

1.6 Count Data–Poisson Regression

Hill's *U/G Econometrics* ch 18.4.3 and EViews 658ff.

The probit and logit models are produced by allowing the probability in the Bernoulli model to vary across individuals as x varies. In the same way regressors can be hung on the parameters of the other distributions of first year Statistics—the Poisson and exponential.

Poisson or count data models focus on the “number of occurrences” of an event. Here the outcome variable is $y = 0, 1, 2, 3, \dots$

Examples include:

- The number of visits to a doctor a person makes during a year.
- The number of children in a household.

- The number of televisions in a household.

We have a sample of persons, or households which differ in ways (age, health, income,...) that affect the probability distribution of the number of visits, children or TVs

- The probability distribution used is the *Poisson*. From the first year if Y is a Poisson random variable, then its probability function is given by

$$P(Y = y) = p(y) = \frac{\lambda^y e^{-\lambda}}{y!} : y = 0, 1, 2, \dots$$

This probability function has one parameter, λ , which is the mean (and variance) of Y .

- Suppose that units in our sample vary in accordance with the value of some variable x .

- The simplest specification of a relationship between λ_i and x_i would be

$$\lambda_i = \beta_0 + \beta_1 x_i.$$

- However the right hand side might be negative, which is unacceptable. A simple way of guaranteeing that it is positive is to specify

$$\lambda_i = e^{\beta_0 + \beta_1 x_i}.$$

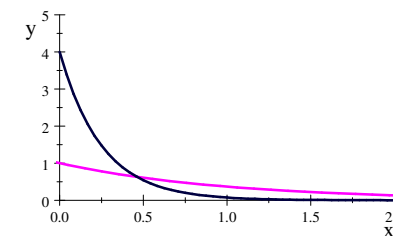
- This is the **Poisson regression model**. The parameters could be estimated by maximum likelihood, i.e. maximise

$$L(\beta_1, \beta_2; y) = \prod \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \text{ where } \lambda_i = e^{\beta_0 + \beta_1 x_i}.$$

Another 1st year distribution is the **exponential** distribution with density

$$f(y) = \lambda e^{-\lambda y}, \text{ for } y \geq 0 \text{ and } \lambda > 0.$$

The expected value is $\frac{1}{\lambda}$.



expos: large λ , small λ

This is the simplest model used for durations or waiting times. Suppose we have a cross-section of individuals who have been unemployed and there is data on their personal characteristics (age, qualifications, etc) and how long they were unemployed. Their experience could be modelled using the exponential with the parameter varying across individuals according to

$$\lambda_i = \beta_0 + \beta_1 x_i.$$

A lot of empirical labour economics is based on this model and extensions of it. EViews does not cover these duration models and we won't consider any empirical examples.

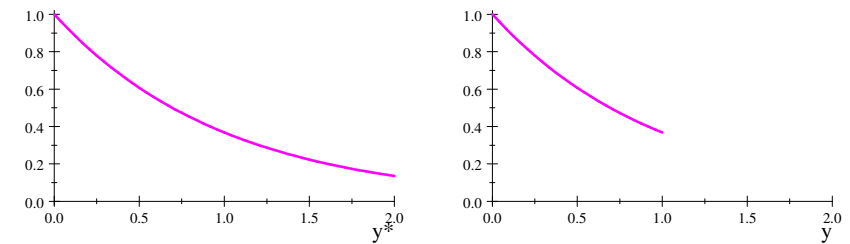
1.7 Censoring—the Tobit Model

Hill's *U/G Econometrics* ch 18.5.1 and EViews pp. 656ff.

There is an important class of models in which an observable outcome, response or dependent variable whose range is *limited* because it is cut-off, or censored, at some particular value.

The first example involves *incomplete recording* of data.

- We have data on the duration of unemployment for the workers made redundant when a factory closed 1 year ago. For those who got jobs after less than 1 year we *know* the length of their unemployment spell. Some, though, are still unemployed and their unemployment is recorded as 1 year. We say that the data is *censored*.



density of latent y^*

censored at one year

- Thus if the distribution of unemployment duration, y^* , is exponential the distribution of recorded durations, y , looks like the second diagram with a blob of probability (an *atom*) on 1 corresponding to $P(y^* \geq 1)$.
- If we use the latent variable formalism, the underlying variable y^* is observed for values less than c (one year in the example): otherwise it is recorded as c

$$y = \begin{cases} y^* & \text{if } y^* < c \\ c & \text{otherwise.} \end{cases}$$

- If we take the average of the *reported* durations, we get a downward biased estimate of the expected duration because durations of at least 1 year are being treated like observations of exactly 1 year.
- If we ignore the observations at 1 and take the average of the rest we get an *even bigger* downward bias. (In this case the data is said to be *truncated*.)
- The standard strategy for analysing censored data is to use information (or make assumptions) about the *distribution* of the period of unemployment and use maximum likelihood for estimating the parameter(s).

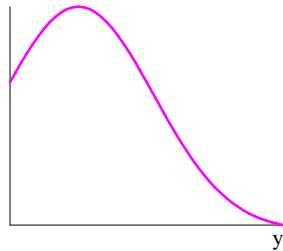
In the next *example* from Hill 18.5.1 there is not incomplete recording but rather a lot of observations on a “boundary”. (A lower boundary in this case, so “left-censoring”.)

- We randomly select individuals and ask them, “How much did you give to charity last year?”
- There are going to be many responses of \$0, and none will be less than \$0. Charitable donation is an example of a dependent variable that is limited in its range.
- We want to model a person’s giving in terms of a variable x , e.g. income.
- The *Tobit* or censored normal regression model has for the i -th person

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

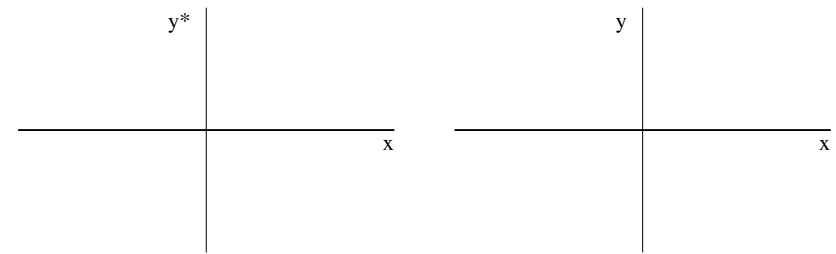
$$y_i^* = \beta_0 + \beta_1 x_i + u_i, \quad u_i \sim N(0, \sigma^2).$$

The density of y looks like this with an atom on 0 corresponding to $P(y^* \leq 0)$



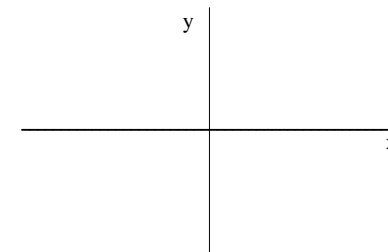
'centre' at $\beta_0 + \beta_1 x$

- The name *Tobit* is from the model's inventor James Tobin and its similarity to the probit model.
- Least squares regression of y on x fails, being biased even in large samples, whether you (a) leave the \$0 in the data and treat them like all other observations, or (b) throw out all the \$0 observations. (Compare with estimating the mean duration of unemployment discussed above.)



scatter plot of y^* and x

(a) scatter plot of y and x



(b) scatter plot y trunc

- Again ML is the standard estimation method.
- The EViews example (p. 651) has the number of extra-marital affairs as the dependent variable.

	Coefficient	Std. Error	z-Statistic	Prob.
C	7.608487	3.905837	1.947979	0.0514
Z1	0.945787	1.062824	0.889881	0.3735
Z2	-0.192698	0.080965	-2.380015	0.0173
Z3	0.533190	0.146602	3.636997	0.0003
Z4	1.019182	1.279524	0.796532	0.4257
Z5	-1.699000	0.405467	-4.190231	0.0000
Z6	0.025361	0.227658	0.111399	0.9113
Z7	0.212983	0.321145	0.663198	0.5072
Z8	-2.273284	0.415389	-5.472657	0.0000
<hr/>				
	Error	Distribution		
SCALE C(10)	8.258432	0.554534	14.89256	0.0000
R-squared	0.151569	Mean dependent var	1.455907	
Adjusted R-squared	0.138649	S.D. dependent var	3.298758	
S.E. of regression	3.061544	Akaike info criterion	2.378473	
Sum squared resid	5539.472	Schwarz criterion	2.451661	
Log likelihood	-704.7311	Hannan-Quinn criter.	2.406961	
Avg. log likelihood	-1.172597			
Left censored obs	451	Right censored obs	0	
Uncensored obs	150	Total obs	601	

Figure 2:

There is the usual panel of estimates, standard errors, z -stats (Wald tests of $\beta_i = 0$) and prob values plus a panel containing the model selection measures, AIC etc.

These limited dependent variable models originated in Statistics but in the last 25 years they have attracted a lot of attention from economists and econometricians, like Heckman and McFadden (Nobel Prizes 2000), have tweaked them so that they serve the purposes of economists better.

The next type of model was created more than 50 years ago and was thought more suited to Economics than the basic regression model. For this work Haavelmo got the Nobel Prize in 1989.