

# The Tobit model

Herman J. Bierens

September 17, 2004

## 1 The model

The Tobit<sup>1</sup> model assumes that the observed dependent variables  $Y_j$  for observations  $j = 1, \dots, n$  satisfy

$$Y_j = \max(Y_j^*, 0), \quad (1)$$

where the  $Y_j^*$ 's are latent variables generated by the classical linear regression model

$$Y_j^* = \beta' X_j + U_j, \quad (2)$$

with  $X_j$  a vector of regressors, possibly including 1 for the intercept, and  $\beta$  the corresponding vector of parameters. The model errors  $U_j$  are assumed to be independent  $N(0, \sigma^2)$  distributed, conditional on the  $X_j$ 's.

Denoting by

$$f(z) = \exp(-z^2/2)/\sqrt{2\pi}$$

the density of the  $N(0, 1)$  distribution, with corresponding cumulative distribution function (c.d.f.)

$$F(z) = \int_{-\infty}^z f(v)dv,$$

---

<sup>1</sup>The model is called *Tobit* because it was first proposed by *Tobin* (1958), and involves aspects of *Probit* analysis. See:

Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables", *Econometrica* 26, 24-36.

the conditional c.d.f. of  $Y_j$  given  $Y_j > 0$  and  $X_j$  is

$$\begin{aligned} H(y|Y_j > 0, X_j, \beta, \sigma) &= P(y_j \leq y \mid Y_j > 0, X_j) \\ &= \frac{P(0 < Y_j^* \leq y \mid x_j)}{P(Y_j^* > 0 \mid X_j)} = \frac{P(-\beta' X_j < U_j \leq y - \beta' X_j \mid X_j)}{P(U_j > -\beta' X_j \mid X_j)} \\ &= \frac{F((y - \beta' X_j)/\sigma) - F(-\beta' X_j/\sigma)}{F(\beta' X_j/\sigma)}, \end{aligned}$$

and the corresponding conditional density is

$$h(y|Y_j > 0, X_j, \beta, \sigma) = \frac{dH(y \mid Y_j > 0, X_j, \beta, \sigma)}{dy} = \frac{f((y - \beta' X_j)/\sigma)}{\sigma F(\beta' X_j/\sigma)}, \quad y > 0.$$

Thus, the conditional distribution of  $Y_j$  given  $Y_j > 0$  and  $X_j$  is **continuous**.

Define the dummy variable  $D_j$  by

$$\begin{aligned} D_j &= 1 \text{ if } Y_j > 0, \\ D_j &= 0 \text{ if } Y_j = 0. \end{aligned}$$

Then

$$\begin{aligned} P[D_j = 1 \mid X_j] &= F(\beta' X_j/\sigma), \\ P[D_j = 0 \mid X_j] &= 1 - F(\beta' X_j/\sigma), \end{aligned}$$

and

$$Y_j = D_j Y_j^*.$$

## 2 Truncation bias

Now the conditional expectation of  $Y_j$  given  $X_j$  and  $D_j = 1$  is

$$\begin{aligned} &E[Y_j \mid X_j, D_j = 1] \\ &= \int_0^\infty y h(y \mid Y_j > 0, X_j, \beta, \sigma) dy \\ &= \frac{1}{\sigma F(\beta' X_j/\sigma)} \int_0^\infty y f((y - \beta' X_j)/\sigma) dy \\ &= \frac{1}{F(\beta' X_j/\sigma)} \int_{-\beta' X_j/\sigma}^\infty (\beta' X_j + \sigma z) f(z) dz \end{aligned}$$

$$\begin{aligned}
&= \frac{(\beta' X_j) \int_{-\beta' X_j/\sigma}^{\infty} f(z) dz + \sigma \int_{-\beta' X_j/\sigma}^{\infty} z f(z) dz}{F(\beta' X_j/\sigma)} \\
&= \frac{(\beta' X_j) F(\beta' X_j) - \sigma \int_{-\beta' X_j/\sigma}^{\infty} f'(z) dz}{F(\beta' X_j/\sigma)} \\
&= \beta' X_j + \sigma \frac{f(\beta' X_j/\sigma)}{F(\beta' X_j/\sigma)}.
\end{aligned}$$

Therefore, if you regress only the positive  $Y_j$ 's on the corresponding  $X_j$ 's then, due to the latter term, the OLS parameters estimate of  $\beta$  will be biased and inconsistent.

Moreover, note that

$$E[Y_j|X_j] = (\beta' X_j) F(\beta' X_j/\sigma) + \sigma f(\beta' X_j/\sigma). \quad (3)$$

Therefore, if you treat the zero values of  $Y_j$  as regular dependent variable values in a linear regression model the OLS parameters estimate of  $\beta$  will be biased and inconsistent as well.

### 3 The log-likelihood

The conditional c.d.f. of  $Y_j$  given  $X_j$  is now

$$\begin{aligned}
G(y|X_j, \beta, \sigma) &= P[Y_j \leq y | X_j] = P[Y_j \leq y | X_j, D_j = 1] P[D_j = 1 | X_j] \\
&\quad + P[Y_j \leq y | X_j, D_j = 0] P[D_j = 0 | X_j] \\
&= I(y > 0) H(y | Y_j > 0, X_j, \beta, \sigma) F(\beta' X_j/\sigma) \\
&\quad + I(y = 0) (1 - F(\beta' X_j/\sigma)),
\end{aligned}$$

where  $I(\cdot)$  is the indicator function:  $I(true) = 1$ ,  $I(false) = 0$ . Hence, the corresponding conditional "density" is

$$\begin{aligned}
g(y|X_j, \beta, \sigma) &= I(y > 0) h(y | Y_j > 0, X_j, \beta, \sigma) F(\beta' X_j/\sigma) \\
&\quad + I(y = 0) (1 - F(\beta' X_j/\sigma))
\end{aligned}$$

Therefore, the log-likelihood function of the Tobit model is

$$\mathcal{L}(\beta, \sigma) = \sum_{j=1}^n \ln [g(Y_j|X_j, \beta, \sigma)]$$

$$\begin{aligned}
&= \sum_{j=1}^n D_j \ln [h(Y_j | Y_j > 0, X_j, \beta, \sigma)] \\
&\quad + \sum_{j=1}^n D_j \ln F(\beta' X_j / \sigma) + \sum_{j=1}^n (1 - D_j) \ln (1 - F(\beta' X_j / \sigma)) \\
&= \sum_{j=1}^n D_j \left( -\frac{1}{2} (Y_j - \beta' X_j)^2 / \sigma^2 - \ln(\sigma) \right) \\
&\quad + \sum_{j=1}^n (1 - D_j) \ln (1 - F(\beta' X_j / \sigma)) - \sum_{j=1}^n D_j \ln(\sqrt{2\pi})
\end{aligned}$$

However, before maximizing this likelihood function it is convenient to reparametrize it by replacing  $\beta$  by  $\sigma\gamma$  and  $\sigma$  by  $1/\theta$ , so that

$$\begin{aligned}
\mathcal{L}^*(\gamma, \theta) &= \mathcal{L}(\sigma\gamma, 1/\sigma) \\
&= -\frac{1}{2} \sum_{j=1}^n D_j \left( (\theta Y_j - \gamma' X_j)^2 + \ln(\theta^2) \right) \\
&\quad + \sum_{j=1}^n (1 - D_j) \ln (1 - F(\gamma' X_j)) - \sum_{j=1}^n D_j \ln(\sqrt{2\pi})
\end{aligned}$$

Moreover, it can be shown<sup>2</sup> (but this is nontrivial) that the Hessian matrix

$$\frac{\partial^2 \mathcal{L}^*(\gamma, \theta)}{\partial(\gamma', \theta)' \partial(\gamma', \theta)}$$

is negative definite for all values of  $\gamma$  and  $\theta > 0$ , so that the log-likelihood  $\mathcal{L}^*(\gamma, \theta)$  has no local maxima. This is the main reason for the reparametrization involved.

## 4 The pseudo $R^2$

EasyReg computes the (pseudo)  $R^2$  of the model as follows. Given the maximum likelihood estimators  $\hat{\beta}$  and  $\hat{\sigma}$ , the "residuals" are computed as

$$\hat{U}_j = Y_j - (\hat{\beta}' X_j) F(\hat{\beta}' X_j / \hat{\sigma}) - \hat{\sigma} f(\hat{\beta}' X_j / \hat{\sigma})$$

---

<sup>2</sup>Reference:

Olsen, R. (1978), "A Note on the Uniqueness of the Maximum Likelihood Estimator in the Tobit Model", *Econometrica* 46, 1211-1215.

[see (3)], and then the  $R^2$  is computed in the same way as for OLS:

$$R^2 = 1 - \frac{\sum_{j=1}^n \hat{U}_j^2}{\sum_{j=1}^n (Y_j - \bar{Y})^2},$$

where  $\bar{Y}$  is the sample mean of the  $Y_j$ 's.

## 5 When does EasyReg refuse to conduct Tobit analysis?

There are three cases for which EasyReg does not allow you to conduct Tobit analysis:

1. If the dependent variables  $Y_j$  take negative values.
2. If the dependent variables  $Y_j$  take only positive values.
3. If the dependent variables  $Y_j$  are non-negative, with some of the  $Y_j$ 's zero, but all the  $Y_j$ 's integer valued.

In the first two cases the reason for refusing to conduct Tobit analysis is obvious, but I have gotten quite a few queries from EasyReg users why in the last case *EasyReg* refuses to continue.

In the third case the  $Y_j$ 's satisfy

$$P[Y_j \in \{0, 1, 2, \dots\}] = 1,$$

and therefore the conditional distribution of  $Y_j$  given  $Y_j > 0$  and  $X_j$  is **discrete**, i.e., the conditional c.d.f.  $H(y|Y_j > 0, X_j, \beta, \sigma)$  in this case is a **step function**, with jumps at some integer values of  $y$ . This violates the basic assumption of the Tobit model, i.e., (1), (2) and the normality of the  $U_j$ 's, and therefore *EasyReg* will not allow you to conduct Tobit analysis.

However, if you insist on conducting Tobit analysis with this dependent variable there is a trick to fool *EasyReg*: Multiply the  $Y_j$ 's by a factor  $10^{-m}$ , where  $m > 0$  is such that at least one of the new variables  $10^{-m}Y_j$  has one or more decimal digits.<sup>3</sup> For example, suppose that the original  $Y_j$ 's are dollar amounts, rounded off to multiples of 1000 dollar, and suppose that there exists at least one  $Y_j$  value with a non-zero digit next to the last three zeros, say

---

<sup>3</sup>This transformation can be done in *EasyReg*: Click Menu  $\rightarrow$  Input  $\rightarrow$  Transform variables  $\rightarrow$  Linear combination of variables, double click the variable involved, click "Selection OK", enter the coefficient involved, and click "OK".

$Y_j = 123000$ . Then  $10^{-4}Y_j = 0.0001 \times 123000 = 12.3$ , hence at least one of the new variables  $10^{-4}Y_j$  has a decimal digit. Since *EasyReg* validates your data by scanning for decimal digits in the values of the dependent variable, it will allow you to conduct Tobit analysis for the rescaled dependent variable involved.