Applied econometrics:

Corner solution responses, sample selection and count responses

Måns Söderbom*

University of Gothenburg

Fall, 2021

---

* I apologize for any errors in these lecture notes. Feel free to contact me at mans.soderbom@economics.gu.se if you spot mistakes or if you have questions on the material in these notes. Thanks.

CHAPTER 1:

CORNER SOLUTION RESPONSES

1. Tobit Estimation of Corner Solution Models

Reference: Wooldridge (2010), Chapter 17.1-17.6.3.

In this lecture we consider econometric issues that arise when the dependent variable is bounded but continuous within the bounds:

$$lo \leq y_i \leq hi,$$

where lo denotes the lower bound (limit) and hi the higher bound, and where these bounds are the result of real economic constraints. The most common case is when a nonnegative response variable y has a continuous distribution over strictly positive values but $\Pr(y = 0) > 0$, resulting in a pileup of observations of y at zero. You will often find this in micro data, e.g. household expenditure on education, health, alcohol,...

When we are modeling a variable expressed as a fraction (percentage of output exported by firms; fraction of charitable contributions made to religious organizations, etc.), we may have $lo = 0$ and $hi = 1$, in which case it makes sense to treat y as having a continuous distribution over the open interval (0,1).

We can think of this type of variable as a hybrid between a continuous variable (for which the linear model is appropriate) and a binary variable (for which one would typically use a binary choice model). Indeed, as we shall see, the econometric model designed to model corner solution variables looks like a hybrid between OLS and the probit model. In what follows we focus on the case where $lo = 0$, $hi = \infty$, however generalizing beyond this case is reasonably straightforward.

The distinction between censored and corner responses Unlike most other authors, Wooldridge makes a clear distinction between "censored responses" and "corner responses". The word "censored" implies that we are not *observing* the entire range of the response variable. For example, if our outcome variable is the demand for tickets for a concert, we won't observe demand whenever the concert sells out.

That is not the case for corner responses - these are the actual outcomes. For example, in a model of charitable contributions the outcome might be zero but this does not mean that contributions are "censored" at zero. The same econometric techniques can be used for analyzing a censored variable as for a corner response model, however the objects of interest are typically different.

Can we use linear regression if y is a corner response variable? Let y be a variable that is equal to zero for some non-zero proportion of the population, and continuous and positive if not equal to zero. As usual, we want to model y as a function of a set of variables $x_2, ..., x_k$ - or in matrix notation:

$$\boldsymbol{x} = \begin{bmatrix} 1 & x_2 & x_3 & ... & x_k \end{bmatrix}.$$

For binary choice models OLS can be a useful starting point (yielding the linear probability model), even though the dependent variable is not continuous. We now have a variable which is 'closer' to being a continuous variable - it's discrete in the sense that it is either in the corner (equal to zero) or not (in which case it's continuous). If we are interested in the effect of x_j on the mean response $E(y|\boldsymbol{x})$, why not use OLS?

Recall that there are a number of reasons why we may not prefer to estimate binary choice models using OLS. For similar reasons OLS may not be an ideal estimator for corner response models:

- Based on OLS estimates we can get **negative predictions** (or, more generally, predictions outside the bounds lo, hi) which doesn't make sense (if we are modelling household expenditure on education, for instance, negative predicted values do not make sense).

- Conceptually, the idea that a corner solution variable is **linearly** related to a continuous independent variable for all possible values is a bit suspect. It seems more likely that for observations close to the corner (close to zero), changes in some continuous explanatory variable x_j has a smaller effect on the outcome than for observations far away from the corner. So if we are interested in understanding how y depends on x_j among low values of y, linearity is not attractive.

- A third (and less serious) problem is that the residual u is likely to be heteroskedastic - but we can deal with this by simply correcting the standard errors.

- A fourth and related problem is that, because the distribution of y has a 'spike' at zero, the residual cannot be normally distributed. This means that OLS point estimates are unbiased, but inference in small samples cannot be based on the usual suite of normality-based distributions such as the t test.

All of this is very similar to the problems identified with the linear probability model. Naturally, if for some reason we feel these problems are not very important, we may opt for a linear regression approach. In the special case where the model is saturated, so that the set of explanatory variables consists of dummy variables representing each mutually exclusive and exhaustive category present in the data, the OLS estimate of $E(y|\boldsymbol{x})$ will be numerically identical to the sub-sample average of y for each category. No assumptions about the functional form relationship between the dependent variable and the explanatory variables are needed in this case, and negative predictions will never arise.

Example: Charitable contributions Example 17.1 in Wooldridge (2010) is a nice illustration of possible behavioral underpinnings of an empirical corner response model. Suppose we study the determinants of charitable giving, and suppose the utility function of individual i is given by

$$util_i\left(c_i, q\right) = c_i + a_i \log\left(1 + q_i\right),$$

where c is consumption and q is charitable giving. The variable a_i determines the marginal utility of giving for individual i. Maximizing utility subject to the following constraints

$$\begin{aligned} c_i + p_i q_i &= m_i, \\ c_i &\geq 0, \\ m_i &\geq 0, \end{aligned}$$

where m_i is income and p_i is the price of a "unit" of charitable contributions, gives a solution for q_i:

$$q_i = \left\{ \begin{array}{l} 0 \text{ if } a_i/p_i \leq 1 \\ \\ a_i/p_i - 1 \text{ if } a_i/p_i > 1 \end{array} \right\},$$

which we can write as

$$1 + q_i = \max\left(1, a_i/p_i\right).$$

Now specify a_i, the determinant of the marginal utility of giving, as

$$a_i = \exp\left(\boldsymbol{z}_i \boldsymbol{\gamma} + u_i\right),$$

where $\boldsymbol{z}_i$ is a vector of explanatory variables $\boldsymbol{\gamma}$ is a parameter vector, and u_i is an unobservable. We can now write

$$\log\left(1 + q_i\right) = \max\left(0, \boldsymbol{z}_i \boldsymbol{\gamma} - \log\left(p_i\right) + u_i\right).$$

The main insight from this little theoretical detour is that we have obtained a behavioral model (underpinned by utility maximization) of q that recognizes that a corner response ($q = 0$) may be optimal in theory. To take this (type of) model to the data, it would thus be useful to have an estimator that recognizes the presence of corner outcomes too; that's what tobit does.

1.1. Type I Tobit

We continue to focus on the case where there is one corner, at zero. We write our population model as

$$y = \max\left(0, \boldsymbol{x}\boldsymbol{\beta} + u\right),$$

where the unobserved term u is assumed independent of $\boldsymbol{x}$, mean-zero, homoskedastic and normally distributed. These assumptions define the type I Tobit

It can sometimes be useful for certain derivations.(see below) to write y as a latent variable model,

$$\begin{aligned} y^* &= \boldsymbol{x}\boldsymbol{\beta} + u, \\ y_i &= \max\left(0, y_i^*\right). \end{aligned} \tag{1.1}$$

Note however that y^* is typically not a relevant quantity in corner response models (e.g. it's not obvious $y^* < 0$ would be very meaningful in the context of the charity contributions example above). In contrast, if your outcome variable is censored, it would be meaningful to think about the determinants of y^*; in that case we would be interested in $E\left(y^*|\boldsymbol{x}\right)$.

As noted above, a corner response variable is a kind of hybrid: both discrete and continuous. The discrete part is due to the piling up of observations at zero. Using the latent variable formulation above, the probability that y is equal to zero can be written

$$\begin{aligned} \Pr\left(y = 0|\boldsymbol{x}\right) &= \Pr\left(y^* \leq 0\right), \\ &= \Pr\left(\boldsymbol{x}\boldsymbol{\beta} + u \leq 0\right), \\ &= \Pr\left(u \leq -\boldsymbol{x}\boldsymbol{\beta}\right) \\ &= \Phi\left(\frac{-\boldsymbol{x}\boldsymbol{\beta}}{\sigma}\right) \text{ (integrate; normal distribution)} \\ \Pr\left(y = 0|\boldsymbol{x}\right) &= 1 - \Phi\left(\frac{\boldsymbol{x}\boldsymbol{\beta}}{\sigma}\right) \text{ (by symmetry),} \end{aligned}$$

exactly like the probit model. In contrast, if $y > 0$ then it is continuous:

$$y = \boldsymbol{x}\boldsymbol{\beta} + u.$$

It follows that the conditional density of y is equal to

$$f\left(y|\boldsymbol{x};\boldsymbol{\beta},\sigma\right) = \left[1 - \Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma\right)\right]^{1_{[y(i)=0]}} \left[\phi\left(\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)\right]^{1_{[y(i)>0]}},$$

where $1_{[a]}$ is a dummy variable equal to one if a is true. Thus the contribution of observation i to the sample log likelihood is

$$\ln L_i = 1_{[y(i)=0]} \log\left[1 - \Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma\right)\right] + 1_{[y(i)>0]} \log\left[\phi\left(\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)\right],$$

and the sample log likelihood is

$$\ln L\left(\boldsymbol{\beta},\sigma\right) = \sum_{i=1}^{N} \ln L_i.$$

Estimation is done by means of maximum likelihood; as usual, we assume the sample has been randomly drawn from the population.

1.1.1. Interpretation of Tobit parameters

How do we interpret the parameters $\boldsymbol{\beta}$? We see straight away from the latent variable model that β_j is interpretable as the partial (marginal) effects of x_j on the conditional expected value of latent variable y^*:

$$\frac{\partial E\left(y^*|\boldsymbol{x}\right)}{\partial x_j} = \beta_j,$$

if x_j is a continuous variable, and

$$E\left(y^*|x_j = 1\right) - E\left(y^*|x_j = 0\right) = \beta_j$$

if x_j is a dummy variable (of course if x_j enters the model nonlinearly these expressions need to be modified accordingly). I have omitted i-subscripts for simplicity. If that's what we want to know, then we are home: all we need is an estimate of the relevant parameter β_j.

The point emphasized by Wooldridge in chapter 17 is that that is *not* what we want to know, since the latent variable y^* is not our outcome variable of interest.

Typically we are interested in the partial effect of x_j on the expected **actual outcome** y, rather than on the latent variable. In fact there are two different potentially interesting marginal effects, namely

$$\frac{\partial E\left(y|\boldsymbol{x}\right)}{\partial x_j}, \qquad \text{(Unconditional on y)}$$

and

$$\frac{\partial E\left(y|\boldsymbol{x}, y>0\right)}{\partial x_j}. \qquad \text{(Conditional on y>0)}$$

We need to be clear on which of these we are interested in. Now let's see what these marginal effects look like.

The marginal effects on expected y, conditional on y positive. We want to derive

$$\frac{\partial E\left(y|\boldsymbol{x}, y>0\right)}{\partial x_j}.$$

Recall that the model can be written

$$\begin{aligned} y &= \max\left(y^*, 0\right), \\ y &= \max\left(\boldsymbol{x\beta} + u, 0\right). \end{aligned}$$

We begin by writing down $E\left(y|\boldsymbol{x}, y>0\right)$:

$$\begin{aligned} E\left(y|y>0, \boldsymbol{x}\right) &= E\left(\boldsymbol{x\beta} + u|y>0, \boldsymbol{x}\right), \\ E\left(y|y>0, \boldsymbol{x}\right) &= \boldsymbol{x\beta} + E\left(u|y>0, \boldsymbol{x}\right), \\ E\left(y|y>0, \boldsymbol{x}\right) &= \boldsymbol{x\beta} + E\left(u|u>-\boldsymbol{x\beta}\right) \end{aligned}$$

Because of the truncation (y is always positive, or, equivalently, u is always larger than $-\boldsymbol{x\beta}$), dealing with the second term is not as easy as it may seem. We begin by taking on board the following result for normally distributed variables:

- **A useful result.** If z follows a normal distribution with mean zero, and variance equal to one (i.e. a standard normal distribution), then

$$E\left(z|z>c\right)=\frac{\phi\left(c\right)}{1-\Phi\left(c\right)}, \tag{1.2}$$

  where c is a constant (i.e. the lower bound here), ϕ denotes the standard normal probability density, and Φ is the standard normal cumulative density.

The error term u is not, in general, standard normal because the variance is not necessarily equal to one, but by dividing and multiplying through with its standard deviation σ we can transform u to become standard normal:

$$E\left(y|y>0,x\right)=\boldsymbol{x\beta}+\sigma E\left(u/\sigma|u/\sigma>-\boldsymbol{x\beta}/\sigma\right).$$

That is, (u/σ) is now standard normal, and so we can apply the above 'useful result', i.e. eq (1.2), and write:

$$E\left(u|u>-\boldsymbol{x\beta}\right)=\sigma\frac{\phi\left(-\boldsymbol{x\beta}/\sigma\right)}{1-\Phi\left(-\boldsymbol{x\beta}/\sigma\right)},$$

and thus

$$E\left(y|y>0,\boldsymbol{x}\right)=\boldsymbol{x\beta}+\sigma\frac{\phi\left(-\boldsymbol{x\beta}/\sigma\right)}{1-\Phi\left(-\boldsymbol{x\beta}/\sigma\right)}.$$

With slightly cleaner notation,

$$E\left(y|y>0,\boldsymbol{x}\right)=\boldsymbol{x\beta}+\sigma\frac{\phi\left(\boldsymbol{x\beta}/\sigma\right)}{\Phi\left(\boldsymbol{x\beta}/\sigma\right)},$$

which is often written as

$$E\left(y|y>0,\boldsymbol{x}\right)=\boldsymbol{x\beta}+\sigma\lambda\left(\boldsymbol{x\beta}/\sigma\right), \tag{1.3}$$

where the function λ is defined as

$$\lambda(z) = \frac{\phi(z)}{\Phi(z)}.$$

in general, and known as the **inverse Mills ratio** function.

- Have a look at the inverse Mills ratio function in Section 1 in the appendix, Figure 1.

Equation (1.3) shows that the expected value of y, given that y is not zero, is equal to $\boldsymbol{x\beta}$ **plus** a term $\sigma\lambda(\boldsymbol{x\beta}/\sigma)$ which is strictly positive (how do we know that?).

We can now obtain the partial effect with respect to a continuous explanatory variable x_j:

$$\begin{aligned}\frac{\partial E(y|y>0,\boldsymbol{x})}{\partial x_j} &= \beta_j + \sigma\frac{\partial\lambda(\boldsymbol{x\beta}/\sigma)}{\partial x_j}, \\ &= \beta_j + \sigma\left(\beta_j/\sigma\right)\lambda', \\ &= \beta_j\left(1+\lambda'\right),\end{aligned}$$

where λ' denotes the partial derivative of λ with respect to $(\boldsymbol{x\beta}/\sigma)$ (note: I am assuming of course that x_j is not functionally related to any other variable - i.e. it enters the model linearly - this means that I don't have to worry about higher-order terms). It is tedious but fairly easy to show that

$$\lambda'(z) = -\lambda(z)\left[z+\lambda(z)\right]$$

in general, hence

$$\frac{\partial E(y|y>0,\boldsymbol{x})}{\partial x_j} = \beta_j\left\{1-\lambda(\boldsymbol{x\beta}/\sigma)\left[\boldsymbol{x\beta}/\sigma+\lambda(\boldsymbol{x\beta}/\sigma)\right]\right\}.$$

This shows that the partial effect of x_j on $E(y|y>0,\boldsymbol{x})$ is not determined just by β_j. In fact, it depends on **all parameters $\boldsymbol{\beta}$** in the model as well as on the values of **all explanatory variables $\boldsymbol{x}$**, and the standard deviation of the error term u. The term in $\{\cdot\}$ is often referred to as the **adjustment factor**, and it can be shown that this is always larger than zero and smaller than one (why is this useful to know?).

It should be clear that, just as in the case for probits and logits, we need to evaluate the marginal effects at specific values of the explanatory variables. This should come as no surprise, since one of the reasons we may prefer tobit to OLS is that we have reasons to believe the partial effects may differ depending on how close to the corner (zero) a given observation is (see above). In Stata we can use *mfx compute* or *margins* to obtain estimates of marginal effects. How this is done will be clearer in a moment.

STUDENT EXERCISE:

1. Write down the expression for $\frac{\partial E(y|y>0,\boldsymbol{x})}{\partial z_1}$ if: i) $x_1 = \log(z_1)$; ii) $x_1 = z_1$, $x_2 = z_1^2$; iii) $x_1 = z_1$, $x_2 = z_1 z_2$.

2. If $x_1 = z_1$, how would the elasticity of $E(y|y>0,\boldsymbol{x})$ with respect to z_1 look like?

3. If x_1 is a dummy variable, what's the partial effect of interest?

The marginal effects on expected y, unconditional on the value of y Recall:

$$\begin{aligned} y &= \max(y^*, 0), \\ y &= \max(\boldsymbol{x\beta} + u, 0). \end{aligned}$$

I now need to derive

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j}.$$

We write $E(y|\boldsymbol{x})$ as follows:

$$\begin{aligned} E(y|\boldsymbol{x}) &= \Phi(-\boldsymbol{x\beta}/\sigma) \cdot E(y|y=0,\boldsymbol{x}) + \Phi(\boldsymbol{x\beta}/\sigma) \cdot E(y|y>0,\boldsymbol{x}), \\ E(y|\boldsymbol{x}) &= \Phi(-\boldsymbol{x\beta}/\sigma) \cdot 0 + \Phi(\boldsymbol{x\beta}/\sigma) \cdot E(y|y>0,\boldsymbol{x}), \\ E(y|\boldsymbol{x}) &= \Phi(\boldsymbol{x\beta}/\sigma) \cdot E(y|y>0,\boldsymbol{x}), \end{aligned}$$

i.e. the probability that y is positive times the expected value of y given that y is indeed positive. Recall that $E(y|y>0,\boldsymbol{x})$ was derived above, so we know what the expression looks like. Using the product rule for differentiation,

$$\frac{\partial E(y|x)}{\partial x_j} = \Phi(\boldsymbol{x\beta}/\sigma)\cdot\frac{\partial E(y|y>0,\boldsymbol{x})}{\partial x_j} + \phi(\boldsymbol{x\beta}/\sigma)\frac{\beta_j}{\sigma}\cdot E(y|y>0,\boldsymbol{x}),$$

where

$$\frac{\partial E(y|y>0,\boldsymbol{x})}{\partial x_j} = \beta_j\left\{1-\lambda(\boldsymbol{x\beta}/\sigma)\left[\boldsymbol{x\beta}/\sigma+\lambda(\boldsymbol{x\beta}/\sigma)\right]\right\},$$

and

$$E(y|y>0,\boldsymbol{x}) = \boldsymbol{x\beta}+\sigma\lambda(\boldsymbol{x\beta}/\sigma).$$

Hence

$$\begin{aligned}\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} &= \Phi(\boldsymbol{x\beta}/\sigma)\cdot\beta_j\left\{1-\lambda(\boldsymbol{x\beta}/\sigma)\left[\boldsymbol{x\beta}/\sigma+\lambda(\boldsymbol{x\beta}/\sigma)\right]\right\}\\ &\quad+\phi(\boldsymbol{x\beta}/\sigma)\frac{\beta_j}{\sigma}\cdot\left[\boldsymbol{x\beta}+\sigma\lambda(\boldsymbol{x\beta}/\sigma)\right],\end{aligned}$$

which looks complicated but the good news is that several of the terms cancel out, so that:

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} = \beta_j\Phi(\boldsymbol{x\beta}/\sigma). \tag{1.4}$$

STUDENT EXERCISE: Prove that eq. (1.4) is true.

Equation (1.4) has a straightforward interpretation: the marginal effect of x_j on the expected value of y, conditional on the vector $\boldsymbol{x}$, is simply the parameter β_j times the probability that y is larger than zero. Of course, this probability is smaller than one, so it follows immediately that: i) the marginal effect is strictly smaller than the parameter β_j; ii) that its sign is determined by the sign of β_j.

EXAMPLE: MODELLING ANNUAL HOURS WORKED. Section 2 in the appendix replicates the results discussed in Example 17.2, Wooldridge (2010), pp. 678-680. It also shows how to obtain the relevant partial effects using the Stata *margins* command.

2. Specification Issues in Tobit Models

Neglected heterogeneity doesn't pose a problem if the omitted variable, q, is normally distributed and independent of the vector $\boldsymbol{x}$; we can estimate the APEs by simply ignoring the heterogeneity. Other forms of neglected heterogeneity may cause problems - see Wooldridge, Section 17.5.1 for a brief discussion.

If one of the explanatory variables of the tobit model is **endogenous**, the tobit estimator is inconsistent. Smith and Blundell (1986) propose a 2-step procedure that is analogous to the Rivers-Vuoung method for binary response models and involves estimating the residual component of the endogenous explanatory variable ($\hat{v}_2$) in a first stage and then adding $\hat{v}_2$ to the set of explanatory variables in the tobit model. The usual t-statistic on $\hat{v}_2$ reported by Tobit provides a simple test of the null that endogeneity is not a problem. Section 3 in the appendix shows the mechanics through an example. The Smith-Blundell approach is implemented by the Stata command `ivtobit depvar [varlist1], twostep.` Note that the endogenous explanatory variable is required to be continuous for this approach to work; if it is not continuous, the likelihood function will have to be adjusted to reflect the nature of the endogenous explanatory variable (e.g. it may be discrete, binary, a corner response variable, etc.)

Heteroskedasticity and **nonnormality** imply that the Tobit estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is inconsistent. This should come as no surprise, given heteroskedasticity and nonnormality change the functional forms for $E(y|x)$ and $E(y|y>0,\boldsymbol{x})$ - recall,

$$\begin{aligned} E(y|x) &= \Phi(\boldsymbol{x}\boldsymbol{\beta}/\sigma) \cdot E(y|y>0,\boldsymbol{x}), \\ E(y|y>0,\boldsymbol{x}) &= \boldsymbol{x}\boldsymbol{\beta} + \sigma\lambda(\boldsymbol{x}\boldsymbol{\beta}/\sigma), \end{aligned}$$

which clearly make use of normality and homoskedasticity. Wooldridge emphasizes that the important point is not whether $\boldsymbol{\beta}$ is estimated with bias, but rather whether our Tobit-based estimates of the partial effects of interest are misleading. In general, however, the tobit-based formulae for partial effects *will* be incorrect under normality and/or homoskedasticity, so it is appropriate to carry out some specification tests. Conditional moment tests are convenient to this end. To illustrate the basic idea, suppose we want to test for heteroskedasticity. Our null hypothesis is that the variance of u is constant and the alternative hypothesis is that the variance of u varies with some explanatory variable x_j:

$$\begin{aligned} H_0 &: E\left(x_j \sigma^2\right) = 0 \\ H_1 &: E\left(x_j \sigma_i^2\right) \neq 0. \end{aligned}$$

For *linear* regression, we can estimate u simply as the difference between observed y and $\boldsymbol{x}\widehat{\boldsymbol{\beta}}$, and then use $\hat{u}^2$ as our estimate of σ^2; we can then test whether x_j is significant in a regression of $\hat{u}^2$ on x_j and the other x-variables in the model (plus, possibly, higher order terms and interaction terms; cf. White's test for heteroskedasticity). For Tobit, however, this simple approach is not feasible, since u cannot be estimated simply as the difference y and $\boldsymbol{x}\widehat{\boldsymbol{\beta}}$, since y is specified as

$$y = \max\left(\boldsymbol{x\beta} + u, 0\right),$$

which is nonlinear. We therefore estimate σ^2 based on generalized residuals, which are nonlinear functions of y and $\boldsymbol{x\beta}$. The null hypothesis above, for example, can be tested against the alternative hypothesis by investigating whether the sample analogue of

$$\frac{E\left(u_i^2|y_i\right) - \sigma^2}{\sigma^2} = -1_{[y_i=0]} \frac{\boldsymbol{x}_i\boldsymbol{\beta}}{\sigma} \lambda\left(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma\right) + 1_{[y_i>0]}\left(\left(\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)^2 - 1\right)$$

covaries with x_j. See Pagan and Vella (1989) for details. You can find on my website an ado file called cmt which carries out conditional moment tests for normality and heteroskedasticity. If I use this program to test the tobit specification used for analyzing hours worked, I can reject homoskedasticity at the 1% level (p-value 0.006) and normality at the 5% level (p-value 0.06). A simple White test following OLS estimation of the model strongly suggests the OLS error term is heteroskedastic too.

How can we proceed if the type I Tobit appears to be misspecified? If heteroskedasticity is the problem, we might allow for non-constant variance of u, e.g. by specifying u as normally distributed with mean zero and variance $\sigma^2 \exp(\boldsymbol{x}_1 \boldsymbol{\delta})$. Alternatively, we might opt for Powell's (1984) censored least absolute deviations (CLAD) estimator which estimates $\boldsymbol{\beta}$ by solving

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} |y_i - \max(0, \boldsymbol{x}_i \boldsymbol{\beta})|.$$

This is attractive primarily because *no* distributional assumption is made about u. This estimator can be derived from the latent variable model if we assume the median of u given $\boldsymbol{x}$ is equal to zero, so that

$$Med(y|\boldsymbol{x}) = \max(0, \boldsymbol{x}\boldsymbol{\beta}).$$

In other words we are modeling the conditional median of y rather than the expected value. See section 17.5.2 in Wooldridge for details on this estimator.

3. Two-Part Models

One implication of the type I Tobit model is that the partial effects of an explanatory variable on $\Pr(y > 0|\boldsymbol{x})$, $E(y|\boldsymbol{x})$ and $E(y|\boldsymbol{x}, y > 0)$ must have the same signs. This may be restrictive: as discussed by Wooldridge, age may have positive effect on the likelihood of having life insurance but perhaps a negative effect on the amount of life insurance coverage. Such a situation would violate the type I Tobit model assumptions.

Another restriction implied by Tobit type I is that the relative effects of two continuous explanatory variables x_j and x_h on $\Pr(y > 0|\boldsymbol{x})$, $E(y|\boldsymbol{x})$ and $E(y|\boldsymbol{x}, y > 0)$ are identical and equal to β_j/β_h.

If we want to allow for separate mechanisms that determine the participation decision ($y = 0$ vs. $y > 0$) and the amount decision (the magnitude of y when it is positive), we can use a **two-part model**. Two-part models allow separate mechanisms to determine the 'participation decision' ($y = 0$ vs. $y > 0$) and the 'amount decision' (the magnitude of y when it is positive). We will discuss three distinct two-part models in this section. First, we introduce some important concepts.

- Define $s = 1[y > 0]$, i.e. s is dummy $= 1$ if y is positive and 0 if y is zero.
- Let w^* be a continuously distributed, nonnegative latent variable, and assume

$$y = s \cdot w^*$$

- Assume s and w^* are **independent**, conditional on explanatory variables $\boldsymbol{x}$. Implications:
  - $E(y|\boldsymbol{x}, s) = s \cdot E(w^*|\boldsymbol{x}, s) = s \cdot E(w^*|\boldsymbol{x})$
  - $E(y|\boldsymbol{x}, s = 1) = E(y|\boldsymbol{x}, y > 0) = E(w^*|\boldsymbol{x})$
  - $E(y|\boldsymbol{x}) = P(s = 1|\boldsymbol{x}) E(w^*|\boldsymbol{x})$

which will be useful later. The independence assumption is of course potentially strong, but we will see below how it can be relaxed.

3.1. Truncated normal hurdle model

The first two-part model we will consider is known as the **truncated normal hurdle model** (Cragg, 1971). The first part of the two-part model is a probit model of participation:

$$\Pr(s = 1|\boldsymbol{x}) = \Phi(\boldsymbol{x\gamma}),$$

where $\Phi(.)$ is the cumulative density function for the standard normal distribution, and $\boldsymbol{\gamma}$ is a vector of parameters. The second part is **truncated regression**, which is a model of y given that $y > 0$:

$$y = \boldsymbol{x\beta} + u \text{ if } \boldsymbol{x\beta} + u > 0,$$

where u is mean-zero, and normally distributed with constant variance σ^2. Underlying this model is an assumption that the latent variable $w^* = \boldsymbol{x\beta} + u$ has a truncated normal distribution, hence it is bounded below at zero (thus u is bounded below at $-\boldsymbol{x\beta}$). Negative predicted outcomes of y are thus ruled out, which is an attractive feature of the model.

The truncated hurdle model is obtained by estimating the participation decision using probit with all observations included (this yields probit estimates of the parameter vector $\boldsymbol{\gamma}$), and the amount decision using truncated regression, with only the positive observations included (yielding estimates of $\boldsymbol{\beta}$).[1] One very nice feature of this model is that, in this special case where $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$, it is equivalent to Tobit type I. Testing $H_0 : \boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$ against the alternative $H_1 : \boldsymbol{\gamma} \neq \boldsymbol{\beta}/\sigma$ is straightforward (e.g. by means of a log likelihood ratio test). Section 4 in the appendix shows estimation results based on the 'Mroz' dataset. Comparing the sum of the log likelihood values for the probit and truncated regression to that for tobit type I (appendix section 2), we obtain an LR statistic equal to 54.28, which with 8 d.f. implies rejection of the null at any significance level.

The expected values are straightforward extensions of the standard tobit models:

$$\begin{aligned} E\left(y|\boldsymbol{x},y>0\right) &= \boldsymbol{x\beta} + E\left(u|u > -\boldsymbol{x\beta}\right) \\ E\left(y|\boldsymbol{x},y>0\right) &= \boldsymbol{x\beta} + \sigma\lambda\left(\boldsymbol{x\beta}/\sigma\right), \end{aligned}$$

and hence

$$E\left(y|\boldsymbol{x}\right) = \Phi\left(\boldsymbol{x\gamma}\right)\left[\boldsymbol{x\beta} + \sigma\lambda\left(\boldsymbol{x\beta}/\sigma\right)\right].$$

It follows that the expression for $\frac{\partial E(y|\boldsymbol{x})}{\partial x_j}$ is somewhat involved - see equation (17.48) in Wooldridge (2010). Obtaining standard errors for estimated $\frac{\partial E(y|\boldsymbol{x})}{\partial x_j}$ is probably best done by means of bootstrapping.

Finally, it is important to understand that, while this estimator certainly is more flexible than tobit, the distributional assumptions (normality, homoskedasticity) are still strong.

[1] The user-written command `craggit` could be used to obtain estimates of all model parameters (this program would have to be downloaded; Stata users can find the program by typing 'findit craggit' in the command window).

3.2. Lognormal hurdle model

Clearly for some outcome variables, assuming a lognormal distribution can be more appropriate than a (truncated) normal distribution. Consider the following specification:

$$y = s \cdot w^* = 1\left[\boldsymbol{x\gamma} + v > 0\right] \exp\left(\boldsymbol{x\beta} + u\right),$$

where (u, v) is independent of $\boldsymbol{x}$ with a bivariate normal distribution, and u and v are independent. This implies that

$$u|\boldsymbol{x} \sim \boldsymbol{Normal}\left(0, \sigma^2\right),$$

and thus the latent variable $w^* = \exp\left(\boldsymbol{x\beta} + u\right)$ has a lognormal distribution, and y conditional on $(\boldsymbol{x}, y > 0)$ has a lognormal distribution. It follows that

$$E\left(y|\boldsymbol{x}, y > 0\right) = \exp\left(\boldsymbol{x\beta} + \boldsymbol{\sigma}^2/2\right),$$

and

$$E\left(y|\boldsymbol{x}\right) = \Phi\left(\boldsymbol{x\gamma}\right) \exp\left(\boldsymbol{x\beta} + \boldsymbol{\sigma}^2/2\right).$$

Estimation of the parameters is easy: probit of s_i on $\boldsymbol{x}$ is a consistent estimator of $\boldsymbol{\gamma}$, while OLS of $\log\left(y_i\right)$ on $\boldsymbol{x}$ is a consistent estimator of $\boldsymbol{\beta}$. If we are primarily interested in estimating e.g. $\frac{\partial E(y|\boldsymbol{x})}{\partial x_j}$, and its associated standard error, we can easily do so if we tweak Stata's `heckman` command. More on this in the next subsection.

Exercise 1. *Obtain the analytical expression for* $\frac{\partial E(y|\boldsymbol{x})}{\partial x_j}$.

3.3. Exponential Type II Tobit Model

The assumption that s and w^* are independent conditional on $\boldsymbol{x}$ is potentially strong. It could well be that the unobservable factors determining s are in fact correlated with the unobservables determining w^*. In that case, the lognormal hurdle model is mis-specified. Fortunately, we can modify the lognormal hurdle model to allow for such a correlation. Wooldridge (section 17.6.3) refers to this modified model as the exponential type II Tobit (ET2T) model. If you are familiar with the 'Heckit' sample selection model, you will see that the likelihood function of ET2T is the same as that for Heckit. But interpretation differs: Heckit is used to correct for the fact that data on your outcome variable of interest is partially missing, while ET2T is used to model a corner response variable. This distinction is somewhat subtle, and we shall discuss it again after we have covered the Heckit model (not this lecture).

The ET2T model is specified as

$$y = s \cdot w^* = 1\left[\boldsymbol{x\gamma} + v > 0\right] \exp\left(\boldsymbol{x\beta} + u\right),$$

where (u, v) is independent of $\boldsymbol{x}$ with a bivariate normal distribution, where u and v are potentially correlated. The correlation between u and v is captured by a parameter ρ. In other words, the ET2T model is just like the lognormal hurdle model except that there is now one more parameter ρ measuring the correlation between u and v. The log likelihood function for the model is as follows (see pp. 697-8 in Wooldridge, 2010, for the derivation, which is somewhat complicated):

$$\begin{aligned} l_i\left(\boldsymbol{\theta}\right) &= 1\left[y_i = 0\right] \log\left(1 - \Phi\left(\boldsymbol{x}_i\boldsymbol{\gamma}\right)\right) \\ &\quad +1\left[y_i > 0\right]\left(A_i + B_i - \log\sigma - \log\left(y_i\right)\right) \end{aligned}$$

where

$$\begin{aligned} A_i &= \log\Phi\left(\frac{\boldsymbol{x}_i\,\boldsymbol{\gamma} + \rho/\sigma\left(\log\left(y_i\right) - \boldsymbol{x}_i\boldsymbol{\beta}\right)}{\sqrt{1-\rho^2}}\right) \\ B_i &= \log\left[\phi\left(\frac{\log\left(y_i\right) - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)\right]. \end{aligned}$$

It is straightforward to verify that in the special case where $\rho = 0$ this reduces to the lognormal hurdle likelihood (eq. 17.55 in Wooldridge). Hence, ET2T can be considered more general than the lognormal hurdle model. Unfortunately, the ET2T model can be *very poorly identified* if the set of explanatory variables determining selection is the same as the set of variables determining w^*. We will return to this important issue when discussing sample selection models.

Now return to our objects of interest, i.e. $\frac{\partial E(y|\boldsymbol{x})}{\partial x_j}$. It can be shown that

$$E\left(y|\boldsymbol{x}\right)=\Phi\left(\boldsymbol{x}\boldsymbol{\gamma}+\rho\sigma\right)\exp\left(\boldsymbol{x}\boldsymbol{\beta}+.5\sigma^2\right).$$

(To obtain this expression, note first that $E\left(y|\boldsymbol{x}\right)=\Phi\left(\boldsymbol{x}\boldsymbol{\gamma}\right)E\left(y|\boldsymbol{x},y>0\right)$. Obtaining an expression for $E\left(y|\boldsymbol{x},y>0\right)$ is not entirely straightforward since we are dealing with a truncated log normal distribution - see e.g. Fact 21.72 in Söderlind[2] for details on how to proceed.) Clearly, if $\rho=0$, this reduces to

$$E\left(y|\boldsymbol{x}\right)=\Phi\left(\boldsymbol{x}\boldsymbol{\gamma}\right)\exp\left(\boldsymbol{x}\boldsymbol{\beta}+.5\sigma^2\right),$$

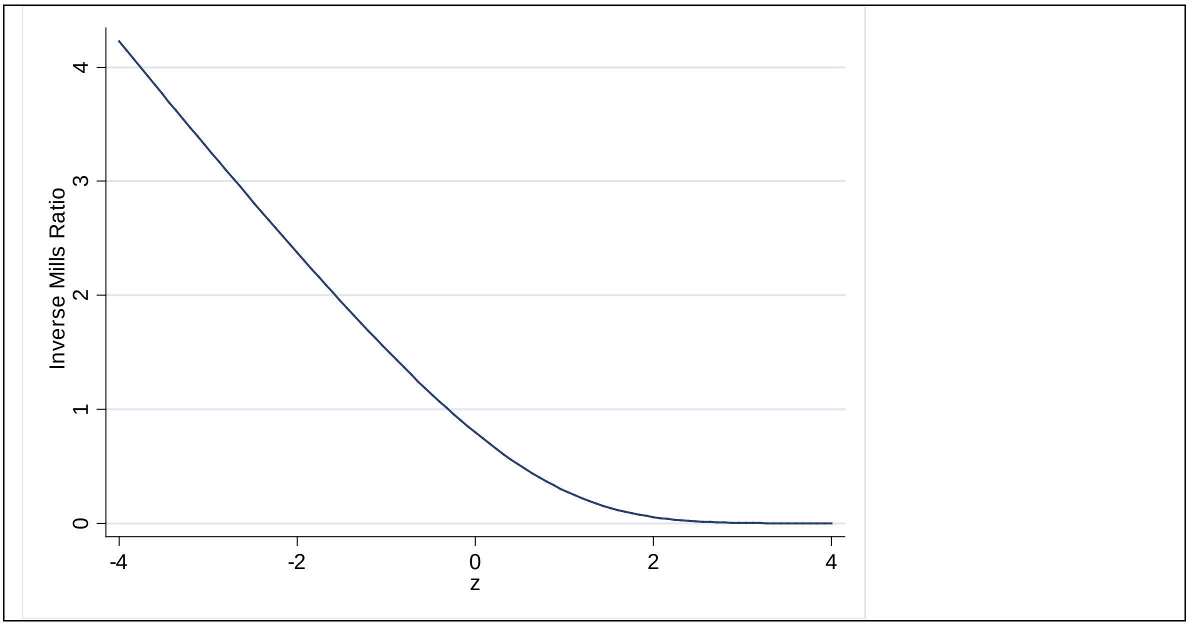
i.e. the lognormal hurdle model.

Estimation of the ET2T model can be done in Stata using the `heckman` command. Helpfully, if for whatever reason we prefer the lognormal hurdle model we can estimate this model using `heckman` with the constraint $\rho=0$ imposed. Moreover, if we use the post-estimation command `margins` we can easily obtain $\frac{\partial E(y|\boldsymbol{x})}{\partial x_j}$ with or without $\rho=0$ imposed. Applications are shown in the appendix, sections 5 and 6.

[2] http://home.datacomm.ch/paulsoderlind/Courses/OldCourses/EcmXSta.pdf

PhD Programme: Econometrics III
Department of Economics, University of Gothenburg
Appendix: Tobit regressions
Måns Söderbom

1. The inverse Mills ratio function



2. OLS and Tobit Estimation of Annual Hours Worked

This section replicates the results discussed in Example 17.2, Wooldridge (2010), pp. 678-680. It also shows how to obtain the relevant partial effects using the Stata *margins* command.

```
. use MROZ.DTA", clear

.
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        inlf |       753    .5683931    .4956295          0          1
       hours |       753    740.5764    871.3142          0       4950
     kidslt6 |       753    .2377158     .523959          0          3
     kidsge6 |       753    1.353254    1.319874          0          8
         age |       753    42.53785    8.072574         30         60
-------------+--------------------------------------------------------
        educ |       753    12.28685    2.280246          5         17
        wage |       428    4.177682    3.310282      .1282         25
     repwage |       753    1.849734    2.419887          0       9.98
      hushrs |       753    2267.271    595.5666        175       5010
      husage |       753    45.12085    8.058793         30         60
-------------+--------------------------------------------------------
     huseduc |       753    12.49137    3.020804          3         17
     huswage |       753    7.482179    4.230559      .4121     40.509
      faminc |       753    23080.59     12190.2       1500      96000
         mtr |       753    .6788632    .0834955      .4415      .9415
    motheduc |       753    9.250996    3.367468          0         17
-------------+--------------------------------------------------------
    fatheduc |       753    8.808765     3.57229          0         17
        unem |       753    8.623506    3.114934          3         14
        city |       753    .6427623    .4795042          0          1
       exper |       753    10.63081     8.06913          0         45
    nwifeinc |       753    20.12896     11.6348  -.0290575         96
-------------+--------------------------------------------------------
       lwage |       428    1.190173    .7231978  -2.054164   3.218876
     expersq |       753    178.0385    249.6308          0       2025
```

2.1 OLS results

```
. reg hours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6

      Source |       SS       df       MS              Number of obs =     753
-------------+------------------------------           F(  7,   745) =   38.50
       Model |   151647606     7  21663943.7           Prob > F      =  0.0000
    Residual |   419262118   745  562767.944           R-squared     =  0.2656
-------------+------------------------------           Adj R-squared =  0.2587
       Total |   570909724   752  759188.463           Root MSE      =  750.18

---------------------------------------------------------------------------------
          hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
       nwifeinc |  -3.446636      2.544    -1.35   0.176    -8.440898    1.547626
           educ |   28.76112   12.95459     2.22   0.027     3.329283    54.19297
          exper |   65.67251   9.962983     6.59   0.000     46.11365    85.23138
                |
c.exper#c.exper |  -.7004939   .3245501    -2.16   0.031    -1.337635   -.0633524
                |
            age |  -30.51163   4.363868    -6.99   0.000    -39.07858   -21.94469
        kidslt6 |  -442.0899    58.8466    -7.51   0.000    -557.6148    -326.565
        kidsge6 |  -32.77923   23.17622    -1.41   0.158     -78.2777    12.71924
          _cons |   1330.482   270.7846     4.91   0.000     798.8906    1862.074
---------------------------------------------------------------------------------
```

>> Since experience enters the model with a quadratic, the partial effect is 65.67 – 2*0.70*exper. The sample average of exper is 10.63, hence the average partial effect is 50.8. Notice that the average partial effect (APE) coincides with the partial effect at the average (PEA) in linear models for a variable entering with a quadratic. If we use factor variable syntax, in this case entering the quadratic term as `c.exper#c.exper` we can use margins to obtain the APE and a standard error directly:

```
. margins, dydx (*)

Average marginal effects                          Number of obs   =        753
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -3.446636      2.544    -1.35   0.175    -8.432784    1.539513
        educ |   28.76112   12.95459     2.22   0.026       3.3706    54.15165
       exper |   50.77888   4.448188    11.42   0.000     42.06059    59.49716
         age |  -30.51163   4.363868    -6.99   0.000    -39.06466   -21.95861
     kidslt6 |  -442.0899    58.8466    -7.51   0.000    -557.4271   -326.7527
     kidsge6 |  -32.77923   23.17622    -1.41   0.157    -78.20378    12.64533
------------------------------------------------------------------------------
```

lower limit of y is zero

2. Tobit results

```
. tobit hours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6, ll(0)

Tobit regression                                  Number of obs   =        753
                                                  LR chi2(7)      =     271.59
                                                  Prob > chi2     =     0.0000
Log likelihood = -3819.0946                       Pseudo R2       =     0.0343

------------------------------------------------------------------------------
          hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
       nwifeinc |  -8.814243   4.459096    -1.98   0.048    -17.56811   -.0603724
           educ |   80.64561   21.58322     3.74   0.000     38.27453    123.0167
          exper |   131.5643   17.27938     7.61   0.000     97.64231    165.4863
                |
c.exper#c.exper |  -1.864158    .5376615    -3.47   0.001    -2.919667   -.8086479
                |
            age |  -54.40501   7.418496    -7.33   0.000    -68.96862    -39.8414
        kidslt6 |  -894.0217   111.8779    -7.99   0.000    -1113.655   -674.3887
        kidsge6 |    -16.218   38.64136    -0.42   0.675    -92.07675    59.64075
          _cons |   965.3053   446.4358     2.16   0.031     88.88528    1841.725
----------------+-------------------------------------------------------------
         /sigma |   1122.022   41.57903                      1040.396    1203.647
------------------------------------------------------------------------------
  Obs. summary:         325  left-censored observations at hours<=0
                        428     uncensored observations
                          0 right-censored observations
```

- Tobit coefficient estimates are the same sign as the corresponding OLS estimates.
- Similar statistical significance.
- Tobit coefficients are much higher than their OLS coefficients – but direct comparisons are misleading. Why?

2.1 Partial effects on E(y|x)

- To obtain tobit-based partial effects that are comparable to those implied by OLS, we look at the effects of changing x-variables on E(y|x). As shown in the lecture notes above, for continuous, non-interacted, variables, the formula for the partial effect looks like this:

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} = \beta_j \Phi(\boldsymbol{x\beta}/\sigma).$$

The Stata margins command computes estimates of these partial effects and their associated standard errors:

```
. margins, dydx (*) predict(ystar(0,.))
```

```
Average marginal effects                          Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours*|hours>0), predict(ystar(0,.))
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -5.188622    2.62141    -1.98   0.048    -10.32649   -.0507525
        educ |   47.47311    12.6214     3.76   0.000     22.73562    72.21061
       exper |   48.79312   3.587271    13.60   0.000      41.7622    55.82404
         age |  -32.02624   4.292112    -7.46   0.000    -40.43862   -23.61385
     kidslt6 |  -526.2779   64.70622    -8.13   0.000    -653.0997    -399.456
     kidsge6 |   -9.54694   22.75225    -0.42   0.675    -54.14054    35.04665
------------------------------------------------------------------------------
```

We now have estimates that are directly comparable to the OLS results above. We see, for example, that one more year of education is estimated to result in 47.5 more hours of work per year, on average, based on the Tobit results, while the OLS estimate of the partial effect is just 28.8. Moreover, the levels of statistical significance differ somewhat across the estimators.

If for some reason we prefer partial effects evaluated at the average, these can be obtained as follows:

```
. margins, dydx (*) predict(ystar(0,.)) atmeans

Conditional marginal effects                      Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours*|hours>0), predict(ystar(0,.))
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6
at           : nwifeinc        =    20.12896 (mean)
               educ            =    12.28685 (mean)
               exper           =    10.63081 (mean)
               age             =    42.53785 (mean)
               kidslt6         =    .2377158 (mean)
               kidsge6         =    1.353254 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -5.687381   2.877882    -1.98   0.048    -11.32793   -.0468357
        educ |   52.03649   13.82013     3.77   0.000     24.94954    79.12345
       exper |   59.31728   5.694847    10.42   0.000     48.15558    70.47897
         age |  -35.10478   4.669466    -7.52   0.000    -44.25676   -25.95279
     kidslt6 |  -576.8666   70.92986    -8.13   0.000    -715.8866   -437.8466
     kidsge6 |  -10.46464   24.93972    -0.42   0.675    -59.34561    38.41632
------------------------------------------------------------------------------
```

For some reason, these turn out to be larger, in absolute terms, than the average partial effects. Given that the APE are more straightforward to interpret, I would have more faith in those.

2.2 Partial effects on `E(y|x,y>0)`

Finally, we look at estimates of average partial effects on the expected value of y given that y is positive. Such estimates can be obtained by modifying the *margins* syntax as can be seen below. In addition to APE, I also show partial effects evaluated at sample averages.

```
. margins, dydx(*) predict(e(0,.))

Average marginal effects                          Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -3.968784   2.007582    -1.98   0.048    -7.903573   -.0339953
        educ |   36.31225   9.703038     3.74   0.000     17.29465    55.32986
       exper |    37.5935   2.965955    12.68   0.000     31.78034    43.40667
         age |  -24.49691   3.362492    -7.29   0.000    -31.08728   -17.90655
     kidslt6 |  -402.5507   50.74877    -7.93   0.000    -502.0164   -303.0849
     kidsge6 |  -7.302468   17.40427    -0.42   0.675     -41.4142    26.80927
------------------------------------------------------------------------------

. margins, dydx(*) predict(e(0,.)) atmeans

Conditional marginal effects                      Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6
at           : nwifeinc        =    20.12896 (mean)
               educ            =    12.28685 (mean)
               exper           =    10.63081 (mean)
               age             =    42.53785 (mean)
               kidslt6         =    .2377158 (mean)
               kidsge6         =    1.353254 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -3.987413   2.017641    -1.98   0.048    -7.941917   -.0329086
        educ |   36.48269   9.689266     3.77   0.000     17.49208    55.47331
       exper |   41.58724   3.988059    10.43   0.000     33.77079    49.40369
         age |   -24.6119   3.273616    -7.52   0.000    -31.02807   -18.19573
     kidslt6 |  -404.4401   49.72179    -8.13   0.000    -501.8931   -306.9872
     kidsge6 |  -7.336744   17.48515    -0.42   0.675    -41.60701    26.93352
------------------------------------------------------------------------------
```

3. Testing exogeneity of other income in the hours equation

Here is an illustration of the two-step procedure proposed by Smith and Blundell (1986):

Step 1

```
. reg nwifeinc huseduc educ exper c.exper#c.exper age kidslt6 kidsge6

      Source |       SS       df       MS              Number of obs =     753
-------------+------------------------------           F(  7,   745) =   27.13
       Model |  20676.7705     7  2953.82436           Prob > F      =  0.0000
    Residual |  81120.3451   745  108.886369           R-squared     =  0.2031
-------------+------------------------------           Adj R-squared =  0.1956
       Total |  101797.116   752  135.368505           Root MSE      =  10.435

--------------------------------------------------------------------------------
       nwifeinc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+---------------------------------------------------------------
        huseduc |   1.178155   .1609449     7.32   0.000     .8621956    1.494115
           educ |   .6746951   .2136829     3.16   0.002     .2552029    1.094187
          exper |  -.3129877   .1382549    -2.26   0.024    -.5844034   -.0415721
                |
c.exper#c.exper |  -.0004776   .0045196    -0.11   0.916    -.0093501     .008395
                |
            age |   .3401521   .0597084     5.70   0.000     .2229354    .4573687
        kidslt6 |   .8262719   .8183785     1.01   0.313    -.7803305    2.432874
        kidsge6 |   .4355289   .3219888     1.35   0.177    -.1965845    1.067642
          _cons |  -14.72048   3.787326    -3.89   0.000    -22.15559   -7.285383
--------------------------------------------------------------------------------

. predict v2hat, res

. tobit hours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6 v2hat, ll(0)

Tobit regression                                  Number of obs   =        753
                                                  LR chi2(8)      =     273.76
                                                  Prob > chi2     =     0.0000
Log likelihood = -3818.0118                       Pseudo R2       =     0.0346

--------------------------------------------------------------------------------
          hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+---------------------------------------------------------------
       nwifeinc |  -31.48215    16.0376    -1.96   0.050    -62.96641    .0021189
           educ |   116.7814   32.75978     3.56   0.000     52.46891    181.0939
          exper |   124.3488   17.87502     6.96   0.000     89.25736    159.4402
                |
c.exper#c.exper |    -1.8972   .5371614    -3.53   0.000     -2.95173   -.8426702
                |
            age |  -46.89244   8.957672    -5.23   0.000    -64.47773   -29.30716
        kidslt6 |  -867.9131   112.9024    -7.69   0.000    -1089.558   -646.2684
        kidsge6 |   -6.32605   39.16561    -0.16   0.872    -83.21414    70.56204
          v2hat |   24.41832   16.58452     1.47   0.141    -8.139637    56.97628
          _cons |   722.1032    475.689     1.52   0.129    -211.7472    1655.954
----------------+---------------------------------------------------------------
         /sigma |   1119.844   41.49319                      1038.387    1201.302
--------------------------------------------------------------------------------
  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
                         0 right-censored observations
```

3.1 Extension: Nonlinear control function

```
. ge nlv2hat=v2hat^2-r(sd)^2

.
. tobit hours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6 v2hat
nlv2hat, ll(0)

Tobit regression                                  Number of obs   =        753
                                                  LR chi2(9)      =     273.88
                                                  Prob > chi2     =     0.0000
Log likelihood = -3817.9538                       Pseudo R2       =     0.0346

------------------------------------------------------------------------------
          hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
       nwifeinc |  -31.50841   16.04047    -1.96   0.050    -62.99839   -.018436
           educ |   115.6293   32.93191     3.51   0.000     50.97881   180.2799
          exper |   124.7863   17.92539     6.96   0.000     89.59591   159.9766
                |
c.exper#c.exper |  -1.904637   .5377187    -3.54   0.000    -2.960264  -.8490106
                |
            age |  -47.09241    8.97805    -5.25   0.000    -64.71774  -29.46708
        kidslt6 |  -871.1396   113.3374    -7.69   0.000    -1093.639  -648.6403
        kidsge6 |  -6.472582   39.16835    -0.17   0.869    -83.36623   70.42106
          v2hat |   23.05771   17.05897     1.35   0.177    -10.43173   56.54715
        nlv2hat |   .0609493    .178337     0.34   0.733    -.2891544   .4110529
          _cons |   742.8585   479.5567     1.55   0.122    -198.5868   1684.304
----------------+-------------------------------------------------------------
         /sigma |   1119.887   41.49608                      1038.423    1201.35
------------------------------------------------------------------------------
  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
                         0 right-censored observations

. test v2hat nlv2hat

 ( 1)  [model]v2hat = 0
 ( 2)  [model]nlv2hat = 0

       F(  2,   744) =    1.14
            Prob > F =    0.3194
```

4. The truncated normal hurdle model

```
. probit anyhours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6

Probit regression                                 Number of obs   =        753
                                                  LR chi2(7)      =     227.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -401.30219                       Pseudo R2       =     0.2206

------------------------------------------------------------------------------
       anyhours |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
       nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
           educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
          exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
                |
c.exper#c.exper |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
                |
            age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
        kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
        kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
          _cons |   .2700768    .508593     0.53   0.595    -.7267473    1.266901
------------------------------------------------------------------------------

. truncreg hours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6 if
anyhours==1, ll(0)

Truncated regression
Limit:   lower =          0                             Number of obs =    428
         upper =       +inf                             Wald chi2(7)  =  59.05
Log likelihood = -3390.6476                             Prob > chi2   = 0.0000

------------------------------------------------------------------------------
          hours |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
       nwifeinc |   .1534399   5.164279     0.03   0.976    -9.968361    10.27524
           educ |  -29.85254   22.83935    -1.31   0.191    -74.61684    14.91176
          exper |   72.62273   21.23628     3.42   0.001     31.00039    114.2451
                |
c.exper#c.exper |  -.9439967   .6090283    -1.55   0.121     -2.13767    .2496769
                |
            age |  -27.44381   8.293458    -3.31   0.001    -43.69869   -11.18893
        kidslt6 |  -484.7109   153.7881    -3.15   0.002      -786.13   -183.2918
        kidsge6 |  -102.6574   43.54347    -2.36   0.018    -188.0011   -17.31379
          _cons |   2123.516   483.2649     4.39   0.000     1176.334    3070.697
----------------+-------------------------------------------------------------
         /sigma |    850.766   43.80097    19.42   0.000     764.9177    936.6143
------------------------------------------------------------------------------
```

>> Judging from these results, does the type I tobit model considered earlier appear correctly specified? If not, why not?

5. Lognormal hurdle model

```
. probit anyhours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6

Probit regression                                 Number of obs   =        753
                                                  LR chi2(7)      =     227.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -401.30219                       Pseudo R2       =     0.2206

------------------------------------------------------------------------------
       anyhours |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
       nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
           educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
          exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
                |
c.exper#c.exper |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
                |
            age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
        kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
        kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
          _cons |   .2700768    .508593     0.53   0.595    -.7267473    1.266901
------------------------------------------------------------------------------

. ge lhours=ln(hours)

. reg lhours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6 if anyhours==1

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  7,   420) =   11.90
       Model |  66.3633428     7  9.48047755           Prob > F      =  0.0000
    Residual |  334.513835   420  .796461511           R-squared     =  0.1655
-------------+------------------------------           Adj R-squared =  0.1516
       Total |  400.877178   427   .93882243           Root MSE      =  .89245

------------------------------------------------------------------------------
         lhours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
       nwifeinc |  -.0019676   .0044436    -0.44   0.658    -.0107021    .0067668
           educ |  -.0385626   .0202098    -1.91   0.057    -.0782876    .0011624
          exper |    .073237   .0179004     4.09   0.000     .0380514    .1084225
                |
c.exper#c.exper |   -.001233   .0005378    -2.29   0.022    -.0022902   -.0001759
                |
            age |  -.0236706    .007248    -3.27   0.001    -.0379175   -.0094237
        kidslt6 |   -.585202   .1186066    -4.93   0.000    -.8183386   -.3520654
        kidsge6 |  -.0694175   .0373355    -1.86   0.064    -.1428053    .0039703
          _cons |   7.896267   .4260789    18.53   0.000     7.058755     8.73378
------------------------------------------------------------------------------
. disp e(ll)

-554.56647

(this is the log likelihood value associated with the OLS estimator)
```

I can obtain exactly these results using the heckman command with rho=0 imposed:

```
constraint 1 [athrho]_b[_cons] = 0

heckman lhours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6,
select(nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6)  constraint(1)


Heckman selection model                         Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(7)       =     84.91
Log likelihood = -955.8687                      Prob > chi2        =    0.0000

 ( 1)  [athrho]_cons = 0
---------------------------------------------------------------------------------
         lhours |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
lhours          |
       nwifeinc |  -.0019676   .0044019    -0.45   0.655    -.0105951    .0066599
           educ |  -.0385626     .02002    -1.93   0.054    -.0778011     .000676
          exper |    .073237   .0177323     4.13   0.000     .0384822    .1079917
                |
c.exper#c.exper |   -.001233   .0005328    -2.31   0.021    -.0022773   -.0001888
                |
            age |  -.0236706   .0071799    -3.30   0.001     -.037743   -.0095981
        kidslt6 |   -.585202   .1174929    -4.98   0.000    -.8154839   -.3549201
        kidsge6 |  -.0694175    .036985    -1.88   0.061    -.1419067    .0030717
          _cons |   7.896267   .4220781    18.71   0.000      7.06901    8.723525
----------------+----------------------------------------------------------------
select          |
       nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
           educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
          exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
                |
c.exper#c.exper |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
                |
            age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
        kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
        kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
          _cons |   .2700768    .508593     0.53   0.595    -.7267473    1.266901
----------------+----------------------------------------------------------------
        /athrho |          0  (constrained)
       /lnsigma |  -.1232225   .0341793    -3.61   0.000    -.1902127   -.0562323
----------------+----------------------------------------------------------------
            rho |          0  (omitted)
          sigma |   .8840669   .0302168                      .8267833    .9453195
         lambda |          0  (omitted)
---------------------------------------------------------------------------------
Wald test of indep. eqns. (rho = 0): chi2(1) =        .   Prob > chi2 =      .
```

```
and I can now obtain average partial (marginal) effects for the lognormal hurdle
model:

. margins, dydx(*) expression(  normal(predict(xbsel) +
exp([lnsigma]_cons)*tanh([athrho]_cons) ) * exp(predict(xb) +
.5*exp(2*[lnsigma]_cons))  )

Average marginal effects                          Number of obs   =        753
Model VCE    : OIM

Expression   : normal(predict(xbsel) + exp([lnsigma]_cons)*tanh([athrho]_cons) ) *
exp(predict(xb) + .5*exp(2*[lnsigma]_cons))
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -6.339119    4.23414    -1.50   0.134    -14.63788    1.959643
        educ |   17.10526   19.35311     0.88   0.377    -20.82614    55.03666
       exper |   61.03048   6.121516     9.97   0.000     49.03253    73.02843
         age |  -40.86435   7.183075    -5.69   0.000    -54.94292   -26.78578
     kidslt6 |  -841.2569   120.3566    -6.99   0.000    -1077.151   -605.3624
     kidsge6 |  -46.18975   36.40202    -1.27   0.204    -117.5364     25.1569
------------------------------------------------------------------------------
```

There are some interesting differences compared to the partial effects obtained after OLS and tobit type I estimation, notice for example that education is no longer statistically significant.

6. Exponential Type II Tobit

```
. heckman lhours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6,
select(nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6)

Heckman selection model                         Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(7)       =     35.50
Log likelihood = -938.8208                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         lhours |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
lhours          |
       nwifeinc |   .0066597   .0050147     1.33   0.184    -.0031689    .0164882
           educ |  -.1193085   .0242235    -4.93   0.000    -.1667858   -.0718313
          exper |  -.0334099   .0204429    -1.63   0.102    -.0734773    .0066574
                |
c.exper#c.exper |   .0006032   .0006178     0.98   0.329    -.0006077    .0018141
                |
            age |   .0142754   .0084906     1.68   0.093    -.0023659    .0309167
        kidslt6 |   .2080079   .1338148     1.55   0.120    -.0542643    .4702801
        kidsge6 |  -.0920299   .0433138    -2.12   0.034    -.1769235   -.0071364
          _cons |   8.670736    .498793    17.38   0.000      7.69312    9.648352
----------------+-------------------------------------------------------------
select          |
       nwifeinc |  -.0096823   .0043273    -2.24   0.025    -.0181637    -.001201
           educ |    .119528   .0217542     5.49   0.000     .0768906    .1621654
          exper |   .0826696   .0170277     4.86   0.000      .049296    .1160433
                |
c.exper#c.exper |  -.0012896   .0005369    -2.40   0.016     -.002342   -.0002372
                |
            age |  -.0330806   .0075921    -4.36   0.000    -.0479609   -.0182003
        kidslt6 |  -.5040406   .1074788    -4.69   0.000    -.7146951    -.293386
        kidsge6 |   .0698201   .0387332     1.80   0.071    -.0060955    .1457357
          _cons |  -.3656166   .4476569    -0.82   0.414    -1.243008    .5117748
----------------+-------------------------------------------------------------
        /athrho |  -2.131542    .174212   -12.24   0.000    -2.472991   -1.790093
       /lnsigma |   .1895611   .0419657     4.52   0.000     .1073099    .2718123
----------------+-------------------------------------------------------------
            rho |  -.9722333   .0095403                     -.9858766   -.9457704
          sigma |   1.208719   .0507247                      1.113279    1.312341
         lambda |  -1.175157   .0560391                     -1.284991   -1.065322
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =    34.10   Prob > chi2 = 0.0000
```

Very large negative estimate of rho, which is not really credible. Probably a sign of the identification problem that often arises when the same regressors are used in the two equations. We probably need an exclusion restriction, i.e. a variable entering the selection equation but not the hours equation. Such an exclusion restriction may be hard to justify, however.

Average partial (marginal) effects on the next page. Education is now significant again. How proceed?

```
heckman lhours nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6,
select(nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6)

margins, dydx(*) expression(  normal(predict(xbsel) +
exp([lnsigma]_cons)*tanh([athrho]_cons) ) * exp(predict(xb) +
.5*exp(2*[lnsigma]_cons))  )


Expression   : normal(predict(xbsel) + exp([lnsigma]_cons)*tanh([athrho]_cons) ) *
exp(predict(xb) + .5*exp(2*[lnsigma]_cons))
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -4.338013   2.551946    -1.70   0.089    -9.339735    .6637086
        educ |   29.39153    11.8269     2.49   0.013     6.211223    52.57183
       exper |   33.96181   4.800083     7.08   0.000     24.55382     43.3698
         age |  -20.34328   5.020793    -4.05   0.000    -30.18385    -10.5027
     kidslt6 |  -316.1543   80.70819    -3.92   0.000    -474.3395   -157.9692
     kidsge6 |   2.619054   24.69581     0.11   0.916    -45.78385    51.02196
------------------------------------------------------------------------------
```

CHAPTER 2:

CENSORING AND SAMPLE SELECTION

1. Introduction

In this lecture we discuss two types of missing data problems: one posed by censoring, the other posed by truncation. The lecture is based on the material presented in Wooldridge (2010), chapter 19.1-6, 19.9.

2. Censored and Truncated Models

In a previous lecture we covered in some detail the tobit model as applied to corner solution models. Recall that a corner solution is an actual economic outcome, e.g. zero expenditure on health by a household in a given period. In this section we discuss briefly two close cousins of the corner solution model, namely the censored regression model and the truncated regression model. The good news is that the econometric techniques used for censored and truncated dependent variables are very similar to what we have already studied.

2.1. Data censoring

In contrast to corner solutions, censoring is essentially a **data problem**. Censoring occurs, for example, if whenever the dependent variable y exceeds some upper threshold c the actual value of y gets **recorded** as equal to c, rather than the true value. Of course, censoring may also occur at the lower end of the dependent variable.

Top coding in income surveys is the most common example of censoring, however. Such surveys are sometimes designed so that people with incomes higher than some upper threshold, say $\$500,000$, are allowed to respond "more than $\$500,000$". In contrast, for people with incomes lower than $\$500,000$ the actual income gets recorded. If we want to run a regression explaining income based on such data, we clearly need to deal with the top coding. A reasonable way of writing down the model might be

$$y^* = \boldsymbol{x}\boldsymbol{\beta} + u,$$

$$y = \min(y^*, c),$$

where y^* is **actual** income (which is not fully observed due to the censoring), u is a normally distributed and homoskedastic error term, and y is measured income, which in this example is bounded above at $c = \$500,000$ due to the censoring produced by the design of the survey.

You now see that the censored regression is very similar to the corner solution model. In fact, if $c = 0$ and this is a lower bound, the econometric model for corner solution models and censored regressions coincide: in both cases we would have the tobit model. If the threshold c is not zero and/or represents an upper rather than a lower bound on what is observed, then we still use tobit but with a simple (and uninteresting) adjustment of the log likelihood.

The only substantive difference between censored regressions models and corner solution models lies in the **interpretation of the results**. That is, suppose we have two models:

- Model 1: the dependent variable is a corner solution variable, with the corner at zero
- Model 2: the dependent variable is censored below at zero.

We could use exactly the same econometric estimator for both models, i.e. the tobit model.

- In the corner solution model we are probably mainly interested in how the expected value of the observed dependent variable varies with the explanatory variable(s). This means we should look at $E(y|\boldsymbol{x}, y > 0)$ or $E(y|\boldsymbol{x})$, and we have seen how to obtain the relevant marginal effects.
- However, for the censored regression model we are mostly interested in learning how the expected value of the **unobserved and censored** variable y^* varies with the explanatory variable(s), i.e. $E(y^*|\boldsymbol{x})$:
$$E(y^*|\boldsymbol{x}) = \boldsymbol{x\beta},$$
and so the relevant partial effect with respect to x_j is simply β_j.

One field in which censored regression models are very common is in the econometric analysis of **duration data.** Duration is the time that elapses between the 'beginning' and the 'end' of some specified state. The most common example is unemployment duration, where the 'beginning' is the day the individual becomes unemployed and the 'end' is when the same individual gets a new job. Data on durations are often censored, either to the right (common) or to the left (not so common) or both (even less common). Right censoring means that we don't know from the data when a certain duration ended; left censoring means that we don't know when it began. I will not cover duration data as part of this course, but you can find an old lecture introducing duration data models on my web page.

2.2. Truncated regression models

A truncated regression model is similar to a censored regression model, but there is one important difference:

- If the dependent variable is truncated we do not observe **any** information about a certain segment in the population.

- In other words, we do not have a representative (random) sample from the population. This can happen if a survey targets a sub-group of the population. For instance when surveying firms in developing countries, the World Bank often excludes firms with less than 10 employees. Clearly if we are modelling employment based on such data we need to recognize the fact that firms with less than 10 employees are not covered in our dataset.

- Alternatively, it could be that we target poor individuals, and so exclude everyone with an income higher than some upper threshold c.

- The standard truncated regression model is written

$$y = \boldsymbol{x}\boldsymbol{\beta} + u,$$

  where the error term u is assumed normally distributed, homoskedastic and uncorrelated with $\boldsymbol{x}$ (the latter assumption can be relaxed if we have instruments). Suppose that all observations for which $y_i > c$ are excluded from the sample. Our objective is to estimate the parameter $\boldsymbol{\beta}$.

- See example in appendix, Section 1.

It is clear from the example in the appendix that ignoring the truncation leads to substantial downward bias in the estimate of $\boldsymbol{\beta}$. Fortunately, we can correct this bias fairly easily, by using the normality assumption in combination with the information about the threshold. The density of y, conditional on $\boldsymbol{x}$ and y observed, takes a familiar form:

$$f\left(y|\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\gamma}\right)=\left[\frac{\phi\left(\left(y-\boldsymbol{x\beta}\right)/\sigma\right)/\sigma}{\Phi\left(\boldsymbol{x\beta}/\sigma\right)}\right],$$

and the individual log likelihood contribution is

$$\ln L_i=\ln\left[\phi\left(\left(y_i-\boldsymbol{x}_i\boldsymbol{\beta}\right)/\sigma\right)/\sigma\right]-\ln\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma\right)$$

The conditional expected value of y is also of a familiar form:

$$E\left(y|y>0,\boldsymbol{x}\right)=\boldsymbol{x\beta}+\sigma_u\lambda\left(\boldsymbol{x\beta}/\sigma_u\right)$$

In Stata we can implement this model using the **truncreg** command (see appendix).

3. Sample Selection Bias

Up to this point we have assumed the availability of a random sample from the underlying population. In practice, however, samples may not be random. In particular, samples are sometimes **truncated** by economic variables.

Example: Suppose you want to study how education impacts on the wage an individual *could potentially* earn in the labour market - i.e. the wage offer. Your plan is to run a regression in which log wage is the dependent variable and education is (let's say) the only explanatory variable. You are primarily interested in the coefficient β_1 on education. Suppose in the population, education is uncorrelated with the error term u_i - i.e. it is exogenous (this can be relaxed, but the model would get more complicated as a result). Thus, with access to a random sample, OLS would be the best estimator.

Suppose your sample contains a non-negligible proportion of unemployed individuals. For these individuals, there is *no information* on earnings, and so the corresponding observations cannot be used when estimating the wage equation (missing values for the dependent variable). Thus you're looking at having to estimate the earnings equation based on a non-random sample - what we shall refer to as a **selected sample**. Can the parameters of the wage offer equation - most importantly β - be estimated without bias based on the selected sample?

The general answer to that question is: It depends! Whenever we have a selected (non-random) sample, it is important to be clear on two things:

- Circumstances under which the OLS estimator (or some other estimator ignoring selection) applied on the selected sample will be suffer from bias - specifically **selectivity bias** - and circumstances when it won't; and
- If there is selectivity bias: how to obtain estimates that are not biased by sample selection.

The most common model accommodating the above sample selection mechanism is one in which the equation of interest (sometimes referred to as the 'structural equation' or the 'primary equation') is written as

$$y_1 = \boldsymbol{x}_1 \boldsymbol{\beta}_1 + u_1, \tag{3.1}$$

and we have a separate model determining selection as follows:

$$y_2 = 1\left[\boldsymbol{x}\boldsymbol{\delta}_2 + v_2 > 0\right].$$

Assumptions:

- $(\boldsymbol{x}, y_2)$ are always observed, but y_1 is observed only when $y_2 = 1$ (this assumption emphasizes the sample selection nature of the problem)
- (u_1, v_2) is independent of $\boldsymbol{x}$ with zero mean (exogeneity)

- v_2˜$Normal\,(0,1)$ (note: explicit distributional assumption; needed to derive a conditional expectation given selection, more on this below)

- $E\,(u_1|v_2) = \gamma_1 v_2$ (linearity; holds e.g. under bivariate normality of (u_1, v_2); note that, since $Var\,(v_2) = 1$, γ_1 is the covariance between u_1 and v_2.)

3.1. When will there be selection bias, and what can be done about it?

The fundamental issue to consider when worrying about sample selection bias is **why** some individuals will not be included in the sample. As we shall see, sample selection bias can be viewed as a special case of **endogeneity bias**, arising when the selection process **generates** endogeneity in the selected sub-sample.

In our model sample selection bias arises when the error term in the selection equation (i.e. v_2) is *correlated* with the error term in the primary equation (i.e. u_1), i.e. whenever $\gamma_1 \neq 0$.

To see this, we will derive the expression for $E\,(y_1|\boldsymbol{x}, y_2 = 1)$, i.e. the expectation of the outcome variable conditional on observable $\boldsymbol{x}$ and selection into the sample, $y_2 = 1$.

We begin by deriving $E\,(y_1|\boldsymbol{x}, v_2)$:

$$\begin{aligned} E\,(y_1|\boldsymbol{x}, v_2) &= \boldsymbol{x}_1\boldsymbol{\beta}_1 + E\,(u_1|\boldsymbol{x}, v_2) \\ E\,(y_1|\boldsymbol{x}, v_2) &= \boldsymbol{x}_1\boldsymbol{\beta}_1 + E\,(u_1|v_2) \\ E\,(y_1|\boldsymbol{x}, v_2) &= \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1 v_2, \end{aligned} \tag{3.2}$$

where the assumption that (u_1, v_2) is independent of $\boldsymbol{x}$ enables us to go from the first to the second line; and the linearity assumption for $E\,(u_1|v_2)$ enables us to go from the second to the third line.

It is now clear that, if and only if $\gamma_1 = 0$,

$$E\,(y_1|\boldsymbol{x}, v_2) = E\,(y_1|\boldsymbol{x}) = E\,(y_1|\boldsymbol{x}_1) = \boldsymbol{x}_1\boldsymbol{\beta}_1,$$

i.e. in this case there is no sample selection problem. But, if $\gamma_1 \neq 0$, $E\,(y_1|\boldsymbol{x}, v_2) \neq \boldsymbol{x}_1\boldsymbol{\beta}_1$.

Since v_2 is not observable, eq (3.2) is not directly usable in applied work (since we can't condition on unobservables when running a regression). To obtain an expression for the expected value of y_1 conditional on observables $\boldsymbol{x}$ and the actual selection outcome y_2, we make use of the law of iterated expectations and write:

$$E\left(y_1|\boldsymbol{x}, y_2 = 1\right) = E\left[E\left(y_1|\boldsymbol{x}, v_2\right)|\boldsymbol{x}, y_2 = 1\right],$$

Hence, using (3.2) we obtain

$$\begin{aligned}
E\left(y_1|\boldsymbol{x}, y_2 = 1\right) &= E\left[\left(\boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1 v_2\right)|\boldsymbol{x}, y_2 = 1\right], \\
E\left(y_1|\boldsymbol{x}, y_2 = 1\right) &= \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1 E\left(v_2|\boldsymbol{x}, y_2 = 1\right), \\
E\left(y_1|\boldsymbol{x}, y_2 = 1\right) &= \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1 h\left(\boldsymbol{x}, y_2 = 1\right),
\end{aligned}$$

where $h\left(\boldsymbol{x}, y_2 = 1\right) = E\left(v_2|\boldsymbol{x}, y_2 = 1\right)$ is a function.

The next challenge is to find $h\left(\boldsymbol{x}, y_2 = 1\right)$. Our model and assumptions imply

$$E\left(v_2|\boldsymbol{x}, y_2 = 1\right) = E\left(v_2|v_2 \geq -\boldsymbol{x}\boldsymbol{\delta}_2\right).$$

Now is the time to use our 'useful result' introduced in a previous lecture:

$$E\left(e|e > c\right) = \frac{\phi\left(c\right)}{1 - \Phi\left(c\right)}, \tag{3.3}$$

where e follows a standard normal distribution, c is a constant, ϕ denotes the standard normal probability density function, and Φ is the standard normal cumulative density function.

Thus

$$\begin{aligned}
E\left(v_2|v_2 \geq -\boldsymbol{x}\boldsymbol{\delta}_2\right) &= \frac{\phi\left(-\boldsymbol{x}\boldsymbol{\delta}_2\right)}{1 - \Phi\left(-\boldsymbol{x}\boldsymbol{\delta}_2\right)} \\
E\left(v_2|v_2 \geq -\boldsymbol{x}\boldsymbol{\delta}_2\right) &= \frac{\phi\left(\boldsymbol{x}\boldsymbol{\delta}_2\right)}{\Phi\left(\boldsymbol{x}\boldsymbol{\delta}_2\right)} \equiv \lambda\left(\boldsymbol{x}\boldsymbol{\delta}_2\right),
\end{aligned}$$

where $\lambda(\cdot)$ is the inverse Mills ratio (see Section 2 in the appendix for a derivation of the inverse Mills ratio).

We now have a *fully parametric expression* for the expected value of y_i, conditional on observable variables $\boldsymbol{w}_i$, and selection into the sample ($z_i = 1$):

$$E(y_1|\boldsymbol{x}, y_2 = 1) = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda(\boldsymbol{x}\boldsymbol{\delta}_2).$$

This equation tells us that the expected value of y_1, given $\boldsymbol{x}_1$ and observability of y_1 (i.e. $y_2 = 1$) is equal to $\boldsymbol{x}_1\boldsymbol{\beta}_1$, **plus** an additional term which is the product of the covariance of the error terms γ_1 and the inverse Mills ratio evaluated at $\boldsymbol{x}\boldsymbol{\delta}_2$. This equation makes it clear that an OLS regression of y_1 on $\boldsymbol{x}_1$ using the selected sample omits the term $\lambda(\boldsymbol{x}\boldsymbol{\delta}_2)$ and generally leads to inconsistent estimation of $\boldsymbol{\beta}_1$.

3.1.1. Exogenous sample selection: $E(u_1|v_2) = 0$

However, if the unobservables determining selection are mean-independent of the unobservables determining the outcome variable of interest, so that $E(u_1|v_2) = 0$, there is no problem - we then say that sample selection is **exogenous**. Then we can estimate the main equation of interest by means of OLS, since

$$E(y_1|\boldsymbol{x}, y_2 = 1) = \boldsymbol{x}_1\boldsymbol{\beta}_1,$$

hence

$$y_1 = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \varsigma_i,$$

where ς_i is a mean-zero error term that is uncorrelated with $\boldsymbol{x}_1$ in the selected sample (recall we assume exogeneity in the population).

Illustrations:

- Suppose sample selection is randomized (or as good as randomized). Imagine an urn containing a lots of balls, where 20% of the balls are red and 80% are black, and imagine participation in the sample depends on the draw from this urn: black ball, and you're in; red ball and you're not. In this case sample selection is independent of **all** other (observable and unobservable) factors (indeed $\boldsymbol{\delta}_2 = 0$). Sample selection is thus exogenous.

- Suppose the variables in the $\boldsymbol{x}$-vector affect the likelihood of selection (i.e. $\boldsymbol{\delta}_2 \neq 0$). Hence individuals with certain observable characteristics are more likely to be included in the sample than others. Still, we've assumed $\boldsymbol{x}$ to be *independent* of the error term in the main equation, u_1, and so sample selection remains **exogenous**. In this case also - no problem.

3.1.2. An example

Based on these insights, let's now think about estimating the following simple wage equation based on a selected sample.

$$\ln w_i = \beta_0 + \beta_1 educ_i + \varepsilon_i,$$

- Always when worrying about endogeneity, you need to be clear on the underlying mechanisms. So begin by asking yourself: What factors are likely to go into the error term ε_i in the wage equation? Clearly individuals with the same levels of education can obtain very different wages in the labour market, and given how we have written the model it follows by definition that the error term ε_i is the source of such wage differences. To keep the example simple, suppose I've convinced myself that the (true) error term ε_i consists of two parts:

$$\varepsilon_i = \theta_1 m_i + e_i,$$

  where m_i is personal 'motivation', which is unobserved (note!) and assumed uncorrelated with education in the population (clearly a debatable assumption, but let's keep things simple), θ_1 is a positive parameter, and e_i reflects the remaining source of variation in wages. Suppose for simplicity that e_i is independent of all variables except wages.

- I now know that the OLS estimator will be biased if the error term in the earnings equation ε_i is correlated with the error term in the selection equation. Let's now relate this insight to economics, sticking to our example. Since motivation (m_i) is (assumed) the only economically interesting part of ε_i, I thus need to ask myself: Is it reasonable to assume that motivation is uncorrelated with education **in the selected sample**? For now, maintain the assumption that motivation and education are uncorrelated in the population - hence had there been no sample selection, education would have been exogenous and OLS would have been fine.

- Still - and this is the *key point* - I may suspect that selection (denoted here by the dummy z) into the labour market depends on education **and** motivation:

$$z_i = \left\{ \begin{array}{ll} 1 & \text{if } \gamma \cdot educ_i + (\theta_2 m_i + \eta_i) \geq 0 \\ 0 & \text{otherwise} \end{array} \right\},$$

  where θ_2 is a positive parameter and η_i is an error term independent of all factors except selection. Because m_i is unobserved it will go into the error term, which will consist of the two terms inside the parentheses (.).

- The big question now is whether the factors determining selection are correlated with the wage error term $\varepsilon_i = \theta_1 m_i + e_i$. There are only three terms determining selection. Two of these are η_i and $educ_i$, and they have been assumed uncorrelated with ε_i. But what about motivation, m_i? Abstracting from the uninteresting case where θ_1 and/or θ_2 are equal to zero, we see that i) motivation determines selection; and ii) motivation is correlated with the wage error term since $\varepsilon_i = \theta_1 m_i + e_i$. So clearly we have endogenous selection.

- Does this imply that education is correlated with ε_i **in the selected sample**? Yes it does. The intuition as to why this is so is straightforward. Think about the characteristics (education and motivation) of the people that are included in the sample.

  - Someone with a **low** level of education must have a **high** level of motivation, otherwise he or she is likely not to be included in the sample (recall: the selection model implies that individuals with **low** levels of education and **low** levels of motivation are those most unlikely to be included in the sample).

  - In contrast, someone with a **high** level of education is fairly likely to participate in the labour market even if he or she happens to have a relatively low level of motivation.

- The implication is that, **in the sample**, the average level of motivation among those with little education will be higher than the average level of motivation with those with a lot of education. In other words, education and motivation are negatively correlated **in the sample**, even though this is not the case in the population.

- And since motivation goes into the error term (since we have no data on motivation - it's unobserved), it follows that education is (negatively) correlated with the error term in the selected sample. And that's why we get selectivity bias.

- Illustration: Figure 2 in the appendix.

3.2. How correct for sample selection bias?

I will now discuss the two most common ways of correcting for sample selection bias.

3.2.1. Method 1: Inclusion of control variables

The first method by which we can correct for selection bias is simple: include in the regression observed variables that control for sample selection. In the wage example above , if we had data on motivation, we could augment the wage model with this variable:

$$\ln w_i = \beta_0 + \beta_1 educ_i + \theta_1 m_i + e_i.$$

More generally, recall that

$$E\left(y_1 | \boldsymbol{x}, v_2\right) = \boldsymbol{x}_1 \boldsymbol{\beta}_1 + \gamma_1 v_2,$$

and so if you have data on v_2, we could just use include this variable in the model as a control variable for selection and estimate the primary equation using OLS. Such a strategy would completely solve the sample selection problem.

Clearly this approach is only feasible if we have data on the relevant factors (e.g. motivation), which may not always be the case. The second way of correcting for selectivity bias is to use the famous **Heckit method**, developed by James Heckman in the 1970s.

3.2.2. Method 2: The Heckit method

We saw above that

$$E(y_1|\boldsymbol{x}, y_2 = 1) = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda(\boldsymbol{x}\boldsymbol{\delta}_2).$$

Using the same line of reasoning as for 'Method 1', it must be that if we had data on $\lambda(\boldsymbol{x}\boldsymbol{\delta}_2)$, we could simply add this variable to the model and estimate by OLS. Such an approach would be fine. Of course, in practice you would never have direct data on $\lambda(\boldsymbol{x}\boldsymbol{\delta}_2)$. However, the functional form $\lambda(\cdot)$ is known - or, rather, assumed (at least in most cases) - and $\boldsymbol{x}$ is observed. If so, the only missing ingredient is the parameter vector γ_1, which can be estimated by means of a probit model. The Heckit method thus consists of the following two steps:

1. Using **all** observations - those for which y_1 is observed (selected observations) and those for which it is not - and estimate a probit model where y_2 is the dependent variable and $\boldsymbol{x}$ are the explanatory variables. Based on the parameter estimates $\hat{\boldsymbol{\delta}}_2$ calculate the inverse Mills ratio for each observation:

$$\lambda\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right) = \frac{\phi\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right)}{\Phi\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right)}.$$

2. Using the selected sample, i.e. all observations for which y_1 is observed, run an OLS regression in which y_1 is the dependent variable and $\boldsymbol{x}_1$ **and** $\lambda\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right)$ are the explanatory variables:

$$y_1 = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right) + \varsigma_i.$$

This will give consistent estimates of the parameter vector $\boldsymbol{\beta}_1$. That is, by including the inverse Mills ratio as an additional explanatory variable, we have corrected for sample selection bias.

Important considerations

- The Heckit procedure gives you an estimate of the parameter γ_1, which measures the covariance between the two error terms u_1 and v_2. Under the null hypothesis that there is no selectivity bias, we have $\gamma_1 = 0$. Hence testing $H_0 : \gamma_1 = 0$ is of interest, and we can do this by means of a conventional t-test. If you cannot reject $H_0 : \gamma_1 = 0$ then this indicates that sample selection does not result in significant bias, and so using OLS on the selected sample without including the inverse Mills ratio is fine - all this, provided the model is correctly specified (i.e. all the underlying assumptions hold), of course.

- We assumed above that the vector $\boldsymbol{x}$ (the determinants of selection) contains all variables that go into the vector $\boldsymbol{x}_1$ (the explanatory variables in the primary equation), and possibly additional variables. In fact, it is highly desirable to specify the selection equation in such a way that there is *at least one* variable that determines selection, and which has no direct effect on y_1 conditional on $\lambda\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right)$. In other words, it is important to impose at least one *exclusion restriction.* The reason is that if $\boldsymbol{x}_1 = \boldsymbol{x}$, the second stage of Heckit is likely to suffer from a *collinearity problem*, with very imprecise estimates as a result. Recall the form of the regression you run in the second stage of Heckit:

$$y_1 = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right) + \varsigma_i.$$

  Clearly, if $\boldsymbol{x}_1 = \boldsymbol{x}$, then

$$y_1 = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda\left(\boldsymbol{x}_1\hat{\boldsymbol{\delta}}_2\right) + \varsigma_i.$$

  Remember that collinearity arises when one explanatory variable can be expressed as a **linear** function of one or several of the other explanatory variables in the model. In the above model $\boldsymbol{x}_1$ enters linearly (the first term) and **non**-linearly (through inverse Mills ratio), which seems to suggest that there will not be perfect collinearity. However, if you look at the graph of the inverse Mills ratio (see appendix for the lecture on corner response models) you see that it is **virtually linear over a wide range of values**. Clearly had it been exactly linear there would be no way of estimating

$$y_1 = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda\left(\boldsymbol{x}_1\hat{\boldsymbol{\delta}}_2\right) + \varsigma_i$$

because $\boldsymbol{x}_1$ would then be perfectly collinear with $\lambda\left(\boldsymbol{x}_1\hat{\boldsymbol{\delta}}_2\right)$. The fact that Mills ratio is virtually linear over a wide range of values means that you can run into problems posed by severe (albeit not complete) collinearity. This problem is solved (or at least mitigated) if $\boldsymbol{x}$ contains one or several variables that are not included in $\boldsymbol{x}_1$. Similar to identification with instrumental variables, the exclusion restriction has to be justified theoretically in order to be convincing. And that, alas, is not always straightforward.

- Finally, always remember that in order to use the Heckit approach, you must have data on the explanatory variables for both selected and non-selected observations. This may not always be the case.

Quantities of interest Now consider partial effects. Suppose we are interested in the effects of changing the variable x_k. It is useful to distinguish between two quantities of interest:

- The effect of a change on x_k on expected y_i in the population:

$$\frac{\partial E\left(y_1|\boldsymbol{x}_1\boldsymbol{\beta}_1\right)}{\partial x_k} = \beta_k$$

  For example, if x_k is education and y_1 is wage offer, then β_k measures the marginal effect of education on expected wage offer in the population.

- The effect of a change on x_k on expected y_i for individuals in the population for whom y_i is observed:

$$\frac{\partial E\left(y_1|\boldsymbol{x}_1\boldsymbol{\beta}_1, y_2 = 1\right)}{\partial x_k} = \beta_k + \gamma_1\frac{\partial\lambda\left(\boldsymbol{x}\boldsymbol{\delta}_2\right)}{\partial x_k}.$$

Recall that

$$\lambda'(c) = -\lambda(c)\left[c + \lambda(c)\right],$$

hence

$$\frac{\partial E\left(y_1|\boldsymbol{x}_1\boldsymbol{\beta}_1, y_2 = 1\right)}{\partial x_k} = \beta_k - \gamma_1\delta_k\lambda\left(\boldsymbol{x}\boldsymbol{\delta}_2\right)\left[\boldsymbol{x}\boldsymbol{\delta}_2 + \lambda\left(\boldsymbol{x}\boldsymbol{\delta}_2\right)\right].$$

It can be shown that $c + \lambda(c) > 0$, hence if γ_1 and δ_k have the **same sign**, this partial effect is lower than that on expected y_1 in the population. In the context of education and wage offers, what is the intuition of this result? [Hint: increase education and less able individuals will work.]

Estimation of Heckit in Stata In Stata we can use the command **heckman** to obtain Heckit estimates. The syntax has the following form

```
heckman y x1, select (z1 x1) twostep
```

where the variable y is **missing** whenever an observation is not included in the selected sample.

If you omit the twostep option you get results from a full information maximum likelihood (FIML) estimator. The assumptions underlying this estimator are stronger than those underpinning Heckman's two-stage estimator; specifically, the requires the error terms u_1 and v_2 are bivariate normal. Under bivariate normality, FIML is more efficient; asymptotically, the two methods (FIML and two-step) are equivalent, but in small samples the results can differ. Simulations have taught us that FIML can be sensitive to mis-specification due to, say, non-normal disturbance terms. In applied work it makes sense to consider both sets of results.

EXAMPLES: See Section 3 in appendix; replicates the results in example 19.6, Wooldridge (2010, pp. 807-8).

3.3. Extensions of the Heckit model

3.3.1. Non-continuous outcome variables

We have focused on the case where y_1, i.e. the outcome variable in the structural equation, is a continuous variable. However, sample selection models can be formulated for many different models - binary response models, censored models, duration models etc. The basic mechanism generating selection bias remains the same: correlation between the unobservables determining selection and the unobservables determining the outcome variable of interest.

Consider the following binary response model with sample selection:

$$\begin{aligned} y_1 &= 1\left[\boldsymbol{x}_1\boldsymbol{\beta}_1 + u_1 > 0\right] \\ y_2 &= 1\left[\boldsymbol{x}\boldsymbol{\delta}_2 + v_2 > 0\right], \end{aligned}$$

where y_1 is observed only if $y_2 = 1$, and $\boldsymbol{x}$ contains $\boldsymbol{x}_1$ and at least one more variable. In this case, probit estimation of $\boldsymbol{\beta}_1$ based on the selected sample will generally lead to inconsistent results, unless u_1 and v_2 are uncorrelated. Assuming that $\boldsymbol{x}$ is exogenous in the population, we can use a two-stage procedure very similar to that discussed above:

1. Obtain $\boldsymbol{\delta}_2$ by estimating the participation equation using a probit model. Construct $\lambda\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right)$.

2. Estimate the structural equation using probit, with $\lambda\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right)$ added to the set of regressors.

This is a good procedure for testing the null hypothesis that there is no selection bias (in which case $\lambda\left(\boldsymbol{x}\hat{\boldsymbol{\delta}}_2\right)$ is insignificant in the structural equation). If, based on this test we decide there is endogenous selection, we should estimate the two equations of the model simultaneously (in Stata: **heckprob**). See Wooldridge, Section 19.6.3 for more details.

Alternatively, it could be that the selection equation is not a binary response model. For example, Bourguignon, Fournier and Gurgand consider the case where selection is modelled by means of a

multinomial logit.[1]

3.3.2. Endogenous explanatory variables

The techniques discussed above can also be extended to allow for endogeneity in the explanatory variables. Wooldridge (2010; Section 19.6.2) focuses on the case where there's a single endogenous explanatory variable y_2; the model looks like this:

$$\begin{aligned} y_1 &= \boldsymbol{z}_1\boldsymbol{\delta}_1 + \alpha y_2 + u_1 \\ y_2 &= \boldsymbol{z}_2\boldsymbol{\delta}_2 + v_2 \\ y_3 &= 1\left[\boldsymbol{z}\boldsymbol{\delta}_3 + v_3 > 0\right], \end{aligned}$$

where the first equation is the structural equation of interest, the second equation is a reduced form equation for the potentially endogenous explanatory variable y_2,. and the third is the selection equation; (u_1, v_2, v_3) are freely correlated.

Exercise: How would you estimate this model? Be specific about what you assume regarding observability of the variables and the exclusion restrictions. Once you have outlined an answer, compare it to Wooldridge's discussion in Section 19.6.2.

3.3.3. Heckit with panel data (optional)

Model:

$$y_{it1} = \boldsymbol{x}_{it1}\boldsymbol{\beta}_1 + c_{i1} + u_{it1}, \qquad \text{(Primary equation)}$$

where selection is determined by the equation

$$s_{it2} = \left\{ \begin{array}{ll} 1 & \text{if } \boldsymbol{x}_i\boldsymbol{\gamma}_{t2} + v_{it2} \geq 0 \\ 0 & \text{otherwise} \end{array} \right\}. \qquad \text{(Selection equation)}$$

[1] François Bourguignon, Martin Fournier, Marc Gurgand "Selection Bias Corrections Based on the Multinomial Logit Model: Monte-Carlo Comparisons" DELTA working paper 2004-20, downloadable at http://www.delta.ens.fr/abstracts/wp200420.pdf

- If selection bias arises because c_{i1} is correlated with v_{it2}, then estimating the main equation using a fixed effects or first differenced approach on the selected sample will produce consistent estimates of $\boldsymbol{\beta}_1$.

- However, if $corr\left(v_{it2}, u_{it1}\right) \neq 0$, we can address the sample selection problem using a panel Heckit approach, where we begin by estimating T different selection probits (i.e. do not use xtprobit here, use pooled probit), and compute T inverse Mills ratios, denoted $\hat{\lambda}_{it2}$.

- Then make a Chamberlain-type assumption:

$$E\left(c_{i1}|\boldsymbol{x}_i, v_{it2}\right) = \boldsymbol{x}_i\boldsymbol{\pi}_1 + \phi_{t1}v_{it2}$$

  and regress y_{it} on $\boldsymbol{x}_{it}$, $\boldsymbol{x}_i, \hat{\lambda}_{it2}, d2_t\hat{\lambda}_{it2}, ..., dT_t\hat{\lambda}_{it2}$.

- This procedure is a consistent estimator of $\boldsymbol{\beta}_1$.

PhD Programme: Econometrics III
Appendix

1. Illustration: The truncated regression model

Consider a simple simulation, obtained by the following Stata code:

```
clear
set seed 2355
set obs 500

ge u=invnorm(uniform())

ge x=2*uniform()

/* true population model: y = -1 + 1*x + u /

ge y=-1+x+u

/* no truncation */
reg y x
predict yh_ols_nt

/* truncation of y at 0.8*/
reg y x if y<.8
predict yh_ols_t

/* truncated regression corrects for the truncation. ul(.) indicates
the upper limit */

truncreg y x, ul(0.8)
```

Consider three different regressions based on these artificial data:

i) OLS using the full sample of 500 observations (i.e. no truncation)

```
. reg y x

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  156.47
       Model |  139.883218     1  139.883218           Prob > F      =  0.0000
    Residual |  445.219899   498  .894015862           R-squared     =  0.2391
-------------+------------------------------           Adj R-squared =  0.2375
       Total |  585.103118   499  1.17255134           Root MSE      =  .94552

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .8940591   .0714753    12.51   0.000     .7536288    1.034489
       _cons |  -.9019037   .0834538   -10.81   0.000    -1.065869   -.7379389
------------------------------------------------------------------------------
```

ii) OLS using the truncated sample of 380 observations

```
. reg y x if y<.8

      Source |       SS       df       MS              Number of obs =     380
-------------+------------------------------           F(  1,   378) =   47.00
       Model |   28.616886     1   28.616886           Prob > F      =  0.0000
    Residual |  230.164146   378  .608899857           R-squared     =  0.1106
-------------+------------------------------           Adj R-squared =  0.1082
       Total |  258.781032   379  .682799556           Root MSE      =  .78032

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .4811388    .070183     6.86   0.000     .3431407    .6191369
       _cons |  -.8577185   .0732374   -11.71   0.000    -1.001722   -.7137147
------------------------------------------------------------------------------
```

Notice coefficient on x is much lower than the true value of one. It is clearly significantly different from one, indicating significant bias.

Figure 3 illustrates the problem of truncation.

iii) Truncated regression which corrects for the truncation

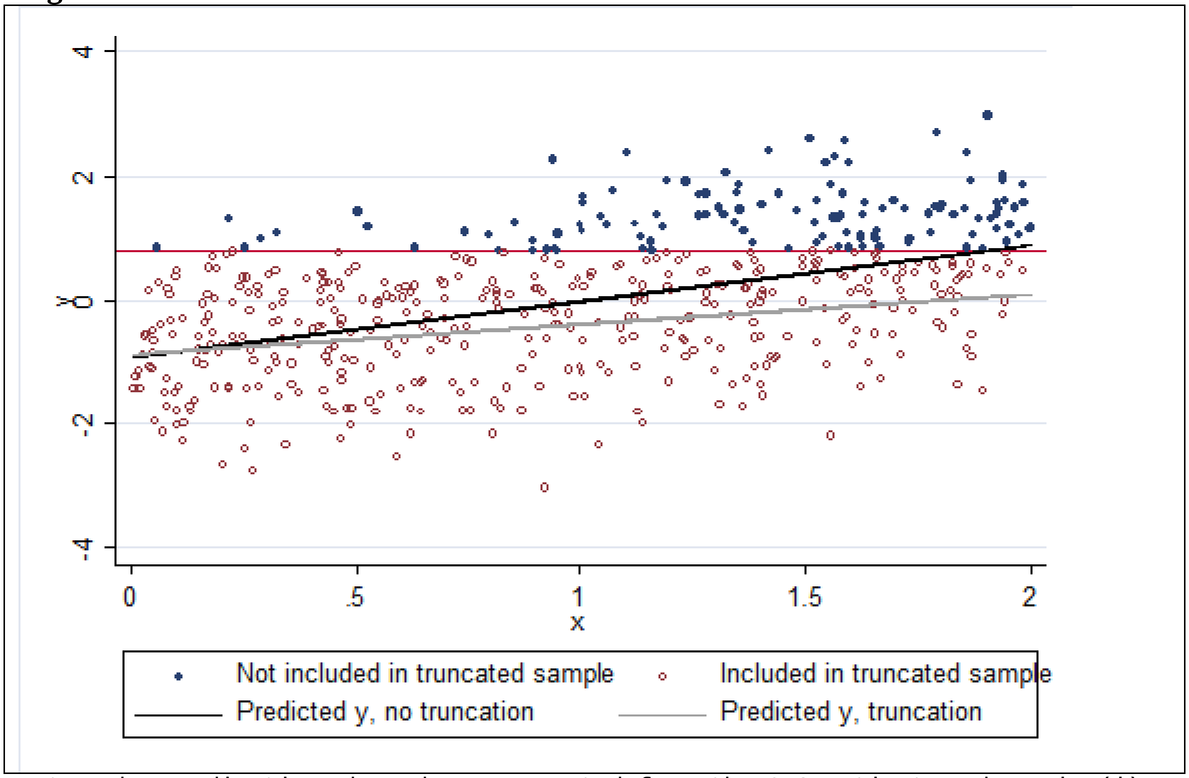
```
. truncreg y x, ul(0.8)
(note: 120 obs. truncated)

Truncated regression
Limit:   lower =       -inf                          Number of obs =    380
         upper =         .8                          Wald chi2(1)  =  37.41
Log likelihood = -398.51329                          Prob > chi2   = 0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
           x |   .8506762   .1390748     6.12   0.000     .5780947    1.123258
       _cons |  -.7836381   .1214471    -6.45   0.000     -1.02167   -.5456061
-------------+----------------------------------------------------------------
sigma        |
       _cons |   1.019341    .067624    15.07   0.000     .8868003    1.151882
------------------------------------------------------------------------------
```

Coefficient increases as a result and is similar to the OLS estimate in (i) and not significantly different from the true value of 1.

Figure 2. The effect of truncation on the OLS estimator



Note: The predications have been generated from the OLS estimates shown in (i) and (ii) above.

2. Derivation of the Inverse Mills Ratio (IMR)

To show $$E(z \mid z > c) = \frac{\phi(c)}{1-\Phi(c)} = \frac{\phi(-c)}{\Phi(-c)}$$

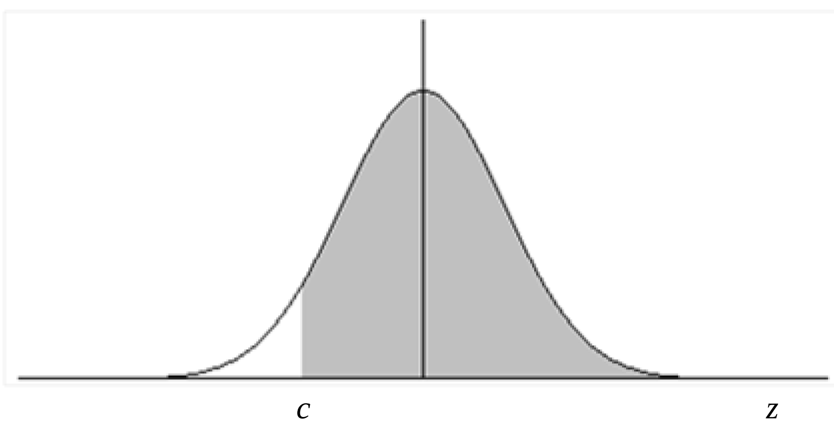
Assume that z is normally distributed:

$$G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \phi(z)dz$$

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})$$

$G(z)$ is the normal cumulative density function (CDF), $\phi(z)$ is the standard normal density function.

We now wish to know the $E(z \mid z > c)$. It is the shaded area in the graph below.



By the characteristics of the normal curve is equal to $[1-\Phi(c)]$. So the density of z is given by

$$\frac{\phi(z)}{[1-\Phi(c)]}, \quad z > c$$

so

$$E(z \mid z > c) = \int_{c}^{\infty} \frac{z\phi(z)}{[1-\Phi(c)]}dz$$

which can be written using the definitions above as:

$$E(z \mid z > c) = \frac{1}{(1-\Phi(c))}\int_c^{\infty} \frac{z}{\sqrt{2\pi}}.\exp(\frac{-z^2}{2})dz$$

This expression can be written as:

$$E(z \mid z > c) = \frac{1}{(1-\Phi(c))}\int_c^{\infty} -(\frac{d\phi(z)}{dz})dz$$

How do we know that:

$$\frac{d\phi(z)}{dz} = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})\cdot -z$$

$$\int_c^{\infty} -(\frac{d\phi(z)}{dz})dz = \int_c^{\infty} -\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2}) = 0 + \frac{1}{\sqrt{2\pi}}\exp(-\frac{c^2}{2}) = \phi(c)$$

So:

Lets evaluate $\int_c^{\infty} \frac{z}{\sqrt{2\pi}}.\exp(\frac{-z^2}{2})dz =$

This can be written as

$$-\frac{1}{[1-\Phi(c)]}\int_c^{\infty} d\Phi(z) = \frac{\phi(c)}{[1-\Phi(c)]}$$

Recall that for the normal distribution $\phi(c) = \phi(-c)$ and $1-\Phi(c) = \Phi(-c)$

From which it follows that

$$E(z \mid z > c) = \int_c^{\infty} \frac{z\phi(z)}{[1-\Phi(c)]}dz = \frac{\phi(-c)}{\Phi(-c)}$$

It is this last expression which is the inverse Mills ratio.

Figure 1: The Inverse Mills Ratio

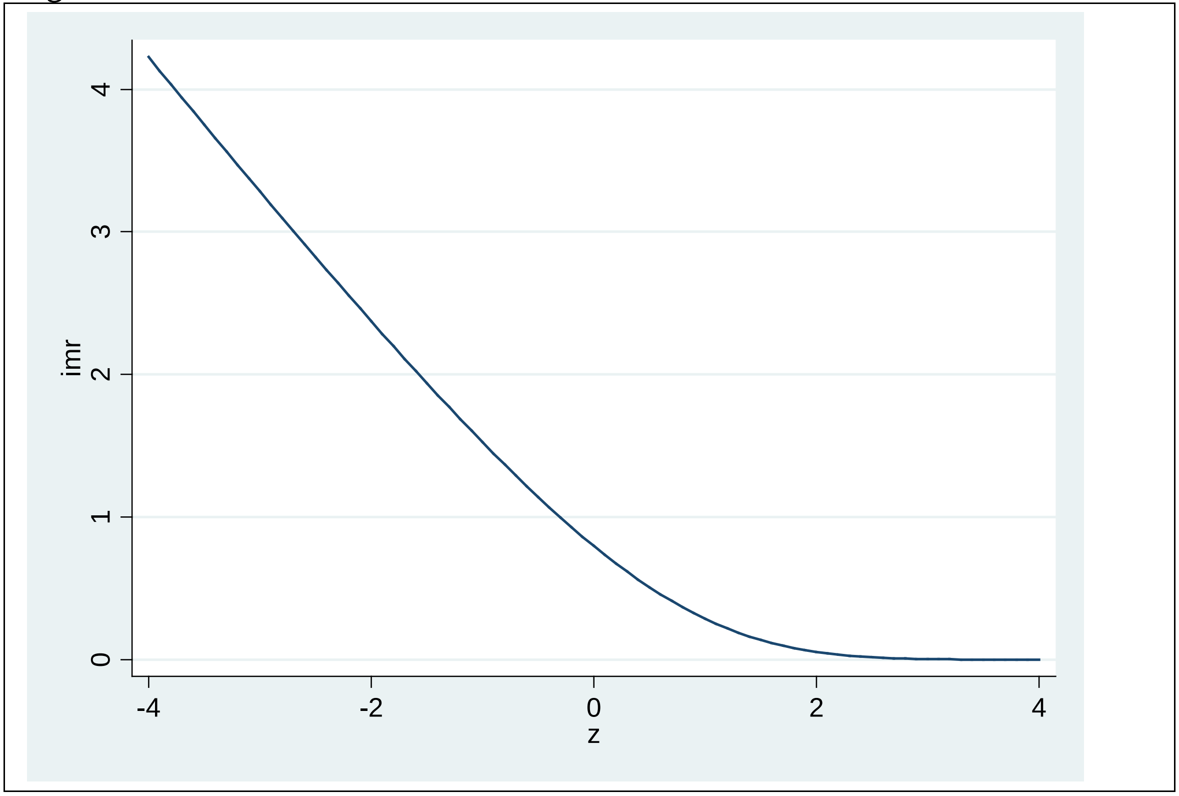
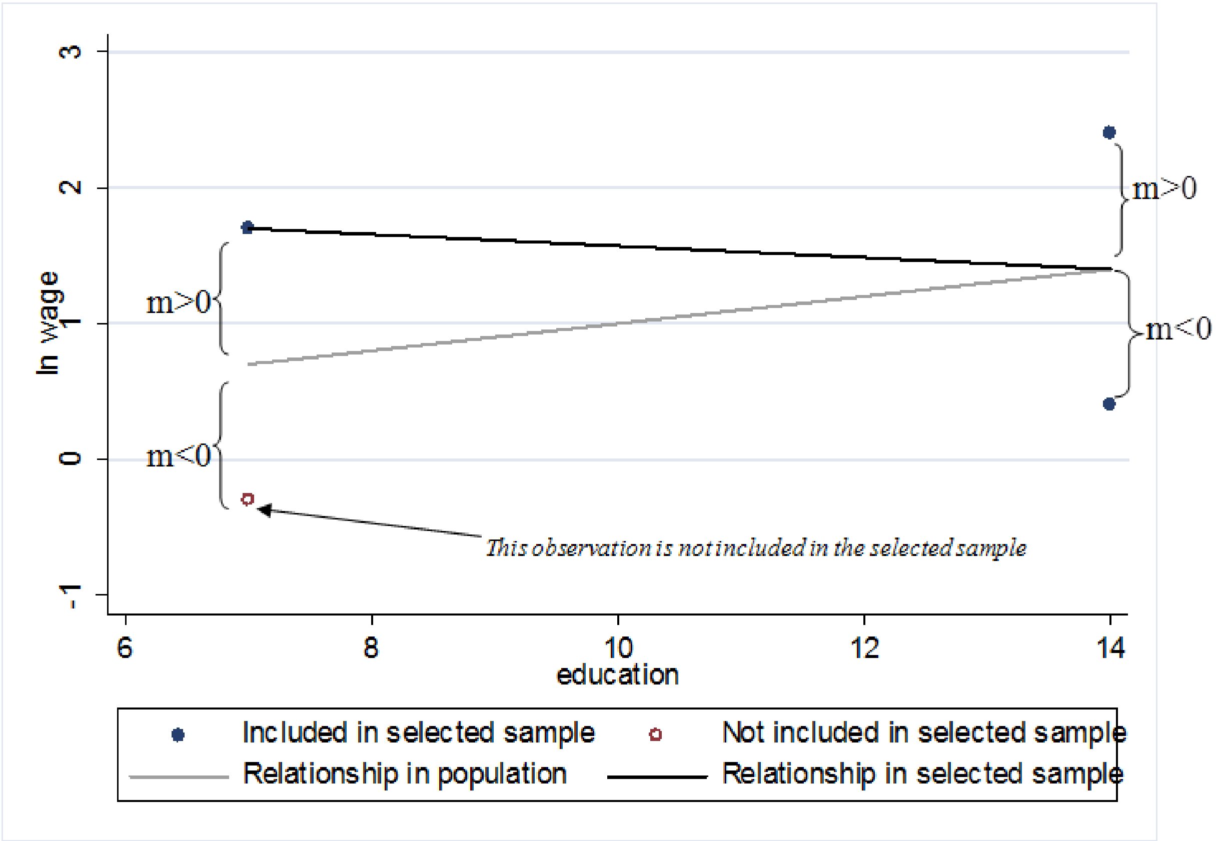


Figure 2: Illustration of Sample Selection Bias



The economic model underlying the graph is

$$\ln w = cons + 0.1educ + m,$$

where w is wage, educ is education and m is unobserved motivation. Selection into the sample is a positive function of educ and m.

3. Empirical illustration of the Heckit model

Earnings regressions for females in the US

This section uses the MROZ dataset.[1] This dataset contains information on 753 women. We observe the wage offer for only 428 women, hence the sample is truncated.

```
use C:\teaching_gbg07\applied_econ07\MROZ.dta
```

1. OLS on selected sample

```
reg  lwage educ exper expersq

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  3,   424) =   26.29
       Model |  35.0223023     3  11.6741008           Prob > F      =  0.0000
    Residual |  188.305149   424  .444115917           R-squared     =  0.1568
-------------+------------------------------           Adj R-squared =  0.1509
       Total |  223.327451   427  .523015108           Root MSE      =  .66642

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1074896   .0141465     7.60   0.000     .0796837    .1352956
       exper |   .0415665   .0131752     3.15   0.002     .0156697    .0674633
     expersq |  -.0008112   .0003932    -2.06   0.040    -.0015841   -.0000382
       _cons |  -.5220407   .1986321    -2.63   0.009    -.9124668   -.1316145
------------------------------------------------------------------------------
```

[1] Source: Mroz, T.A. (1987) "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," Econometrica 55, 765-799.

2. Two-step Heckit

```
. heckman lwage educ exper expersq, select(nwifeinc educ exper expersq age
kidslt6 kidsge6) twostep

Heckman selection model -- two-step estimates   Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(6)       =    180.10
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1090655    .015523     7.03   0.000     .0786411      .13949
       exper |   .0438873   .0162611     2.70   0.007     .0120163    .0757584
     expersq |  -.0008591   .0004389    -1.96   0.050    -.0017194    1.15e-06
       _cons |  -.5781033   .3050062    -1.90   0.058    -1.175904    .0196979
-------------+----------------------------------------------------------------
select       |
    nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
        educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
       exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
     expersq |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
         age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
     kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
     kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
       _cons |   .2700768    .508593     0.53   0.595    -.7267472    1.266901
-------------+----------------------------------------------------------------
mills        |
      lambda |   .0322619   .1336246     0.24   0.809    -.2296376    .2941613
-------------+----------------------------------------------------------------
         rho |    0.04861
       sigma |  .66362876
      lambda |  .03226186   .1336246
------------------------------------------------------------------------------
```

3. Simultaneous estimation of selection model

```
. heckman lwage educ exper expersq, select(nwifeinc educ exper expersq age
kidslt6 kidsge6)

Iteration 0:   log likelihood = -832.89777
Iteration 1:   log likelihood =  -832.8851
Iteration 2:   log likelihood = -832.88509

Heckman selection model                         Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(3)       =     59.67
Log likelihood = -832.8851                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1083502   .0148607     7.29   0.000     .0792238    .1374767
       exper |   .0428369   .0148785     2.88   0.004     .0136755    .0719983
     expersq |  -.0008374   .0004175    -2.01   0.045    -.0016556   -.0000192
       _cons |  -.5526974   .2603784    -2.12   0.034     -1.06303   -.0423652
-------------+----------------------------------------------------------------
select       |
    nwifeinc |  -.0121321   .0048767    -2.49   0.013    -.0216903    -.002574
        educ |   .1313415   .0253823     5.17   0.000     .0815931    .1810899
       exper |   .1232818   .0187242     6.58   0.000     .0865831    .1599806
     expersq |  -.0018863   .0006004    -3.14   0.002     -.003063   -.0007095
         age |  -.0528287   .0084792    -6.23   0.000    -.0694476   -.0362098
     kidslt6 |  -.8673988   .1186509    -7.31   0.000     -1.09995   -.6348472
     kidsge6 |   .0358723   .0434753     0.83   0.409    -.0493377    .1210824
       _cons |   .2664491   .5089578     0.52   0.601    -.7310898    1.263988
-------------+----------------------------------------------------------------
     /athrho |    .026614    .147182     0.18   0.857    -.2618573    .3150854
    /lnsigma |  -.4103809   .0342291   -11.99   0.000    -.4774687   -.3432931
-------------+----------------------------------------------------------------
         rho |   .0266078   .1470778                     -.2560319    .3050564
       sigma |   .6633975   .0227075                      .6203517    .7094303
      lambda |   .0176515   .0976057                     -.1736521    .2089552
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =     0.03   Prob > chi2 = 0.8577
------------------------------------------------------------------------------
```

CHAPTER 3:

COUNT RESPONSES

1. Introduction

A count variable is a variable that takes on nonnegative integer values - e.g. number of times someone is arrested in a year, number of children ever born to a woman, number of visits to the doctor in a year, etc. A common feature of count variables is that there is a lower bound at zero.

If y is a count variable and $\boldsymbol{x}$ is a vector of explanatory variables, we are often interested in the population regression $E(y|\boldsymbol{x})$.

Using OLS for estimating $E(y|\boldsymbol{x})$ is certainly an option, however linear models have shortcomings similar to those for binary responses or corner responses (negative predictions not ruled out etc.). Using a log transformation solves these problems, however this approach is not very useful if y is often equal to zero. With count data, it is better to model $E(y|\boldsymbol{x})$ directly and to choose functional forms that ensure positivity for any values of $\boldsymbol{x}$ and any parameter values.

In this lecture I provide an introduction to the econometrics of count data models. I draw on Chapter 18.1-18.3 in Wooldridge (2010) and, to a lesser extent, on selected parts in Chapter 20, Cameron and Trivedi (2005), *Microeconometrics*.

2. Poisson Regression

Recap: The Poisson Distribution

- The Binomial distribution: If p is the probability of a success and there are n independent trials of some experiment, then the probability of observing z successes is equal to

$$f(z) = B(z; n, p) = \frac{n!}{z!(n-z)!} p^z (1-p)^{n-z},$$

  where the coefficient $\frac{n!}{z!(n-z)!}$, known as the binomial coefficient, captures the fact that there are $\frac{n!}{z!(n-z)!}$ different ways of distributing z successes across n independent trials (recall $x!$ is x factorial (i.e. 4!=4*3*2*1 etc.).

- The Poisson distribution: Let $n \to \infty$ and $p \to 0$ but in such a way that $np = \mu > 0$ for all n and p. We get

$$\lim_{n\to\infty} B(z;\mu) = \frac{\exp(-\mu)\,\mu^z}{z!}.$$

One implications is that $Var(z) = \mu$, i.e. the variance of z is equal to its mean.

Poisson regression Poisson regression involves specifying μ as a function of $\boldsymbol{x}$: $\mu = \mu(\boldsymbol{x}) \equiv E(y|\boldsymbol{x})$. This implies that y given $\boldsymbol{x}$ has a Poisson distribution:

$$f(y|\boldsymbol{x}) = \exp[-\mu(\boldsymbol{x})]\,[\mu(\boldsymbol{x})]^y\,/y!$$

Hence the density of y given $\boldsymbol{x}$ is completely determined by the conditional mean $\mu(\boldsymbol{x}) \equiv E(y|\boldsymbol{x})$. Moreover,

$$Var(y|\boldsymbol{x}) = E(y|\boldsymbol{x}), \tag{2.1}$$

which is a testable (and often rejected) restriction in empirical work. We refer to (2.1) as the **Poisson-variance assumption**. We return to this assumption later.

Example in appendix: The Poisson distribution for $\mu = 2$ and $\mu = 7$.

Given a parametric model $m(\boldsymbol{x};\beta)$ for $\mu(\boldsymbol{x})$, the log likelihood for observation i is

$$\begin{aligned} l_i(\boldsymbol{\beta}) &= \log f(y|\boldsymbol{x}) \\ l_i(\boldsymbol{\beta}) &= -m(\boldsymbol{x};\boldsymbol{\beta}) + y_i \log(m(\boldsymbol{x};\boldsymbol{\beta})) - \log(y!)\,. \end{aligned}$$

Fortunately, we can drop the computationally awkward term $\log(y!)$ because it does not depend on the parameters β, and write the log likelihood simply as

$$l_i(\boldsymbol{\beta}) = -m(\boldsymbol{x};\boldsymbol{\beta}) + y_i \log(m(\boldsymbol{x};\boldsymbol{\beta}))\,.$$

A popular choice for $m(.)$ is the exponential,

$$m(\boldsymbol{x};\boldsymbol{\beta}) = \exp(\boldsymbol{x}\boldsymbol{\beta}),$$

where $x_1 = 1$, yielding

$$l_i(\boldsymbol{\beta}) = -\exp(\boldsymbol{x}_i\boldsymbol{\beta}) + y_i \cdot \boldsymbol{x}_i\boldsymbol{\beta}.$$

Other functional forms than the exponential can be used - see Wooldridge (2010), p. 727 for a brief discussion (punchline: "exponential regression with flexible functions of the explanatory variables is often adequate").

Interpretation of the parameters Interpretation of the parameters is straightforward. Keeping in mind that we have now specified

$$E(y|\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta}),$$

it follows that

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} = \exp(\boldsymbol{x}\boldsymbol{\beta})\beta_j$$

for continuous x_j. Hence the partial effect on $E(y|\boldsymbol{x})$ depends on $\boldsymbol{x}\boldsymbol{\beta}$ and the sign of the effect is determined by the sign of β_j. It also follows that

$$\begin{aligned}
\beta_j &= \frac{\partial E(y|\boldsymbol{x})}{\partial x_j}\frac{1}{\exp(\boldsymbol{x}\boldsymbol{\beta})} \\
\beta_j &= \frac{\partial E(y|\boldsymbol{x})}{\partial x_j}\frac{1}{E(y|\boldsymbol{x})} \\
\beta_j &= \frac{\partial \log E(y|\boldsymbol{x})}{\partial x_j},
\end{aligned}$$

hence $100 \times \beta_j$ is the semielasticity of $E(y|\boldsymbol{x})$ with respect to x_j; if we replace x_j by $\log x_j$, β_j is the elasticity of $E(y|\boldsymbol{x})$ with respect to x_j. Effects of dummy variables or variables that enter $\boldsymbol{x}$ in a nonlinear fashion are easy to write down also (please do).

Computing average partial effects (APEs) of an explanatory variable on the mean is straightforward.

The sample log likelihood looks like this:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(-\exp(\boldsymbol{x}_i \boldsymbol{\beta}) + y_i \cdot \boldsymbol{x}_i \boldsymbol{\beta} \right). \tag{2.2}$$

We maximize the sample log likelihood with respect to the parameters $\boldsymbol{\beta}$, hence the first-order conditions can be written

$$\sum_{i=1}^{N} \boldsymbol{x}_i' \left(-\exp(\boldsymbol{x}_i \boldsymbol{\beta}) + y_i \right) = \mathbf{0},$$

which shows that the residuals $y_i - \exp(\boldsymbol{x}_i \boldsymbol{\beta})$ always sum to zero; hence $\bar{y} = \overline{\hat{y}}$, where $\hat{y}_i = \exp\left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}}\right)$ are the fitted values. The APE referring to x_j is thus

$$\frac{1}{N} \sum_{i=1}^{N} \exp\left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}}\right) \hat{\beta}_j = \bar{y} \hat{\beta}_j,$$

i.e. simply a product of two scalars. Thus, as a rough comparison with linear model estimates, the Poisson coefficients can be multiplied by the average outcome $\bar{y}$.

Simple measures of the goodness of fit are the pseudo R-squared (reported by Stata - how is it defined?) or the squared correlation coefficient between the dependent variable and the predictions $\exp\left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}}\right)$ (proposed by Wooldridge; can also be obtained of course by regressing y_i on a constant and $\exp\left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}}\right)$).

Example in appendix: Determinants of Fertility (replication of Example 18.1 in Wooldridge, 2010)

The Poisson-variance assumption Recall the Poisson-variance assumption:

$$Var(y|\boldsymbol{x}) = E(y|\boldsymbol{x})$$

This is clearly restrictive - and testable. A weaker assumption allows the variance-mean ratio to be a positive constant, σ^2:

$$\begin{aligned}\frac{Var\left(y|\boldsymbol{x}\right)}{E\left(y|\boldsymbol{x}\right)} &= \sigma^2 \\ Var\left(y|\boldsymbol{x}\right) &= \sigma^2 E\left(y|\boldsymbol{x}\right).\end{aligned}$$

If $\sigma^2 > 1$ we have **overdispersion** (relative to the Poisson case), while if $\sigma^2 < 1$ there is **underdispersion**. Clearly, if $\sigma^2 \neq 1$, the Poisson distribution is mis-specified. Does it matter if we have this type of mis-specification problem or not? Key results as follows:

- If $\sigma^2 \neq 1$, the correct log likelihood contribution is **not**

  $$l_i\left(\boldsymbol{\beta}\right) = -\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) + y_i \cdot \boldsymbol{x}_i\boldsymbol{\beta}.$$

  Question: What happens if we nevertheless estimate $\boldsymbol{\beta}$ based on a sample log likelihood function made up of such individual contributions? Because of the mis-specification, we refer to an estimator based on (2.2) as the **Poisson quasi-maximum likelihood estimator (QMLE).** The good news is that, despite the fact that the likelihood is incorrectly specified, QMLE a consistent estimator of the parameters of interest. In particular, if we assume that for some value $\boldsymbol{\beta}_o$,

  $$E\left(y|\boldsymbol{x}\right) = m\left(\boldsymbol{x}, \boldsymbol{\beta}o\right),$$

  it can be shown that $\boldsymbol{\beta}_o$ is the unique solution to $\max_{\boldsymbol{\beta}} E\left[l_i\left(\boldsymbol{\beta}\right)\right]$.

- However, the conventional formula for computing standard errors is based on the assumption that $\sigma^2 = 1$, and will generally be incorrect if $\sigma^2 \neq 1$. Wooldridge (2010) shown in Section 18.2.3 that the variance of $\boldsymbol{\beta}$ with σ^2 *unrestricted* can be estimated as

  $$\widehat{Avar\left(\hat{\boldsymbol{\beta}}\right)} = \hat{\sigma}^2 \left(\sum_{i=1}^{N} \nabla_\beta \hat{m}_i' \nabla_\beta \hat{m}_i / \hat{m}_i\right)^{-1}, \tag{2.3}$$

where the j:th element of the $K \times 1$ vector $\nabla_\beta \hat{m}_i$ is equal to

$$\frac{\partial \hat{m}_i}{\partial \beta_j} = x_j \exp\left(\boldsymbol{x}_i \boldsymbol{\beta}\right),$$

(continuing to assume $m(.)$ is exponential). Wooldridge calls the standard errors obtained from (2.3) **GLM (generalized linear model) standard errors**. The key point to note here is that $\hat{\sigma}^2$ features in (2.3). But the conventional Poisson regression standard errors assume the variance-mean ratio is equal to 1 (i.e. $\hat{\sigma}^2 = 1$will be *imposed*). Hence, if there is overdispersion ($\sigma^2 > 1$), these standard errors will be overestimated, while if there's underdispersion they will be underestimated.

- Fortunately, the solution is obvious: once you have obtained the Poisson standard errors, simply scale them by the square root of $\hat{\sigma}^2$. Alternatively, we may opt for a fully robust asymptotic variance matrix estimator, which does not require $\sigma^2 = 1$. This has the familiar White-type 'sandwich' form - see eq. (18.14) in Wooldridge (2010) for details.

- If σ^2 is far from one, predicted conditional probabilities and sampling distributions based on the Poisson distribution with $\sigma^2 = 1$ imposed can be very misleading.

- It's now clear that σ^2 is a parameter of some interest: it tells us whether the Poisson distribution is correct, and if it isn't, we can use an estimate of σ^2 in order to correct the standard errors for the bias caused by the mis-specification. How, then, do we obtain $\hat{\sigma}^2$? First, note that (unlike OLS) $\hat{\sigma}^2$ is **not** exactly an estimate of the variance of the difference between y_i and $\exp\left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}}\right)$. Remember the definition:

$$Var\left(y|\boldsymbol{x}\right) = \sigma^2 E\left(y|\boldsymbol{x}\right).$$

Appealing to the sample analogy principle using

$$\sigma^2 = \frac{Var\left(y|\boldsymbol{x}\right)}{E\left(y|\boldsymbol{x}\right)}$$

as the starting point, we write

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 / \hat{m}_i,$$

where $\hat{u}_i = y_i - \hat{m}_i$ and $\hat{m}_i = \exp\left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}}\right)$.

Fertility example continued: Having estimated the Poisson regression, I obtain $\hat{\sigma}^2$ as follows:

```
predict xb, xb

generate yhat=exp(xb)

generate sig2hat=((children-yhat)^2)/yhat

summarize sig2hat
```

The sample average of `sig2hat` is equal to 0.749, and the square root of that (which is my estimate of σ, i.e. $\hat{\sigma}$) is 0.866 (this estimate is also reported in Table 18.1 in Wooldridge, 2010). Hence, we have $\hat{\sigma} < 1$ implying underdispersion in the data; and the standard errors shown in Table 2.2 should all be multiplied by 0.866, yielding the GLM standard errors (and obviously higher z statistics). I can obtain the corrected standard errors using the Stata `glm` command with scale(x2) added as an option:

```
glm children educ age agesq evermarr urban electric tv, family(poisson) scale(x2)
```

Results are shown in Table 2.3.

Once we have obtained reliable estimates of the standard errors, hypothesis testing is straightforward. Testing a hypothesis regarding an individual parameter is based on the reported z-values; multiple hypotheses are probably best tested using Wald tests. A (quasi) log likelihood ratio test, defined as follows

$$QLR \equiv 2\left[\log L\left(\hat{\boldsymbol{\beta}}_{\text{Unrestricted}}\right) - \log L\left(\hat{\boldsymbol{\beta}}_{\text{Restricted}}\right)\right] / \hat{\sigma},$$

may alternatively be used, but then we really do need to assume $Var\,(y|\boldsymbol{x}) = \sigma^2 E\,(y|\boldsymbol{x})$ which, as we have seen, is not necessary for consistency of the QMLE.

You are encouraged to use the FERTIL2 data to verifying the following statements:

- Fully robust standard errors for the Poisson QMLE are similar to the GLM standard errors

- If you multiply children by some constant and re-run the Poisson regression, the reported standard errors will change but the parameter estimates will not.

3. Negative Binomial Regression Models

A popular alternative to Poisson QMLE is full maximum likelihood estimation of the "NegBin I" model proposed by Cameron and Trivedi (1986). This model is parameterized through the slope parameters $\boldsymbol{\beta}$ and an additional parameter to be estimated $\eta^2 > 0$, where $\sigma^2 = 1 + \eta^2$.

At first glance, this may look like an improvement over the Poisson regression in the sense that the Poisson mean-variance assumption $Var(y|\boldsymbol{x}) = E(y|\boldsymbol{x})$ is relaxed. But remember that we don't really need this assumption - the Poisson QMLE will be consistent anyway. Moreover, Wooldridge asserts (I admit to not having studied the proof) that a joint estimator of $\boldsymbol{\beta}$ and η^2 generally is *inconsistent* if $\sigma^2 = 1 + \eta^2$ is an incorrect assumption. Therefore, Poisson QMLE is more robust than NegBin I if the goal is to estimate the parameters $\boldsymbol{\beta}$.

What if conditional probabilities need to be estimated? Then we really do need to estimate parameters summarizing the extent of dispersion in the data, and base our predictions on probabilities more general than the Poisson function $f(y|\boldsymbol{x}) = \exp[-\mu(\boldsymbol{x})][\mu(\boldsymbol{x})]^y / y!$ NegBin I would be a step in the right direction, but may not be fully satisfactory either.

A more interesting extension of the Poisson regression is the "NegBin II" model (Cameron and Trivedi, 1986). This model can be derived from a Poisson model with unobserved heterogeneity. Key assumptions:

- Conditional on the vector of observables $\boldsymbol{x}_i$, and an unobserved heterogeneity term $c_i > 0$, y_i follows a Poisson distribution with mean $c_i \exp(\boldsymbol{x}_i\boldsymbol{\beta})$.

- c_i is independent of $\boldsymbol{x}_i$, and has a **gamma distribution** with mean equal to 1 and $Var(c_i) = \eta^2$.

It can then be shown that

$$\begin{aligned} E(y_i|\boldsymbol{x}_i) &= \exp(\boldsymbol{x}_i\boldsymbol{\beta}) \\ Var(y_i|\boldsymbol{x}_i) &= \exp(\boldsymbol{x}_i\boldsymbol{\beta}) + \eta^2 [\exp(\boldsymbol{x}_i\boldsymbol{\beta})]^2, \end{aligned}$$

in other words the variance is a **quadratic** in the conditional mean. Since $\eta^2 > 0$ (follows from $Var(c_i) = \eta^2$ and $Var(c_i) > 0$), NegBin II implies overdispersion, increasing with $E(y_i|\boldsymbol{x}_i)$.

Example in appendix: Contacts with medical doctor (Cameron & Trivedi, *Microeconometrics*, 2005, section 20.3)

4. A Two-Part Model for Count Responses

Reference: Cameron & Trivedi (2005), Chapter 20.4.5.

For both applications discussed above we observed **excess zeros**, i.e. the presence of more zeros in the data than predicted by Poisson or NegBin II. Recall the two-part generalization of the tobit type I model discussed previously in this course, in which the zero vs. non-zero outcome is modeled separately from amount differences amongst the positives. This approach can be used for count data too:

- We model the zeros by the density $f_1(.)$, so that $\Pr(y = 0) = f_1(0)$, where f_1 can be probit, logit or some other density suitable for modeling binary outcomes.
- We model the positive counts using a truncated density

$$f_2(y|y > 0) = \frac{f_2(y)}{1 - f_2(0)},$$

which is multiplied by $\Pr(y > 0) = 1 - f_1(0)$ to ensure the probabilities sum to one.

Thus:

$$g(y) = \left\{ \begin{array}{ll} f_1(0) & \text{if } y = 0 \\ (1 - f_1(0)) \frac{f_2(y)}{1 - f_2(0)} & \text{if } y > 0 \end{array} \right\}.$$

The probability of observing a zero given the Poisson distribution is easily obtained:

$$f_2(y|\boldsymbol{x}) = \exp[-\mu(\boldsymbol{x})][\mu(\boldsymbol{x})]^y / y!$$

implies

$$f_2(0|\boldsymbol{x}) = \exp[-\mu(\boldsymbol{x})],$$

hence

$$1 - f_2(0|\boldsymbol{x}) = 1 - \exp[-\mu(\boldsymbol{x})],$$

hence the truncated density for the Poisson model is given by

$$\frac{f_2(y)}{1 - f_2(0)} = \frac{\exp[-\mu(\boldsymbol{x})][\mu(\boldsymbol{x})]^y}{y!(1 - \exp(-\mu(\boldsymbol{x})))}. \tag{4.1}$$

The likelihood function can now be constructed combining terms like (4.1) with probit or logit probability expressions. Estimation of the two models is done separately. A central object of interest is the partial effect of x_j on $E(y_i|\boldsymbol{x}_i)$. We have

$$E(y_i|\boldsymbol{x}_i) = \Pr(y_i = 0|\boldsymbol{x}_i) \cdot 0 + \Pr(y_i > 0|\boldsymbol{x}_i) E(y_i|\boldsymbol{x}_i, y_i > 0),$$

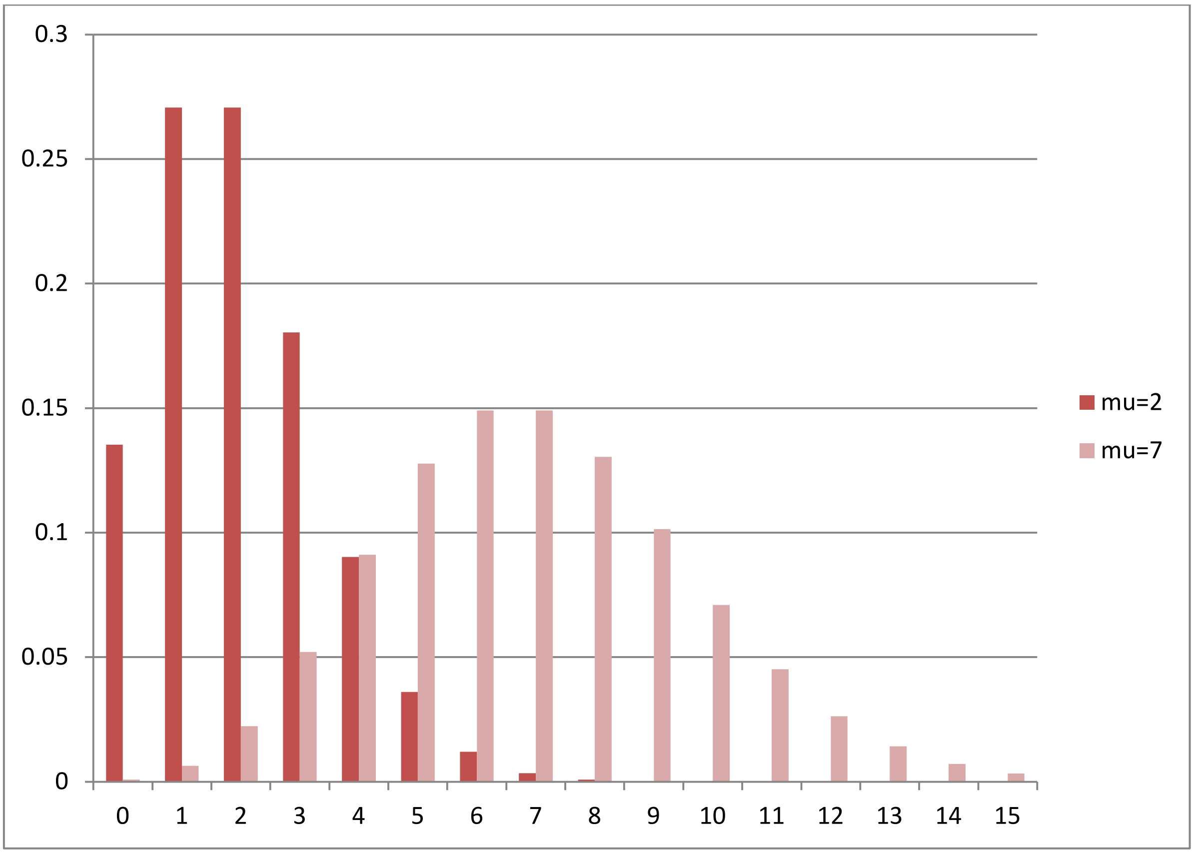
thus

$$\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_j} = \Pr(y_i > 0|\boldsymbol{x}_i) \frac{\partial E(y_i|\boldsymbol{x}_i, y_i > 0)}{\partial x_j} + \frac{\partial \Pr(y_i > 0|\boldsymbol{x}_i)}{\partial x_j} E(y_i|\boldsymbol{x}_i, y_i > 0).$$

Partial effects such as this one are straightforward to compute. Obtaining standard errors appears to be more awkward, however. See Section 4 in the appendix for some results.

PhD Programme: Econometrics II
Department of Economics, University of Gothenburg
Appendix
Måns Söderbom

1. The Poisson Distribution: Illustrations



Matlab code:

```
z=0:1:15; mu=7; p=exp(-mu)*mu.^z./factorial(z);
[z' p']
```

Some early insights:

- **If your data contain a lot of zeros, then the sample mean should be close to zero, or the distribution cannot be Poisson.**
- **The distribution is more asymmetric if mu is low**

2. Determinants of Fertility

The results in this section replicate and extend the results in example 18.1 in Wooldridge (2010), pp. 730-32). The dataset is called FERTIL2; it contains information on women in Botswana; summary statistics and variable labels as follows:

Variable	Obs	Mean	Std. Dev.	Min	Max
children	4361	2.267828	2.222032	0	13
educ	4361	5.855996	3.927075	0	20
age	4361	27.40518	8.685233	15	49
agesq	4361	826.46	526.9232	225	2401
evermarr	4361	.4767255	.4995153	0	1
urban	4361	.5166246	.4997808	0	1
electric	4358	.1402019	.3472363	0	1
tv	4359	.0929112	.2903413	0	1

variable name	storage type	display format	value label	variable label
children	byte	%8.0g		number of living children
educ	byte	%8.0g		years of education
age	byte	%8.0g		age in years
agesq	int	%8.0g		age^2
evermarr	byte	%9.0g		=1 if ever married
urban	byte	%8.0g		=1 if live in urban area
electric	byte	%8.0g		=1 if has electricity
tv	byte	%8.0g		=1 if has tv

Observation:

The variance of children is more than twice as high as the mean. This suggests children cannot follow an unconditional Poisson distribution. For the Poisson regression model, we assume that children, given the explanatory variables, has a Poisson distribution (which of course is not the same as assuming children is unconditionally Poisson).

For a regression model with no explanatory variables, there would be _____-dispersion in the children data (fill in the blank).

Figure 2: Sample distribution of children. Implied distribution if Poisson and mu=2.27

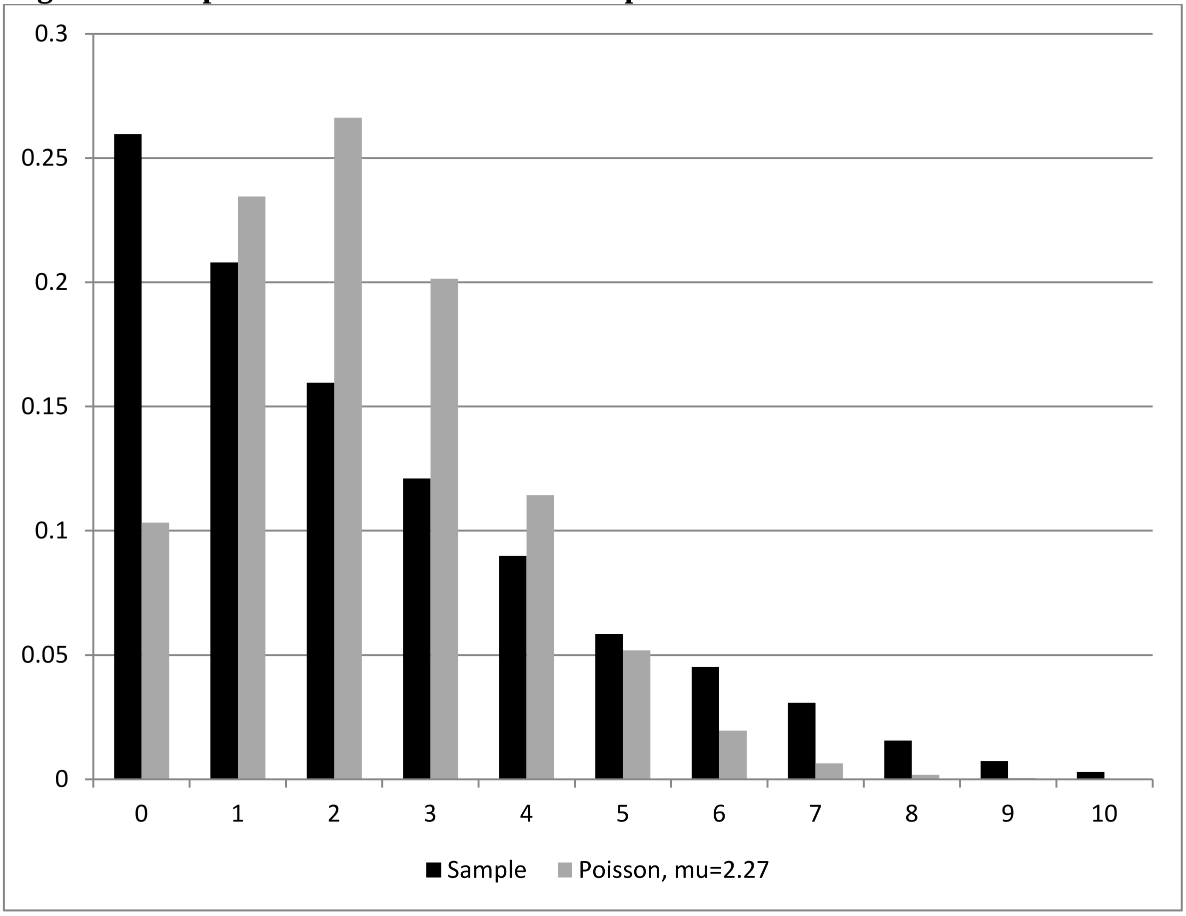
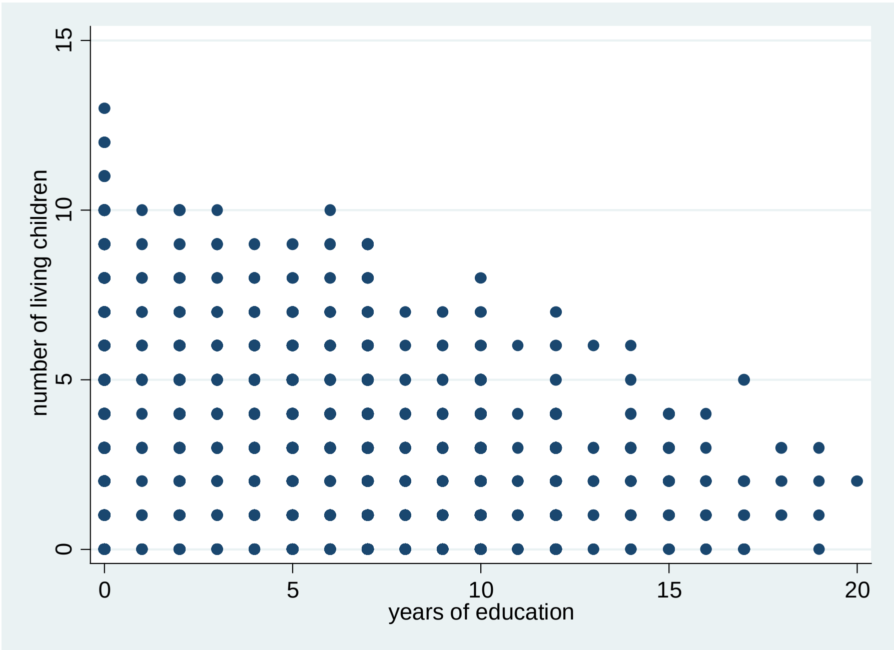


Figure 3: Fertility and Education - Heteroskedasticity



2.1 OLS results

```
. regress children educ age agesq evermarr urban electric tv

      Source |       SS       df       MS              Number of obs =    4358
-------------+------------------------------           F(  7,  4350) =  893.91
       Model |  12688.9349     7  1812.70499           Prob > F      =  0.0000
    Residual |  8821.09719  4350  2.02783843           R-squared     =  0.5899
-------------+------------------------------           Adj R-squared =  0.5892
       Total |  21510.0321  4357  4.93689055           Root MSE      =   1.424

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0644086   .0063199   -10.19   0.000    -.0767987   -.0520184
         age |   .2724736    .017019    16.01   0.000     .2391077    .3058395
       agesq |  -.0019067    .000274    -6.96   0.000    -.0024438   -.0013696
    evermarr |   .6822725    .052167    13.08   0.000     .5799986    .7845463
       urban |  -.2278933   .0458653    -4.97   0.000    -.3178126    -.137974
    electric |  -.2617394   .0758688    -3.45   0.001     -.410481   -.1129979
          tv |  -.2499509   .0901474    -2.77   0.006    -.4266858   -.0732161
       _cons |   -3.39384   .2445496   -13.88   0.000    -3.873281   -2.914398
------------------------------------------------------------------------------

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of children

         chi2(1)      =  1873.85
         Prob > chi2  =   0.0000

. regress children educ age agesq evermarr urban electric tv, robust

Linear regression                                      Number of obs =    4358
                                                       F(  7,  4350) =  885.79
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5899
                                                       Root MSE      =   1.424

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0644086   .0063525   -10.14   0.000    -.0768628   -.0519544
         age |   .2724736   .0198484    13.73   0.000     .2335606    .3113866
       agesq |  -.0019067   .0003555    -5.36   0.000    -.0026036   -.0012098
    evermarr |   .6822725   .0526617    12.96   0.000     .5790287    .7855162
       urban |  -.2278933   .0447829    -5.09   0.000    -.3156907   -.1400959
    electric |  -.2617394   .0729908    -3.59   0.000    -.4048385   -.1186404
          tv |  -.2499509   .0821469    -3.04   0.002    -.4110007   -.0889012
       _cons |   -3.39384   .2515591   -13.49   0.000    -3.887024   -2.900656
------------------------------------------------------------------------------
```

2.2 Results from Poisson regression

```
. poisson children educ age agesq evermarr urban electric tv

Poisson regression                                Number of obs   =       4358
                                                  LR chi2(7)      =    6167.34
                                                  Prob > chi2     =     0.0000
Log likelihood = -6497.0599                       Pseudo R2       =     0.3219

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0216645   .0029131    -7.44   0.000     -.027374   -.0159549
         age |   .3373308   .0099365    33.95   0.000     .3178556     .356806
       agesq |  -.0041158   .0001453   -28.33   0.000    -.0044006   -.0038311
    evermarr |    .314751   .0244473    12.87   0.000     .2668352    .3626668
       urban |  -.0860549   .0216487    -3.98   0.000    -.1284855   -.0436243
    electric |  -.1205347    .038839    -3.10   0.002    -.1966578   -.0444116
          tv |  -.1447046   .0473875    -3.05   0.002    -.2375824   -.0518268
       _cons |  -5.374829   .1628673   -33.00   0.000    -5.694043   -5.055615
------------------------------------------------------------------------------

Computing Wooldridge's R-squared:

. predict xb, xb
(3 missing values generated)

. ge yhat=exp(xb)
(3 missing values generated)

. reg children yhat

      Source |       SS       df       MS              Number of obs =    4358
-------------+------------------------------           F(  1,  4356) = 6468.24
       Model |  12853.7443     1  12853.7443           Prob > F      =  0.0000
    Residual |   8656.2878  4356  1.98721024           R-squared     =  0.5976
-------------+------------------------------           Adj R-squared =  0.5975
       Total |  21510.0321  4357  4.93689055           Root MSE      =  1.4097

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yhat |   .9675659   .0120306    80.43   0.000     .9439798     .991152
       _cons |   .0735461   .0346438     2.12   0.034     .0056267    .1414656
------------------------------------------------------------------------------
```

Average partial effects of education:

i) By hand:

```
predict xb, xb

ge dEy_deduc=_b[educ]*exp(xb)

sum dEy_deduc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   dEy_deduc |      4358   -.0491254    .0384582   -.138027  -.0036272
```

[NOTE: Instead of exp(xb), we could just use the sample mean of children]

```
ii) Using margins:

margins, dydx (*)

Average marginal effects                          Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : educ age agesq evermarr urban electric tv

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0491254   .0066241    -7.42   0.000    -.0621083   -.0361424
         age |   .7649159   .0238093    32.13   0.000     .7182506    .8115811
       agesq |  -.0093329   .0003425   -27.25   0.000    -.0100042   -.0086615
    evermarr |    .713715   .0558986    12.77   0.000     .6041558    .8232741
       urban |  -.1951341   .0491288    -3.97   0.000    -.2914247   -.0988436
    electric |   -.273319   .0881125    -3.10   0.002    -.4460163   -.1006217
          tv |  -.3281255   .1075044    -3.05   0.002    -.5388303   -.1174206
------------------------------------------------------------------------------
```

```
Warning: The reported standard errors are NOT correct, since there is
underdispersion in the data.
```

2.3 Results from Poisson regression using `glm`

```
. glm children educ age agesq evermarr urban electric tv, family(poisson) scale(x2)

Iteration 0:   log likelihood = -6614.4491
Iteration 1:   log likelihood = -6497.4043
Iteration 2:   log likelihood = -6497.0599
Iteration 3:   log likelihood = -6497.0599

Generalized linear models                          No. of obs      =      4358
Optimization     : ML                              Residual df     =      4350
                                                   Scale parameter =         1
Deviance         =   3908.76293                    (1/df) Deviance =  .8985662
Pearson          =  3265.867362                    (1/df) Pearson  =  .7507741

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  2.985342
Log likelihood   = -6497.059873                    BIC             = -32543.23

------------------------------------------------------------------------------
             |                 OIM
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0216645   .0025241    -8.58   0.000    -.0266117   -.0167173
         age |   .3373308   .0086097    39.18   0.000     .3204561    .3542055
       agesq |  -.0041158   .0001259   -32.70   0.000    -.0043625   -.0038691
    evermarr |    .314751   .0211829    14.86   0.000     .2732333    .3562688
       urban |  -.0860549    .018758    -4.59   0.000    -.1228198     -.04929
    electric |  -.1205347   .0336529    -3.58   0.000    -.1864933   -.0545762
          tv |  -.1447046     .04106    -3.52   0.000    -.2251807   -.0642285
       _cons |  -5.374829     .14112   -38.09   0.000    -5.651419   -5.098239
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion.)
```

Note: These standard errors are corrected for underdispersion. `margins` will now give me the right standard errors for the APEs (see next page).

```
. margins, dydx (*)

Average marginal effects                          Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted mean children, predict()
dy/dx w.r.t. : educ age agesq evermarr urban electric tv

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0491254   .0057396    -8.56   0.000    -.0603747    -.037876
         age |   .7649159   .0206301    37.08   0.000     .7244817      .80535
       agesq |  -.0093329   .0002968   -31.44   0.000    -.0099146   -.0087511
    evermarr |    .713715   .0484345    14.74   0.000      .618785    .8086449
       urban |  -.1951341   .0425687    -4.58   0.000    -.2785673    -.111701
    electric |   -.273319    .076347    -3.58   0.000    -.4229564   -.1236815
          tv |  -.3281255   .0931496    -3.52   0.000    -.5106954   -.1455556
------------------------------------------------------------------------------
```

3. Contacts with Medical Doctor

This example is taken from Cameron & Trivedi, Microeconometrics, 2005, section 20.3. The dataset is called **randdata.dta** and can be obtained from:

http://cameron.econ.ucdavis.edu/mmabook/mmadata.html

The main objective of the original research based on these data was to assess how the use of health services is affected by types of randomly assigned health insurance (Deb and Trivedi, 2002). The data file consists of utilization, expenditures, demographic characteristics, health status, and insurance status variables. Variable labels and summary statistics as follows:

```
              storage  display     value
variable name   type   format      label      variable label
---------------------------------------------------------------------------
mdvis           float  %9.0g                  number face-to-fact md visits
logc            float  %9.0g                  log(coinsurance+1)
idp             float  %9.0g                  individual deductible plan
lpi             float  %9.0g                  log participation incentive
fmde            float  %9.0g                  function of mdeoff
physlm          float  %9.0g                  physical limitations --
baselin
disea           float  %9.0g                  count of chronic diseases --
ba
hlthg           float  %9.0g                  good health
hlthf           float  %9.0g                  fair health
hlthp           float  %9.0g                  poor health
linc            float  %9.0g
lfam            float  %9.0g                  log of family size
xage            float  %9.0g                  age that year
female          float  %9.0g                  female
child           float  %9.0g                  child
femchild        float  %9.0g
black           float  %9.0g                  black
educdec         float  %9.0g                  education of decision maker
```

```
. summarize mdvis logc idp lpi fmde physlm disea hlthg hlthf hlthp linc
lfam xage female child femchild black educdec, sep(0)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       mdvis |     20186    2.860696    4.504765          0         77
        logc |     20186    2.383588    2.041713          0   4.564348
         idp |     20186    .2599822    .4386354          0          1
         lpi |     20186    4.708827    2.697293          0   7.163699
        fmde |     20186    4.030322    3.471234          0   8.294049
      physlm |     20186    .1235247    .3220437          0          1
       disea |     20186     11.2445    6.741647          0       58.6
       hlthg |     20186    .3620826    .4806144          0          1
       hlthf |     20186    .0772813    .2670439          0          1
       hlthp |     20186    .0149609    .1213992          0          1
        linc |     20186    8.708167     1.22841          0   10.28324
        lfam |     20186    1.248404    .5390681          0   2.639057
        xage |     20186    25.71844    16.76759          0   64.27515
      female |     20186    .5169424    .4997252          0          1
       child |     20186    .4014168    .4901972          0          1
    femchild |     20186    .1937481    .3952436          0          1
       black |     20186    .1815343    .3827365          0          1
     educdec |     20186    11.96681    2.806255          0         25
```

3.1 Results from Poisson regression

```
. poisson mdvis logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage
female child femchild black educdec


Poisson regression                                Number of obs   =      20186
                                                  LR chi2(17)     =   13106.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -60087.622                       Pseudo R2       =     0.0983

------------------------------------------------------------------------------
       mdvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logc |  -.0427332   .0060785    -7.03   0.000    -.0546469   -.0308195
         idp |  -.1613169   .0116218   -13.88   0.000    -.1840952   -.1385385
         lpi |   .0128511   .0018362     7.00   0.000     .0092523    .0164499
        fmde |   -.020613   .0035521    -5.80   0.000     -.027575   -.0136511
      physlm |   .2684048   .0123624    21.71   0.000     .2441749    .2926347
       disea |    .023183   .0006081    38.12   0.000     .0219912    .0243749
       hlthg |   .0394004   .0095884     4.11   0.000     .0206074    .0581934
       hlthf |   .2531119    .016212    15.61   0.000     .2213369    .2848869
       hlthp |   .5216034   .0272382    19.15   0.000     .4682176    .5749892
        linc |   .0834099   .0051656    16.15   0.000     .0732854    .0935343
        lfam |  -.1296626   .0089603   -14.47   0.000    -.1472245   -.1121008
        xage |   .0023756   .0004311     5.51   0.000     .0015306    .0032206
      female |   .3487667   .0113504    30.73   0.000     .3265203     .371013
       child |   .3361904   .0178194    18.87   0.000     .3012649    .3711158
    femchild |  -.3625218   .0179396   -20.21   0.000    -.3976827   -.3273608
       black |  -.6800518   .0155484   -43.74   0.000    -.7105262   -.6495775
     educdec |   .0176149   .0016387    10.75   0.000     .0144031    .0208268
       _cons |  -.1898766   .0491731    -3.86   0.000    -.2862541    -.093499
------------------------------------------------------------------------------
```

3.2 Results from Poisson regression with glm

```
. glm mdvis logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage female
child femchild black educdec, family(poisson) scale(x2)

Generalized linear models                          No. of obs      =     20186
Optimization     : ML                              Residual df     =     20168
                                                   Scale parameter =         1
Deviance         =  79279.53229                    (1/df) Deviance =  3.930957
Pearson          =  119800.1596                    (1/df) Pearson  =  5.940111

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  5.955179
Log likelihood   = -60087.62207                    BIC             = -120640.7

------------------------------------------------------------------------------
             |                 OIM
       mdvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logc |  -.0427332   .0148148    -2.88   0.004    -.0717698   -.0136967
         idp |  -.1613169   .0283251    -5.70   0.000     -.216833   -.1058007
         lpi |   .0128511   .0044752     2.87   0.004     .0040799    .0216223
        fmde |   -.020613   .0086572    -2.38   0.017    -.0375809   -.0036451
      physlm |   .2684048   .0301301     8.91   0.000     .2093508    .3274588
       disea |    .023183   .0014821    15.64   0.000     .0202783    .0260878
       hlthg |   .0394004   .0233693     1.69   0.092    -.0064025    .0852033
       hlthf |   .2531119   .0395125     6.41   0.000     .1756688    .3305551
       hlthp |   .5216034   .0663858     7.86   0.000     .3914896    .6517172
        linc |   .0834099   .0125898     6.63   0.000     .0587343    .1080854
        lfam |  -.1296626   .0218384    -5.94   0.000     -.172465   -.0868603
        xage |   .0023756   .0010508     2.26   0.024     .0003161    .0044351
      female |   .3487667   .0276635    12.61   0.000     .2945471    .4029862
       child |   .3361904   .0434301     7.74   0.000     .2510688    .4213119
    femchild |  -.3625218    .043723    -8.29   0.000    -.4482172   -.2768263
       black |  -.6800518   .0378952   -17.95   0.000     -.754325   -.6057787
     educdec |   .0176149    .003994     4.41   0.000     .0097869    .0254429
       _cons |  -.1898766   .1198465    -1.58   0.113    -.4247713    .0450182
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion.)
```

Estimate of sigma2hat. Much higher than 1.

Note: **Considerable overdispersion. Adjusting the standard errors makes a big difference.**

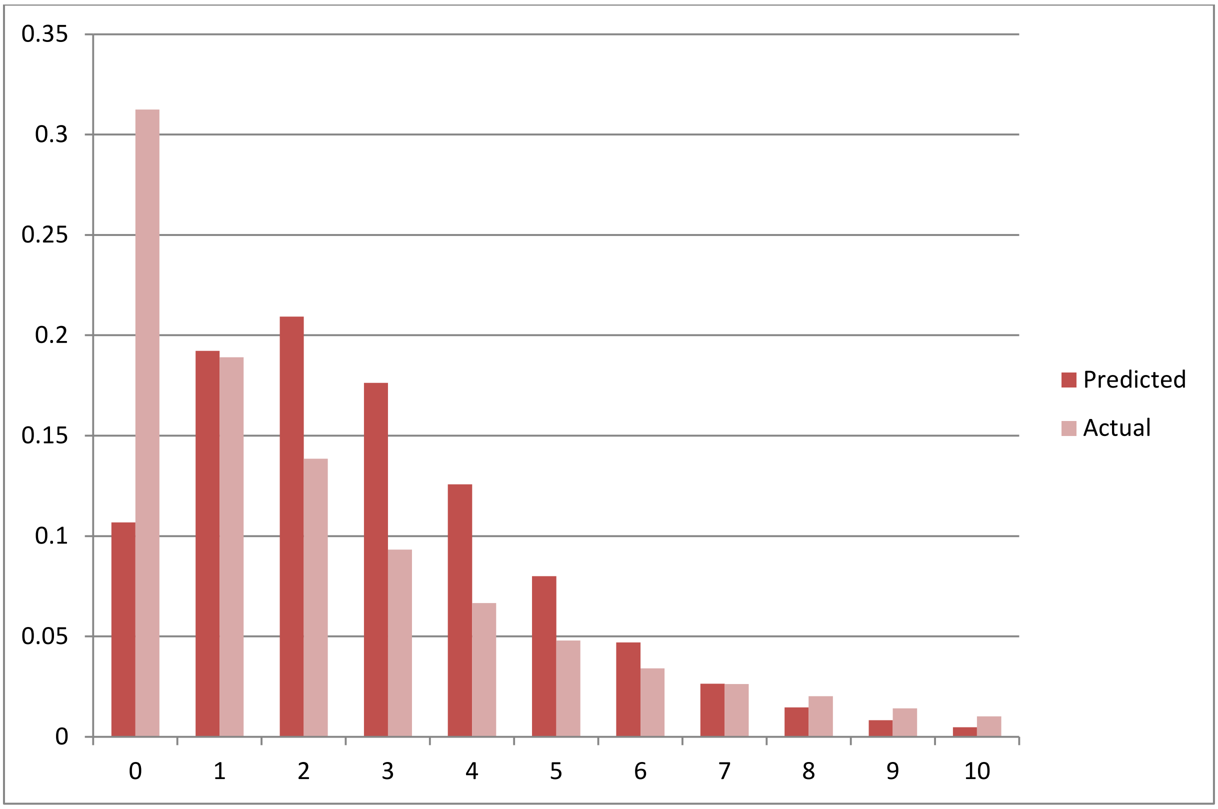
```
margins, dydx (*)

Average marginal effects                          Number of obs   =      20186
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage
female
               child femchild black educdec

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logc |  -.1453519   .0371484    -3.91   0.000    -.2181615   -.0725423
         idp |   -.425325   .0734597    -5.79   0.000    -.5693033   -.2813467
         lpi |   .0456313   .0117184     3.89   0.000     .0226636     .068599
        fmde |  -.0614799   .0216688    -2.84   0.005      -.10395   -.0190098
      physlm |   .7929485   .0861892     9.20   0.000     .6240208    .9618762
       disea |   .0747362   .0044514    16.79   0.000     .0660116    .0834609
       hlthg |   .0188377   .0582752     0.32   0.747    -.0953796     .133055
       hlthf |   .6825604   .1082641     6.30   0.000     .4703668    .8947541
       hlthp |   1.226593   .2145904     5.72   0.000     .8060039    1.647183
        linc |   .2435473   .0249019     9.78   0.000     .1947405    .2923541
        lfam |  -.3535108   .0558729    -6.33   0.000    -.4630197   -.2440019
        xage |   .0074759   .0027207     2.75   0.006     .0021435    .0128084
      female |   1.058397   .0709652    14.91   0.000     .9193082    1.197487
       child |   .8818771   .1117681     7.89   0.000     .6628157    1.100938
    femchild |  -1.082205   .1082329   -10.00   0.000    -1.294338   -.8700727
       black |  -2.047233    .084124   -24.34   0.000    -2.212113   -1.882353
     educdec |   .0468504   .0100629     4.66   0.000     .0271275    .0665733
------------------------------------------------------------------------------
```

Figure 3.1: Contacts with Medical Doctor: Observed and Fitted Frequencies based on Poisson Regression



Observations and insights:

- **The Poisson regression seriously underpredicts the proportion of zero visits and overestimates the proportion of positive number of visits up to seven.**

- **If our goal is to characterize the <u>distribution</u> of visits to the doctor, the Poisson regression is a bad approach. But if we are primarily interested in β, it may not be so bad.**

3.2 Results from NegBin II regression

```
. nbreg mdvis logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage
female child femchild black educdec


Negative binomial regression                      Number of obs   =      20186
                                                  LR chi2(17)     =    2828.01
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -42777.611                       Pseudo R2       =     0.0320

------------------------------------------------------------------------------
       mdvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logc |  -.0504405   .0128694    -3.92   0.000    -.0756641   -.0252169
         idp |  -.1475976   .0254099    -5.81   0.000    -.1974001   -.0977951
         lpi |   .0158351   .0040586     3.90   0.000     .0078805    .0237898
        fmde |   -.021335   .0075119    -2.84   0.005     -.036058   -.0066119
      physlm |   .2751715   .0295572     9.31   0.000     .2172404    .3331026
       disea |   .0259352   .0014827    17.49   0.000     .0230292    .0288412
       hlthg |   .0065371   .0202235     0.32   0.747    -.0331002    .0461744
       hlthf |   .2368643   .0374086     6.33   0.000     .1635448    .3101837
       hlthp |   .4256563   .0741812     5.74   0.000     .2802638    .5710488
        linc |   .0845165   .0085659     9.87   0.000     .0677277    .1013053
        lfam |  -.1226764    .019308    -6.35   0.000    -.1605195   -.0848333
        xage |   .0025943   .0009433     2.75   0.006     .0007455    .0044432
      female |   .3672884    .024005    15.30   0.000     .3202395    .4143373
       child |   .3060317   .0385618     7.94   0.000      .230452    .3816115
    femchild |  -.3755503   .0371392   -10.11   0.000    -.4483418   -.3027587
       black |  -.7104372   .0274929   -25.84   0.000    -.7643223   -.6565521
     educdec |   .0162582   .0034846     4.67   0.000     .0094285    .0230879
       _cons |  -.2069298   .0899431    -2.30   0.021    -.3832151   -.0306445
-------------+----------------------------------------------------------------
    /lnalpha |   .1674206   .0147901                      .1384326    .1964087
-------------+----------------------------------------------------------------
       alpha |   1.182251   .0174856                      1.148472    1.217024
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 3.5e+04 Prob>=chibar2 = 0.000
```

Observations:

- **The estimate of η^2 is reported as alpha: the results thus imply Var(y|x) = exp(xb)+1.18*(exp(xb)^2).**
- **Note that the estimated coefficients are quite similar to what was obtained from the simpler Poisson regression.**
- **The fit of the data is much better, however.**

3.3 Results from Zero Inflated NegBin II regression

```
zinb mdvis logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage female
child femchild black educdec, inf(logc idp lpi fmde physlm disea hlthg hlthf hlthp
linc lfam xage female child femchild black educdec)

Zero-inflated negative binomial regression        Number of obs   =      20186
                                                  Nonzero obs     =      13878
                                                  Zero obs        =       6308

Inflation model = logit                           LR chi2(17)     =    1505.67
Log likelihood  = -42493.84                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
       mdvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mdvis        |
        logc |  -.0086059   .0139112    -0.62   0.536    -.0358714    .0186595
         idp |  -.1686335   .0267574    -6.30   0.000    -.2210771   -.1161899
         lpi |   .0077349   .0042245     1.83   0.067     -.000545    .0160148
        fmde |  -.0292258   .0078858    -3.71   0.000    -.0446817   -.0137699
      physlm |   .2742958   .0296884     9.24   0.000     .2161077    .3324839
       disea |   .0232821   .0015173    15.34   0.000     .0203082    .0262559
       hlthg |   .0117103   .0209513     0.56   0.576    -.0293534    .0527741
       hlthf |   .2254197   .0393802     5.72   0.000      .148236    .3026035
       hlthp |    .404586   .0746336     5.42   0.000     .2583068    .5508651
        linc |   .0691123   .0096208     7.18   0.000      .050256    .0879687
        lfam |  -.1162018   .0205384    -5.66   0.000    -.1564564   -.0759473
        xage |   .0025331   .0009808     2.58   0.010     .0006107    .0044554
      female |   .2851505   .0262158    10.88   0.000     .2337686    .3365325
       child |   .2980832   .0422492     7.06   0.000     .2152763    .3808901
    femchild |  -.2692997   .0396782    -6.79   0.000    -.3470676   -.1915319
       black |  -.3630949   .0348303   -10.42   0.000     -.431361   -.2948288
     educdec |   .0112041   .0036065     3.11   0.002     .0041354    .0182727
       _cons |    .054865   .0992224     0.55   0.580    -.1396074    .2493374
-------------+----------------------------------------------------------------
inflate      |
        logc |    .576912   .0744402     7.75   0.000     .4310119    .7228121
         idp |  -.4390761   .1707392    -2.57   0.010    -.7737188   -.1044333
         lpi |  -.1550697   .0270002    -5.74   0.000     -.207989   -.1021504
        fmde |  -.0362254   .0410157    -0.88   0.377    -.1166146    .0441638
      physlm |  -.1717491   .2139769    -0.80   0.422    -.5911361    .2476379
       disea |  -.0660192   .0160202    -4.12   0.000    -.0974183   -.0346201
       hlthg |  -.1785634   .1473313    -1.21   0.226    -.4673275    .1102007
       hlthf |   .0690902   .2084244     0.33   0.740    -.3394141    .4775946
       hlthp |  -.3201465   .4916775    -0.65   0.515    -1.283817    .6435236
        linc |  -.0412511   .0324192    -1.27   0.203    -.1047915    .0222893
        lfam |   .0454647   .1317032     0.35   0.730    -.2126688    .3035983
        xage |  -.0076207   .0077923    -0.98   0.328    -.0228933     .007652
      female |  -1.598575   .2253219    -7.09   0.000    -2.040198   -1.156952
       child |  -.4843173   .3051426    -1.59   0.112    -1.082386    .1137511
    femchild |   1.970006   .2701394     7.29   0.000     1.440542    2.499469
       black |   2.666002   .1846913    14.43   0.000     2.304013     3.02799
     educdec |  -.0926334   .0224535    -4.13   0.000    -.1366415   -.0486252
       _cons |  -1.324506   .5758135    -2.30   0.021     -2.45308   -.1959322
-------------+----------------------------------------------------------------
    /lnalpha |    .024474   .0190476     1.28   0.199    -.0128587    .0618067
-------------+----------------------------------------------------------------
       alpha |   1.024776   .0195196                      .9872236    1.063757
------------------------------------------------------------------------------
```

```
. margins, dydx (*)

Average marginal effects                          Number of obs   =      20186
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage
female
               child femchild black educdec

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logc |  -.0950083    .037225    -2.55   0.011    -.1679679   -.0220486
         idp |  -.4289201   .0727991    -5.89   0.000    -.5716037   -.2862366
         lpi |   .0410499   .0115475     3.55   0.000     .0184172    .0636827
        fmde |  -.0792003   .0217921    -3.63   0.000    -.1219121   -.0364885
      physlm |   .8057607   .0837612     9.62   0.000     .6415917    .9699296
       disea |   .0746685   .0042856    17.42   0.000     .0662689    .0830681
       hlthg |   .0552906   .0571903     0.97   0.334    -.0568004    .1673816
       hlthf |   .6365347    .108132     5.89   0.000     .4245999    .8484695
       hlthp |   1.196648   .2085426     5.74   0.000      .787912    1.605384
        linc |   .2027749   .0267022     7.59   0.000     .1504394    .2551103
        lfam |    -.33802   .0556753    -6.07   0.000    -.4471416   -.2288984
        xage |   .0081773   .0026603     3.07   0.002     .0029633    .0133914
      female |   1.010896   .0684784    14.76   0.000     .8766804    1.145111
       child |    .911955   .1114265     8.18   0.000     .6935631    1.130347
    femchild |   -1.01086   .1063752    -9.50   0.000    -1.219351   -.8023682
       black |  -1.364137   .0920164   -14.82   0.000    -1.544486   -1.183789
     educdec |   .0433583   .0098894     4.38   0.000     .0239753    .0627412
------------------------------------------------------------------------------
```

4. Visits to the Doctor: Results from Two-Part Models and Zero-Inflated Models

4.1: Zero truncated Poisson regression and logit

```
. ztp mdvis logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage female
child femchild black educdec if mdvis>0

Zero-truncated Poisson regression                 Number of obs   =      13878
                                                  LR chi2(17)     =    4567.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -42502.152                       Pseudo R2       =     0.0510

------------------------------------------------------------------------------
       mdvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logc |  -.0036367   .0064401    -0.56   0.572     -.016259    .0089857
         idp |  -.0794352   .0124901    -6.36   0.000    -.1039152   -.0549551
         lpi |   .0041128   .0019263     2.14   0.033     .0003373    .0078883
        fmde |  -.0233077   .0037312    -6.25   0.000    -.0306207   -.0159947
      physlm |   .2256891   .0125523    17.98   0.000     .2010872    .2502911
       disea |   .0176813   .0006326    27.95   0.000     .0164413    .0189212
       hlthg |   .0382901   .0100467     3.81   0.000     .0185988    .0579813
       hlthf |    .206954   .0167213    12.38   0.000     .1741809    .2397271
       hlthp |   .3998769   .0276018    14.49   0.000     .3457785    .4539754
        linc |   .0445074   .0053239     8.36   0.000     .0340728     .054942
        lfam |  -.1157228   .0094271   -12.28   0.000    -.1341995    -.097246
        xage |    .001439   .0004481     3.21   0.001     .0005608    .0023172
      female |   .1510196   .0118412    12.75   0.000     .1278113    .1742278
       child |   .1664605   .0189914     8.77   0.000     .1292381    .2036828
    femchild |   -.148791   .0188057    -7.91   0.000    -.1856495   -.1119325
       black |  -.2525526   .0165395   -15.27   0.000    -.2849694   -.2201358
     educdec |   .0045414   .0016868     2.69   0.007     .0012354    .0078474
       _cons |    .744609   .0506351    14.71   0.000     .6453661    .8438519
------------------------------------------------------------------------------
```

```
. logit anyvisit logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage
female child femchild black educdec


Logistic regression                               Number of obs   =      20186
                                                  LR chi2(17)     =    2774.48
                                                  Prob > chi2     =     0.0000
Log likelihood = -11149.911                       Pseudo R2       =     0.1107

------------------------------------------------------------------------------
    anyvisit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        logc |  -.1444157   .0227271    -6.35   0.000      -.18896   -.0998713
         idp |  -.2838512   .0444716    -6.38   0.000    -.3710139   -.1966885
         lpi |    .038765   .0073672     5.26   0.000     .0243256    .0532043
        fmde |  -.0090285   .0135284    -0.67   0.505    -.0355436    .0174867
      physlm |   .3088853   .0603161     5.12   0.000     .1906678    .4271028
       disea |   .0350224   .0029613    11.83   0.000     .0292184    .0408265
       hlthg |    .013004   .0366707     0.35   0.723    -.0588693    .0848773
       hlthf |   .2039548   .0697303     2.92   0.003     .0672859    .3406238
       hlthp |   .5762548   .1632094     3.53   0.000     .2563702    .8961394
        linc |   .0980477   .0146601     6.69   0.000     .0693144     .126781
        lfam |  -.0856836   .0349847    -2.45   0.014    -.1542525   -.0171148
        xage |   .0046859   .0017757     2.64   0.008     .0012056    .0081662
      female |   .8021234   .0446338    17.97   0.000     .7146429     .889604
       child |   .6311402   .0677326     9.32   0.000     .4983867    .7638937
    femchild |  -.8717343   .0673375   -12.95   0.000    -1.003713   -.7397553
       black |  -1.259927   .0443694   -28.40   0.000    -1.346889   -1.172965
     educdec |    .054107   .0064066     8.45   0.000     .0415503    .0666638
       _cons |   -1.07807   .1581403    -6.82   0.000    -1.388019   -.7681208
------------------------------------------------------------------------------
```

4.2 Partial effect of logc at the average

Stata code:

```
ztp mdvis logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage female
child femchild black educdec if mdvis>0

margins, predict(cm) atmeans nose
mat junk=r(b)
scalar ey1=junk[1,1]

margins, dydx(logc) predict(cm) atmeans
mat junk=r(b)
scalar dey1=junk[1,1]

logit anyvisit logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage
female child femchild black educdec

margins, predict(p) atmeans nose
mat junk=r(b)
scalar py1=junk[1,1]

margins, dydx(logc) atmeans
mat junk=r(b)
scalar dpy1=junk[1,1]

scalar PEA=py1*dey1+dpy1*ey1
scalar list PEA
```

```
. scalar list PEA
       PEA =  -.1282192
```

Hence I estimate the PEA with respect to logc at -0.128. This is quite similar to the previous estimates. I have now learned, however, that logc primarily seems to affect whether there are any visits to the doctor or not; conditional on positives, logc doesn't seem to affect the number of visits.

To determine whether logc is statistically significant, I use bootstrapping – see next page.

Bootstrapping the standard error of the PEA:

```
ztp mdvis logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage female
child femchild black educdec if mdvis>0

logit anyvisit logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage
female child femchild black educdec

keep if e(sample)==1

save tempdat, replace
mat store=J(100,1,0)

set seed 3246
qui{
forvalues k=1(1)100{
noi disp `k'
use tempdat, clear
bsample

ztp mdvis logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage female
child femchild black educdec if mdvis>0

margins, predict(cm) atmeans nose
mat junk=r(b)
scalar ey1=junk[1,1]

margins, dydx(logc) predict(cm) atmeans nose
mat junk=r(b)
scalar dey1=junk[1,1]

logit anyvisit logc idp lpi fmde physlm disea hlthg hlthf hlthp linc lfam xage
female child femchild black educdec

margins, predict(p) atmeans nose
mat junk=r(b)
scalar py1=junk[1,1]

margins, dydx(logc) atmeans nose
mat junk=r(b)
scalar dpy1=junk[1,1]

scalar PEA=py1*dey1+dpy1*ey1

mat store[`k',1]=PEA
}
}

svmat store
tabstat store*, s(sd)
```

```
. tabstat store*, s(sd)

    variable |        sd
-------------+----------
      store1 |  .0444498
------------------------
. disp -.1282192/.0444498
-2.8845844
```

Hence I estimate the PEA with respect to logc at -0.128, and infer from the bootstrapped standard error that the effect of logc is statistically significant at the 1% level.

Naturally, one could estimate the two-part model using NebBin II if that is thought more appropriate than simple Poisson.