

11

Limited dependent variable regression models

In the logit and probit models we discussed previously the dependent variable assumed values of 0 and 1, 0 representing the absence of an attribute and 1 representing the presence of that attribute, such as smoking or not smoking, or owning a house or not owning one, or belonging or not belonging to a union. As noted, the logit model uses the logistic probability distribution and the probit the normal distribution. We saw in Chapter 8 how one estimates and interprets such models, using the example of cigarette smoking.

But now consider this problem: how many packs of cigarettes does a person smoke, given his or her socio-economic variables? Now this question is meaningful only if a person smokes; a nonsmoker may have no interest in this question. In our smoker example discussed in Chapter 8 we had a sample of 1,196 people, of which about 38% smoked and 62% did not smoke. Therefore we can obtain information about the number of packs smoked for only about 38% of the people in the sample.

Suppose we only consider the sample of smokers and try to estimate a demand function for the number of packs of cigarettes smoked per day based on socio-economic information of the smokers only. How reliable will this demand function be if we omit 62% of the people in our sample of 1,196? As you might suspect, such a demand function may not be reliable.

The problem here is that we have a **censored sample**, a sample in which information on the regressand is available only for some observations but not all, although we may have information on the regressors for all the units in the sample. It may be noted that the regressand can be **left-censored** (i.e. it cannot take a value below a certain threshold, typically, but not always, zero) or it may be **right-censored** (i.e. it cannot take a value above a certain threshold, say, people making more than one million dollars of income), or it can be both left- and right-censored.

A closely related but somewhat different model from the censored sample model is the **truncated sample model**, in which information on both the regressand and regressors is not available on some observations. This could be by design, as in the New Jersey negative income tax experiment where data for those with income higher than 1.5 times the 1967 poverty line income were not included in the sample.¹

Like the censored sample, the truncated sample can be left-censored, right-censored or both right- and left-censored.

¹ See J. A. Hausman and D. A. Wise, *Social Experimentation*, NBER Economic Research Conference Report, University of Chicago Press, Chicago, 1985.

How then do we estimate such models, which are also known as **limited dependent variable regression models** because of the restriction put on the values taken by the regressand? Initially we will discuss the censored regression model and then discuss briefly the truncated regression model. As in the various models in this book, our emphasis will be on practical applications.

11.1 Censored regression models

A popularly used model in these situations is the **Tobit model**, which was originally developed by James Tobin, a Nobel laureate economist.² Before we discuss the Tobit model, let us first discuss OLS (ordinary least squares) applied to a censored sample. See Table 11.1, available on the companion website.

OLS estimation of censored data

For this purpose we use the data collected by Mroz.³ His sample gives data on 753 married women, 428 of whom worked outside the home and 325 of whom did not work outside the home, and hence had zero hours of work.

Some of the socio-economic variables affecting the work decision considered by Mroz are age, education, experience, squared experience, family income, number of kids under age 6, and husband's wage. Table 11.1 gives data on other variables considered by Mroz.

Applying OLS to hours of work in relation to the socio-economic variables for all the observations, we obtained the results in Table 11.2.

The results in this table are to be interpreted in the framework of the standard linear regression model. As you know, in the linear regression model each slope coefficient gives *the marginal effect* of that variable on the *mean* or average value of the dependent variable, holding all other variables in the model constant. For example, if husband's wages go up by a dollar, the average hours worked by married women declines by about 71 hours, *ceteris paribus*. Except for education, all the other coefficients seem to be highly statistically significant. But beware of these results, for in our sample 325 married women had zero hours of work.

Suppose, instead of using all observations in the sample, we only use the data for 428 women who worked. The OLS results based on this (censored) sample are given in Table 11.3.

If you compare the results in Tables 11.2 and 11.3, you will see some of the obvious difference between the two.⁴ The education variable now seems to be highly significant, although it has a negative sign. But we should be wary about these results also.

This is because OLS estimates of censored regression models, whether we include the whole sample (Figure 11.1) or a subset of the sample (Figure 11.2), are *biased* as

² James Tobin (1958) Estimation of Relationship for Limited Dependent Variables. *Econometrica*, vol. 26, pp. 24–36.

³ See T. A. Mroz, (1987) The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, vol. 55, pp. 765–99. Recall that we used these data in Chapter 4 while discussing multicollinearity.

⁴ In the traditional regression model the mean value of the error term u_i is assumed to be zero, but there is no guarantee that this will be the case if we only use a subset of the sample values, as in this example.

Table 11.2 OLS estimation of the hours worked function.

Dependent Variable: HOURS
 Method: Least Squares
 Sample: 1 753
 Included observations: 753

	Coefficient	Std. Error	t-Statistic	Prob.
C	1298.293	231.9451	5.597413	0.0000
AGE	-29.55452	3.864413	-7.647869	0.0000
EDUC	5.064135	12.55700	0.403292	0.6868
EXPER	68.52186	9.398942	7.290380	0.0000
EXPERSQ	-0.779211	0.308540	-2.525480	0.0118
FAMINC	0.028993	0.003201	9.056627	0.0000
KIDSLT6	-395.5547	55.63591	-7.109701	0.0000
HUSWAGE	-70.51493	9.024624	-7.813615	0.0000

R-squared 0.338537 Mean dependent var 740.5764
 Adjusted R-squared 0.332322 S.D. dependent var 871.3142
 S.E. of regression 711.9647 Akaike info criterion 15.98450
 Sum squared resid 3.78E+08 Schwarz criterion 16.03363
 Log likelihood -6010.165 Hannan-Quinn criter. 16.00343
 F-statistic 54.47011 Durbin-Watson stat 1.482101
 Prob(F-statistic) 0.000000

III

Table 11.3 OLS estimation of hours function for working women only.

Dependent Variable: HOURS
 Method: Least Squares
 Sample: 1 428
 Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	1817.334	296.4489	6.130345	0.0000
AGE	-16.45594	5.365311	-3.067100	0.0023
EDUC	-38.36287	16.06725	-2.387644	0.0174
EXPER	49.48693	13.73426	3.603174	0.0004
EXPERSQ	-0.551013	0.416918	-1.321634	0.1870
FAMINC	0.027386	0.003995	6.855281	0.0000
KIDSLT6	-243.8313	92.15717	-2.645821	0.0085
HUSWAGE	-66.50515	12.84196	-5.178739	0.0000

R-squared 0.218815 Mean dependent var 1302.930
 Adjusted R-squared 0.205795 S.D. dependent var 776.2744
 S.E. of regression 691.8015 Akaike info criterion 15.93499
 Sum squared resid 2.01E+08 Schwarz criterion 16.01086
 Log likelihood -3402.088 Hannan-Quinn criter. 15.96495
 F-statistic 16.80640 Durbin-Watson stat 2.107803
 Prob(F-statistic) 0.000000

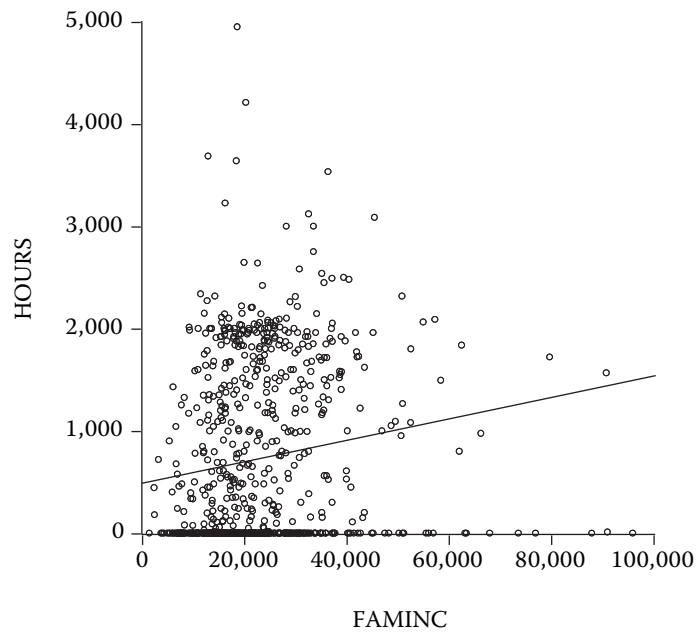


Figure 11.1 Hours worked and family income, full sample.

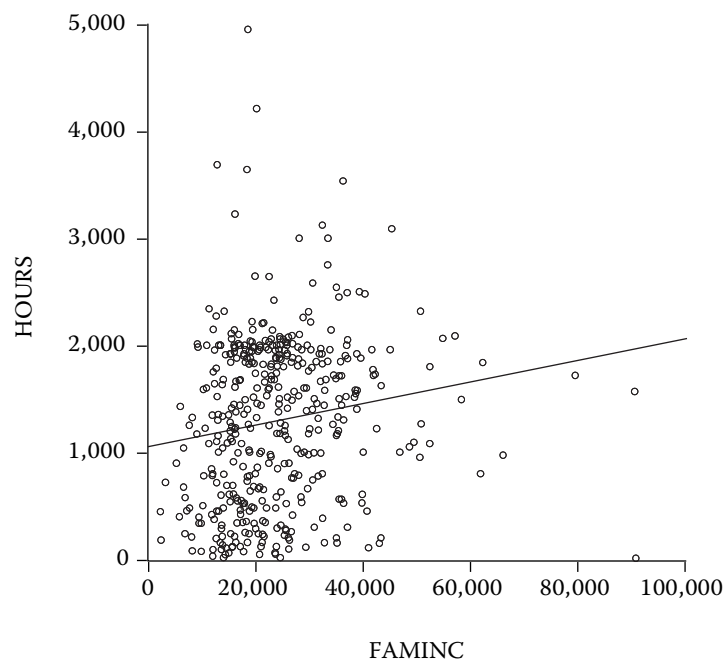


Figure 11.2 Hours vs. family income for working women.

well as *inconsistent* – that is, no matter how large the sample size is, the estimated parameters will not converge to their true values.⁵ The reason for this is the fact that in censored as well as truncated regression models the conditional mean of the error term, u_t , is nonzero and the error is correlated with the regressors. As we know, if the error term and the regressors are correlated, the OLS estimators are biased as well as inconsistent.

To give a glimpse of why the OLS estimates may be biased as well as inconsistent, we plot hours worked against family income in Figure 11.1 and hours worked and family income only for working women in Figure 11.2.

In Figure 11.1 there are several observations (actually 325) that lie on the horizontal axis because for these observations the hours worked are zero.

In Figure 11.2, none of the observations lie on the horizontal axis, for these observations are for 428 working women. The slope coefficients of the regression lines in the two figures will obviously be different.

A popularly used model to deal with censored samples is the Tobit model, which we now discuss.

11.2 Maximum likelihood (ML) estimation of the censored regression model: the Tobit model

III

One of the popularly used censored sample regression model is the Tobit model. There are several variants of the Tobit model, but we consider here the simplest one, the so-called standard Tobit model.⁶ We will continue with the Mroz data.

To see how the censored observations are dealt with, we proceed as follows: Let

$$Y_i^* = B_1 + B_2 \text{Age}_i + B_3 \text{Edu}_i + B_4 \text{Exp}_i + B_5 \text{Kids6} + B_6 \text{Faminc} + B_7 \text{Huswage} + u_i \quad (11.1)$$

where Y_i^* are *desired* hours of work. Now

$$Y_i = 0 \quad \text{if } Y_i^* \leq 0 \\ = Y_i^* \quad \text{if } Y_i^* > 0 \quad (11.2)$$

where $u_i \sim N(0, \sigma^2)$ and where Y_i are the realized or actual hours worked.⁷ The regressors are, respectively, age in years, education in years of schooling, work experience in years, number of kids under age 6, family income in thousands of dollars, and husband's hourly wage.

The variable Y_i^* is called a **latent variable**, the variable of primary interest. Of course, we do not actually observe this variable for all the observations. We only

⁵ For a rigorous proof, see Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, South-Western, USA, 4th edn, 2006, Ch. 17. See also Christiaan Heij, Paul de Boer, Philip Hans Franses, Teun Kloek, and Herman K. van Dijk, *Econometric Methods with Applications in Business and Economics*, Oxford University Press, Oxford, UK, 2004, Chapter 6.

⁶ A detailed, but somewhat advanced discussion can be found in A. Colin Cameron and Pravin K. Trivedi, *Microeconometrics: Methods and Applications*, Cambridge University Press, New York, 2005, Chapter 16.

⁷ One can use the logistic or the extreme value probability distribution in lieu of the normal distribution.

observe it for those observations with positive hours of work because of censoring. Recall that we discussed the concept of latent variables in the previous chapter.⁸

Notice that we are assuming that the error term is normally distributed with zero mean and constant (or homoscedastic) variance. We will have more to say about this assumption later.

Before we proceed further, it is useful to note the difference between the probit model and the Tobit model. In the probit model, $Y_i = 1$ if Y_i^* is greater than zero, and it is equal to zero if the latent variable is zero. In the Tobit model Y_i may take any value as long as the latent variable is greater than zero. That is why the Tobit model is also known as Tobin's probit.

To estimate a model where some observations on the regressand are censored (because they are not observed), the Tobit model uses the method of **maximum likelihood (ML)**, which we have encountered on several occasions.⁹ The actual mechanics of Tobit ML method is rather complicated, but *Stata*, *Eviews* and other software packages can estimate this model very easily.¹⁰

Using *Eviews6* we obtained the results in Table 11.4 for our example.

Interpretation of the Tobit estimates

How do we interpret these results? If you only consider the signs of the various regressors, you will see that they are the same in Tables 11.2 and 11.3. And qualitatively they make sense. For example, if the husband's wages go up, on average, a woman will work less in the labor market, *ceteris paribus*. The education variable is not significant in Table 11.2, but it is in Table 11.3, although it has a negative sign. In Table 11.4 it is significant and has a positive sign, which makes sense.

The slope coefficients of the various variables in Table 11.4 give the marginal impact of that variable on the *mean value of the latent variable*, Y_i^* , but in practice we are interested in the marginal impact of a regressor on the mean value of Y_i , the actual values observed in the sample.

Unfortunately, unlike the OLS estimates in Table 11.2, we *cannot interpret the Tobit coefficient of a regressor as giving the marginal impact of that regressor on the mean value of the observed regressand*. This is because in the Tobit type censored regression models a unit change in the value of a regressor has two effects: (1) the effect on the mean value of the observed regressand, and (2) the effect on the probability that Y_i^* is actually observed.¹¹

Take for instance the impact of age. The coefficient for age of about -54 in Table 11.4 means that, holding other variables constant, if age increases by a year, its direct impact on the hours worked per year will be a decrease by about 54 hours per year and the probability of a married woman entering the labor force will also decrease. So we have to multiply -54 by the probability that this will happen. Unless we know the latter, we will not be able to compute the aggregate impact of an increase in age on the

⁸ In the present context we can interpret the latent variable as a married woman's propensity or desire to work.

⁹ There are alternatives to ML estimation, some of which may be found in the book by Greene, *op cit*.

¹⁰ The details of Tobin's ML method can be found in Christiaan Heij, *op cit*.

¹¹ That is, $\partial[Y | X_i] / \partial X_i = B_{ix} \Pr(0 < Y_i^* < \infty)$ and the latter probability depends on all the regressors in the model and their coefficients.

Table 11.4 ML estimation of the censored regression model.

Dependent Variable: HOURS
 Method: ML - Censored Normal (TOBIT) (Quadratic hill climbing)
 Sample: 1 753
 Included observations: 753
 Left censoring (value) at zero
 Convergence achieved after 6 iterations
 Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	1126.335	379.5852	2.967279	0.0030
AGE	-54.10976	6.621301	-8.172074	0.0000
EDUC	38.64634	20.68458	1.868365	0.0617
EXPER	129.8273	16.22972	7.999356	0.0000
EXPERSQ	-1.844762	0.509684	-3.619422	0.0003
FAMINC	0.040769	0.005258	7.754009	0.0000
KIDSLT6	-782.3734	103.7509	-7.540886	0.0000
HUSWAGE	-105.5097	15.62926	-6.750783	0.0000

Error Distribution
 SCALE:C(9) 1057.598 39.06065 27.07579 0.0000
 Mean dependent var 740.5764 S.D. dependent var 871.3142
 S.E. of regression 707.2850 Akaike info criterion 10.08993
 Sum squared resid 3.72E+08 Schwarz criterion 10.14520
 Log likelihood -3789.858
 Avg. log likelihood -5.033012
 Left censored obs 325 Right censored obs 0
 Uncensored obs 428 Total obs 753

Note: The scale factor is the estimated scale factor σ , which may be used to estimate the standard deviation of the residual, using the known variance of the assumed distribution, which is 1 for the normal distribution, $\pi^2 / 3$ for the logistic distribution and $\pi^2 / 6$ for the extreme value (Type I) distribution.

III

hours worked. And this probability calculation depends on all the regressors in the model and their coefficients.

Interestingly, the slope coefficient gives directly the marginal impact of a regressor on the latent variable, Y_i^* , as noted earlier. Thus, the coefficient of the age variable of -54 means if age increases by a year, the *desired* hours of work will decrease by 54 hours, *ceteris paribus*. Of course, we do not actually observe the desired hours of work, for it is an abstract construct.

In our example we have 753 observations. It is a laborious task to compute the marginal impact of each regressor for all the 753 observations. In practice, one can compute the marginal impact at the *average value* of each regressor.

Since the probability of Y^* must lie between zero and one, the *product* of each slope coefficient multiplied by this probability will be smaller (in absolute value) than the slope coefficient itself. As a result, the marginal impact of a regressor on the mean value of the *observed* regressand will be smaller (in absolute value) than indicated by the value of the slope coefficient given in Table 11.4. The sign of the marginal impact will depend on the sign of the slope coefficient, for the probability of observing Y_i^* is always positive. Packages like *Stata* and *Eviews* can compute the marginal impact of each regressor.

Statistical significance of the estimated coefficients

Table 11.4 presents the standard errors, the Z -statistics (standard normal distribution values) and the p values of each estimated coefficient.¹² As the table shows all the coefficients are statistically significant at the 10% or lower level of significance.

For the Tobit model there is no conventional measure of R^2 . This is because the standard linear regression model estimates parameters by minimizing the residual sum of squares (RSS), whereas the Tobit model maximizes the likelihood function. But if you want to compute an R^2 equivalent to the conventional R^2 , you can do so by squaring the coefficient of correlation between the actual Y values and the Y values estimated by the Tobit model.

The test of the omitted variables or superfluous variables can be conducted in the framework of the usual large sample tests, such as the likelihood ratio, Wald, or Lagrange Multiplier (L). Try this by adding the experience-squared variable to the model or father's education and mother's education variables to the model.

Caveats

In the Tobit model it is assumed that the error term follows the normal distribution with zero mean and constant variance (i.e. homoscedasticity).

Non-normality of error term

In the censored regression models under non-normality of the error term the estimators are not consistent. Again, some remedial methods are suggested in the literature. One is to change the error distribution assumption. For example, *Eviews* can estimate such regression models under different probability distribution assumptions for the error term (such as logistic and extreme value). For a detailed discussion, see the books by Maddala and Wooldridge.¹³

Heteroscedasticity

In the usual linear regression model, if the error term is heteroscedastic, the OLS estimators are consistent, though not efficient. In Tobit-type models, however, the estimators are *neither consistent nor efficient*. There are some methods to deal with this problem, but a detailed discussion of them would take us far afield.¹⁴ However, statistical packages, such as *Stata* and *Eviews*, can compute *robust* standard errors, as shown in Table 11.5.

As you can see, there are no vast differences in the estimated standard errors in the two tables, but this need not always be the case.

¹² Because of the large sample size, we use the standard normal than the t distribution.

¹³ For detailed, but somewhat advanced, discussion, see G. S. Maddala, *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, UK, 1983, and Wooldridge, J. M., *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA, 2002.

¹⁴ For an advanced discussion, see Maddala and Wooldridge, *op cit*.

Table 11.5 Robust estimation of the Tobit model.

Dependent Variable: HOURS				
Method: ML – Censored Normal (TOBIT) (Quadratic hill climbing)				
Sample: 1 753				
Included observations: 753				
Left censoring (value) at zero				
Convergence achieved after 6 iterations				
QML (Huber/White) standard errors & covariance				
	Coefficient	Std. Error	z-Statistic	Prob.
C	1126.335	386.3109	2.915618	0.0035
AGE	-54.10976	6.535741	-8.279056	0.0000
EDUC	38.64634	20.30712	1.903094	0.0570
EXPER	129.8273	17.27868	7.513728	0.0000
EXPERSQ	-1.844762	0.536345	-3.439505	0.0006
FAMINC	0.040769	0.005608	7.269982	0.0000
KIDSLT6	-782.3734	104.6233	-7.478004	0.0000
HUSWAGE	-105.5097	16.33276	-6.460007	0.0000
Error Distribution				
SCALE:C(9)	1057.598	42.80938	24.70482	0.0000
Mean dependent var	740.5764	S.D. dependent var	871.3142	
S.E. of regression	707.2850	Akaike info criterion	10.08993	
Sum squared resid	3.72E+08	Schwarz criterion	10.14520	
Log likelihood	-3789.858	Avg. log likelihood	-5.033012	
Left censored obs	325	Right censored obs	0	
Uncensored obs	428	Total obs	753	

III

11.3 Truncated sample regression models

Earlier we discussed the difference between censored and truncated sample regression models. Having discussed the censored sample regression model, we now turn our attention to truncated sample regression models.

In truncated samples if we do not have information on the regressand, we do not collect information on the regressors that may be associated with the regressand. In our illustrative example, we do not have data on hours worked for 325 women. Therefore we may not consider information about socio-economic variables for these observations, even though we have that information on them in the current example.

Why, then, not estimate the hours function for the sub-sample of 428 working women only using the method of OLS? As a matter of fact, we did that in Table 11.2. However, the OLS estimators are inconsistent in this situation. Since the sample is truncated, the assumption that the error term in this model is normally distributed with mean μ and variance σ^2 distributed cannot be maintained. Therefore, we have to use what is known as the **truncated normal distribution**. In that case we have to use a nonlinear method of estimation, such as the ML method.

Using ML, we obtain the results in Table 11.6. If you compare these results with the OLS results give in Table 11.2, you will see the obvious differences, although the signs of the coefficients are the same.

Table 11.6 ML estimation of the truncated regression model.

Dependent Variable: HOURS				
Method: ML – Censored Normal (TOBIT) (Quadratic hill climbing)				
Sample (adjusted): 1 428				
Included observations: 428 after adjustments				
Truncated sample				
Left censoring (value) at zero				
Convergence achieved after 6 iterations				
QML (Huber/White) standard errors & covariance				
	Coefficient	Std. Error	z-Statistic	Prob.
C	1864.232	397.2480	4.692867	0.0000
AGE	-22.88776	7.616243	-3.005125	0.0027
EDUC	-50.79302	20.77250	-2.445205	0.0145
EXPER	73.69759	22.42240	3.286784	0.0010
EXPERSQ	-0.954847	0.575639	-1.658761	0.0972
FAMINC	0.036200	0.006947	5.210857	0.0000
KIDSLT6	-391.7641	193.4270	-2.025385	0.0428
HUSWAGE	-93.52777	19.11320	-4.893360	0.0000
Error Distribution				
SCALE:C(9)	794.6310	56.36703	14.09744	0.0000
Mean dependent var	1302.930	S.D. dependent var	776.2744	
S.E. of regression	696.4534	Akaike info criterion	15.78988	
Sum squared resid	2.03E+08	Schwarz criterion	15.87524	
Log likelihood	-3370.035	Avg. log likelihood	-7.873913	
Left censored obs	0	Right censored obs	0	
Uncensored obs	428	Total obs	428	
<i>Note:</i> The standard errors presented in this table are robust standard errors.				

If you compare the results of the censored regression given in Table 11.5 with the truncated regression given in Table 11.6, you will again see differences in the magnitude and statistical significance of the coefficients. Notice particularly that the education coefficient is positive in the censored regression model, but is negative in the truncated regression model.

Interpretation of the truncated regression coefficients

As in the Tobit model, an individual regression coefficient measures the marginal effect of that variable on the mean value of the regressand for *all* observations – that is, including the non-included observations. But if we consider only the observations in the (truncated) sample, then the relevant (partial) regression coefficient has to be multiplied by a factor which is smaller than 1. Hence, the within-sample marginal effect of a regressor is smaller (in absolute value) than the value of the coefficient of that variable, as in the case of the Tobit model.

Tobit vs. truncated regression model

Now, between censored and truncation regression models, which is preferable? Since the Tobit model uses more information (753 observations) than the truncated

regression model (428 observations), estimates obtained from Tobit are expected to be more efficient.¹⁵

11.4 Summary and conclusions

In this chapter we discussed the nature of censored regression models. The key here is the concept of a *latent variable*, a variable which, although intrinsically important, may not always be observable. This results in a censored sample in which data on the regressand is not available for several observations, although the data on the explanatory variables is available for all the observations.

In situations like this OLS estimators are biased as well as inconsistent. Assuming that the error term follows the normal distribution with zero mean and constant variance, we can estimate censored regression models by the method of maximum likelihood (ML). The estimators thus obtained are consistent.

The slope coefficients estimated by ML need to be interpreted carefully. Although we can interpret the slope coefficient as giving the marginal impact of a variable on the mean value of the *latent* variable, holding other variables constant, we cannot interpret it so with respect to the *observed value* of the latent variable. Here we have to multiply the slope coefficient by the probability of observing the latent variable. And this probability depends on all the explanatory variables and their coefficients. However, modern statistical software packages do this relatively easily.

One major caveat is that the ML estimators are consistent only if the assumptions about the error term are valid. In cases of heteroscedasticity and non-normal error term, the ML estimators are inconsistent. Alternative methods need to be devised in such situations. Some solutions are available in the literature. We can, however, compute robust standard errors, as illustrated by a concrete example.

The truncated regression model differs from the censored regression model in that in the former we observe values of the regressors only if we have data on the regressand. In the censored regression model, we have data on the regressors for all the values of the regressand including those values of the regressand that are not observed or set to zero or some such limit.

In practice, censored regression models may be preferable to the truncated regression models because in the former we include all the observations in the sample, whereas in the latter we only include observations in the truncated sample.

Finally, the fact that we have software to estimate censored regression models does not mean that Tobit-type models are appropriate in all situations. Some of the situations where such models may not be applicable are discussed in the references cited in this chapter.

Exercises

11.1 Include the Faminc-squared variable in both the censored and truncated regression models discussed in the chapter and compare and comment on the results.

¹⁵ Technically, this is the result of the fact that the Tobit likelihood function is the sum of the likelihood functions of truncated regression model and the probit likelihood function.

11.2 Expand the models discussed in this chapter by considering interaction effects, for example, education and family income.

11.3 The data given in Table 11.1 includes many more variables than are used in the illustrative example in this chapter. See if adding one or more variables to the model in Tables 11.4 and 11.6 substantially alters the results given in these tables.