\partial L}{\partial \alpha} = n_1(\sigma^2\alpha + \alpha^{-1}) - \sum_{I} u_i - \sigma \sum_{I} G_i^{-1} f_i$$

(2.11)
$$\frac{\partial L}{\partial \alpha_2} = n_2(\sigma^2\alpha_2 + \alpha_2^{-1}) + \sum_{II} u_i - \sigma \sum_{II} G_{2i}^{-1} f_{2i}$$

Although analytical expressions for the second derivatives of L can be obtained by differentiation, they are rather tedious to write down and to program in the computer. We therefore recommend the use of numerical second derivatives which

can be calculated as the rates of change of the above analytical first derivatives with respect to small changes in the parameter values.

Writing $\theta' = (\beta', \sigma, \alpha, \alpha_2)$, we apply Newton's iteration formula

$$\theta^{(r+1)} = \theta^{(r)} - \left(\frac{\partial^2 L}{\partial\theta\partial\theta'} \right)^{-1} \left(\frac{\partial L}{\partial\theta} \right) \tag{2.12}$$

where the superscript denotes the iteration number and the derivatives are evaluated at $\theta^{(r)}$. For initial values of β and σ^2 , we may use the least squares estimates obtained from the observations in group III only. Given these initial values, we can set (2.10) and (2.11) equal to zero and solve for α and α_2 respectively by any univariate iterative method, including Newton's method where the required second derivatives can again be obtained numerically. The results can serve as initial values for α and α_2. As the sample sizes n_1 , n_2 , and n_3 increase, the sequencies of these initial estimators are consistent because they are obtained by the method of maximum likelihood. One then has the option of using the solution of only one iteration of (2.12) as a set of linearized maximum likelihood estimates, rather than iterating until (2.12) converges. An asymptotic covariance matrix of the parameter estimates is given by $-\frac{\partial^2 L}{\partial\theta\partial\theta'}$ as usual.

3. Half Truncated Normal Distribution for the One-Sided Error

We now assume that for group I the density of η_i is twice the density of a normal random variable with mean zero and variance α for $\eta_i > 0$ and is zero otherwise, and that η_i and ε_i are independent. The sum $u_i = \varepsilon_i + \eta_i$ has the cummulative distribution function, with σ^2 written as v ,

$$F(u) = P(u_i < u) = \int_0^{\infty} \frac{2}{\sqrt{2\pi\alpha}} e^{-(\eta^2/2\alpha)} \int_{-\infty}^{u-\eta} \frac{1}{\sqrt{2\pi v}} e^{-(x^2/2v)} dx\, d\eta \tag{3.1}$$

The density function for u_i is

$$(3.2) \qquad \frac{dF}{du} = (2/\pi)^{.5}(v+\alpha)^{-.5}e^{-u^2/2(v+\alpha)}G(-\alpha^{.5}[v^2+v\alpha]^{-.5}u)$$

where the function G has been defined by (2.3)

If for group II, the distribution of $-\eta_i$ is half-truncated normal with parameter α_2, and if the numbers of observations in the three groups are again n_1, n_2 and n_3 respectively, the log-likelihood function will become, with u_i again denoting $y_i - x_i\beta_i$,

$$(3.3) \qquad L = \text{const.} - .5n_1\ln(v+\alpha) - .5(v+\alpha)^{-1}\sum_I u_i^2 + \sum_I \ln G(-\alpha^{.5}[v^2+v\alpha]^{-.5}u_i)$$

$$- .5n_2\ln(v+\alpha_2) - .5(v+\alpha_2)^{-1}\sum_{II} u_i^2 + \sum_{II} \ln G(\alpha_2^{.5}[v^2+v\alpha_2]^{-.5}u_i)$$

$$- .5n_3\ln v - .5v^{-1}\sum_{III} u_i^2$$

Using the abbreviations

$$(3.4) \qquad G_i = G(-\alpha^{.5}[v^2+v\alpha]^{-.5}u_i) \quad , \quad f_i = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\alpha[v^2+v\alpha]^{-1}u_i^2}$$

$$G_{2i} = G(\alpha_2^{.5}[v^2+v\alpha_2]^{-.5}u_i) \quad , \quad f_{2i} = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\alpha_2[v^2+v\alpha_2]^{-1}u_i^2}$$

we can write the first derivatives of L as

$$(3.5) \qquad \frac{\partial L}{\partial \beta} = (v+\alpha)^{-1}\sum_I u_i x_i' - \alpha^{.5}[v^2+v\alpha]^{-.5}\sum_I G_i^{-1}f_i x_i' + (v+\alpha_2)^{-1}\sum_{II} u_i x_i'$$

$$+ \alpha_2^{.5}[v^2+v\alpha_2]^{-.5}\sum_{II} G_{2i}^{-1}f_{2i}x_i' + v^{-1}\sum_{III} u_i x_i'$$

$$(3.6)\qquad \frac{\partial L}{\partial v} = -.5n_1(v+\alpha)^{-1} + .5(v+\alpha)^{-2}\sum_{I} u_i^2 - .5\alpha^{.5}(v^2+v\alpha)^{-1.5}(2v+\alpha)\sum_{I} G_i^{-1}f_i u_i$$

$$- .5n_2(v+\alpha_2)^{-1} + .5(v+\alpha_2)^{-2}\sum_{II} u_i^2$$

$$+ .5\alpha_2^{.5}(v^2+v\alpha_2)^{-1.5}(2v+\alpha_2)\sum_{II} G_{2i}^{-1}f_{2i}u_i - .5n_3v^{-1} + .5v^{-2}\sum_{III} u_i^2$$

$$(3.7)\qquad \frac{\partial L}{\partial \alpha} = -.5n_1(v+\alpha)^{-1} + .5(v+\alpha)^{-2}\sum_{I} u_i^2 + .5v^2(v^2+v\alpha)^{-1.5}\alpha^{-.5}\sum_{I} G_i^{-1}f_i u_i$$

$$(3.8)\qquad \frac{\partial L}{\partial \alpha_2} = -.5n_2(v+\alpha_2)^{-1} + .5(v+\alpha_2)^{-2}\sum_{II} u_i^2 - .5v^2(v^2+v\alpha_2)^{-1.5}\alpha_2^{-.5}\sum_{II} G_{2i}^{-1}f_{2i}u_i$$

Again we recommend using numerical second derivatives together with the above analytical first derivatives of L and applying Newton's method as suggested at the end of section 2.

4. Other Distributions for the One-Sided Error

Three other obvious candidates for the distribution of η_i having only one parameter are the rectangular distribution, with density $1/\alpha$ for $(0 < \eta < \alpha)$, the symmetrical triangular distribution, and the one-sided triangular distribution, with density $(2/\alpha) - (2\eta/\alpha^2)$ for $(0 < \eta < \alpha)$. A fourth possibility is to truncate a normal distribution with mean zero and variance α not at the mid-point zero but at a γ percentage point. If γ is greater than 50 per cent (for the right-tail probability), the effect of the random one-sided error will have higher densities at some positive values than at zero. For these distributions, and assuming independence of ε_i and η_i, one can work out the required likelihood functions without difficulty.

One simple, and possibly useful, extension of the methods of this paper is to allow for both random and non-random differences between the three groups specified in section 1. The random differences are captured by η_i. The non-random differences may be accounted for by introducing additional dummy and/or continuous variables and allowing for differences in subsets of regression coefficients for the different groups, as is customarily done in regression analysis.

Footnote

1. I would like to thank Roger Gordon and John Ham for useful discussions and to acknowledge financial support from the National Science Foundation.