MLE - Limited Dependent Variable Models

Introduction

- Frequently, our data exhibit some non-standard characteristics. In this lecture we will cover three related models:
  - Censored regression models.
  - Truncated regression models
  - Sample selection regression models (Incidental truncation).
- Easiest to understand with examples...

Introduction

- **Censoring**: imagine a survey which collects data on income. Suppose that for those who have an income of more than 1mil DKK, we code it as simple 1 mil + (i.e. not, say, 1.36 mil).
- **Truncation**: now, instead of coding it as 1 mil +, it is simply dropped from the whole analysis. (This removal of data can happen either when the data is collected or when we come to analyse it).
- **Sample selection**: now suppose that we are interested in the amount of tax people pay on this income. However, for privacy reasons, we only observe how much tax people pay if they have an income of less than 1 mil DKK.

Introduction

- The difference between truncation and censoring: in the censored model, everything below (or above) a certain threshold is lumped into one value. For truncated data, everything below (or above) this threshold is not in the data at all, x and y are missing.
- Sample selection is kind of half way: below the threshold we only have access to y or x, not both.

Examples to Test Yourself

- The demand for beer.
- Criminal sentencing... The problem is, we only observe the sentence a defendant receives if they are found guilty. For those not found guilty, we don't know what punishment they would have received.
- Police records... by definition everyone in this sample has at least one arrest. You conduct some causal analysis but want to generalise this to the entire population, i.e. also people with zero arrests.
- Suppose you are now looking at **convictions** from police records (those arrested), rather than arrests, and you care about generalising to all people who have been arrested.

Censored Regression Model

- Although we can handle models without making distributional assumptions, we assume $\epsilon \sim N(0, \sigma^2)$ here

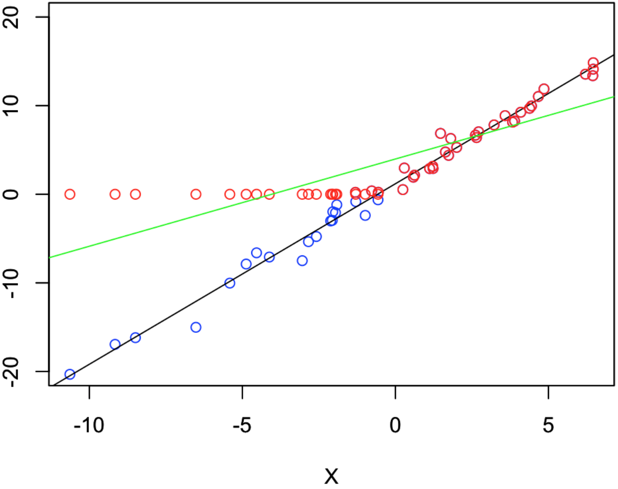
$$y^* = X\beta + \epsilon.$$

- However, we do not observe y^*, we observe $y_i = \max(y_i^*, c)$. In our previous example, y^* was the desire for beer, y is the amount of beer bought, and $c = 0$.
- It is possible to allow c to be different for each observation. In this case we must then assume that c_i is independent of ϵ_i.
- It is also possible to allow for censoring from above, or even from both above and below. Think, for example, demand for a cinema (they only have so many seats).

Censored Regression Model

- If we attempted to remove the problem of censoring by running OLS only on the uncensored observations we would move into the truncated model area (and we'll see later why naive OLS doesn't work there)
- Or, if we were to plough ahead and just use OLS on this censored data we will obtain a bias. This can be seen easily in a plot (throughout we will assume that y is censored from below):

Censored Regression Model



Censored Regression Model

- As with probit/logit, we use MLE to estimate this model.
- For this, we need the density of $y_i|X_i$. For uncensored observations, $y_i = y_i^*$. The density of $y_i^*|X_i$, is given by

$$(2\pi\sigma^2)^{-1/2} exp\left(\frac{-\left(y_i^* - X_i\beta\right)^2}{2\sigma^2}\right)$$

  which can be written as $\frac{1}{\sigma}\phi\left([y_i^* - X_i\beta]/\sigma\right)$, where $\phi(\cdot)$ is the standard normal pdf.

- For censored observations, we need the probability that $y_i = c$ given X_i:

$$\begin{aligned} Pr(y_i = c|X) &= Pr(y_i^* \leq c|X) \\ &= Pr(\epsilon_i \leq c - X\beta|X) \\ &= Pr(\frac{\epsilon_i}{\sigma} \leq \frac{c - X\beta}{\sigma}|X) \\ &= \Phi\left[(c - X\beta)/\sigma\right]. \end{aligned}$$

Censored Regression Model

- Combining, these two parts, the log-likelihood can be written as

$$
\begin{aligned}
lnL(\beta, \sigma; y) &= \sum_{i=1}^{n} I(y_i = c) log\left(\Phi\left[(c - X\beta)/\sigma\right]\right) + \\
&\quad I(y_i > c) log\left(\frac{1}{\sigma}\phi\left([y_i - X_i\beta]/\sigma\right)\right),
\end{aligned}
$$

  where $I(\cdot)$ is the indicator function.

- To maximise this, we need numerical optimsation; we do not get a closed-form solution. (Notice that we also need to estimate σ).

- This is known as the **Tobit estimator**. (It was pioneered by James Tobin)

Interpreting the Coefficients

- It is tempting to interpret the coefficients in the same way as with OLS. However, things aren't quite so simple.
- The β_j's measure the partial effect of x_j on $E[y^*|X]$ not $E[y|X]$. Typically, we are actually interested in the effect on y rather than y^*. Think about the demand for beer, I don't care about how someone's desire changes, I want to know if people will buy more!
- Generally we are interested in two objects
  - $E\left[y|y > c, X\right]$
  - $E\left[y|X\right]$

Interpreting the Coefficients

- If we know $E[y|y > c, X]$ then it is easy to calculate $E[y|X]$ as

$$\begin{aligned} E[y|X] &= Pr(y > c)E[y|y > c, X] + Pr(y \leq c)c \\ &= \Phi\left(\frac{X\beta - c}{\sigma}\right) E[y|y > c, X] + \left[1 - \Phi\left(\frac{X\beta - c}{\sigma}\right)\right] c \end{aligned}$$

- To obtain $E[y|y > c, X]$ we use a result concerning standard normal variables. If $z \sim N(0, 1)$, then $E[z|z > h] = \frac{\phi(h)}{1-\Phi(h)}$ for some constant h.
- Note, in many cases $c = 0$ and things simplify quite nicely.

Interpreting the Coefficients

- We proceed as follows

$$\begin{aligned}
E[y|y>c,X]=E[y^*|y^*>c,X] &= X\beta+E[\epsilon|y^*>c,X] \\
&= X\beta+E[\epsilon|\epsilon>c-X\beta] \\
&= X\beta+\sigma E\left[\frac{\epsilon}{\sigma}|\epsilon>c-X\beta\right] \\
&= X\beta+\sigma E\left[\frac{\epsilon}{\sigma}|\frac{\epsilon}{\sigma}>\frac{c-X\beta}{\sigma}\right] \\
&= X\beta+\sigma\frac{\phi(\{c-X\beta\}/\sigma)}{1-\Phi(\{c-X\beta\}/\sigma)} \\
&= X\beta+\sigma\frac{\phi\left(\{X\beta-c\}/\sigma\right)}{\Phi\left(\{X\beta-c\}/\sigma\right)} \\
&\equiv X\beta+\sigma\lambda\left(\{X\beta-c\}/\sigma\right),
\end{aligned}$$

since $\phi(-h)=\phi(h)$ and $1-\Phi(-h)=\Phi(h)$. $\lambda(\cdot)$ is known as the inverse Mills ratio.

Interpreting the Coefficients

- What we have shown is that the naive OLS regression of y on X, will be biased because it does not include the omitted variable $\lambda(\{X\beta - c\}/\sigma)$ (which, in general is correlated with X).

- Taking $c = 0$ for ease, we can obtain partial effects as follows:

$$\frac{\partial E[y|y > 0, X]}{\partial x_j} = \beta_j + \beta_j \left(\frac{\partial}{\partial z} \lambda(z) \right),$$

  it can be shown that $\frac{\partial}{\partial z}\lambda(z) = \frac{-\lambda(z)}{z+\lambda(z)}$.

- Hence

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

Comments

- The β_j gets multiplied by some adjustment factor. It can be shown that this adjustment lies in $[0, 1]$.

- Using similar techniques, it is possible to show

$$\frac{\partial E[y|X]}{\partial x_j} = \beta_j \Phi\left(X\beta/\sigma\right).$$

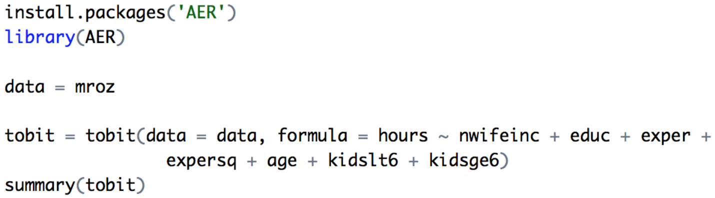
- We can report an R^2 as in a linear regression and think of it in the same way, however, it isn't quite the same and we shouldn't interpret it as such.

- It's calculated as the square of the correlation between y and $\hat{y}$.

- Note, the Tobit estimator requires Normality and Homoskedasticty for it to be unbiased/consistent.

Example: Women's Labour Supply (mroz)

- To try out our Tobit estimator we look at womens' hours worked.
- If a woman chooses to not enter the formal labour market, her hours are 0, otherwise it is something positive.



```
install.packages('AER')
library(AER)

data = mroz

tobit = tobit(data = data, formula = hours ~ nwifeinc + educ + exper +
                    expersq + age + kidslt6 + kidsge6)
summary(tobit)
```

Example: Women's Labour Supply

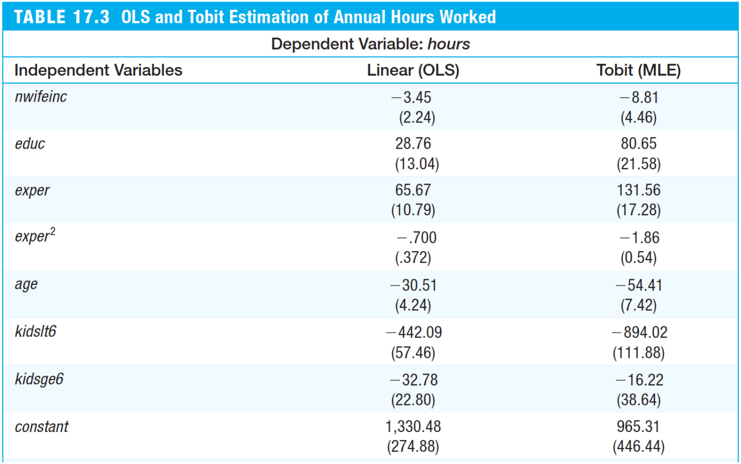


TABLE 17.3 OLS and Tobit Estimation of Annual Hours Worked

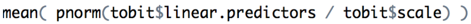
Dependent Variable: *hours*		
Independent Variables	Linear (OLS)	Tobit (MLE)
nwifeinc	−3.45 (2.24)	−8.81 (4.46)
educ	28.76 (13.04)	80.65 (21.58)
exper	65.67 (10.79)	131.56 (17.28)
exper^2	−.700 (.372)	−1.86 (0.54)
age	−30.51 (4.24)	−54.41 (7.42)
kidslt6	−442.09 (57.46)	−894.02 (111.88)
kidsge6	−32.78 (22.80)	−16.22 (38.64)
constant	1,330.48 (274.88)	965.31 (446.44)

(Credit: Wooldridge 2016)

Example: Women's Labour Supply

- We should be careful not to compare these coefficients directly.

- We should multiply by the adjustment factor $\Phi\left(\frac{X\beta}{\sigma}\right)$. We can calculate this in R as $\frac{1}{n}\sum_i \Phi\left(\frac{X\beta}{\sigma}\right)$



```
mean( pnorm(tobit$linear.predictors / tobit$scale) )
```

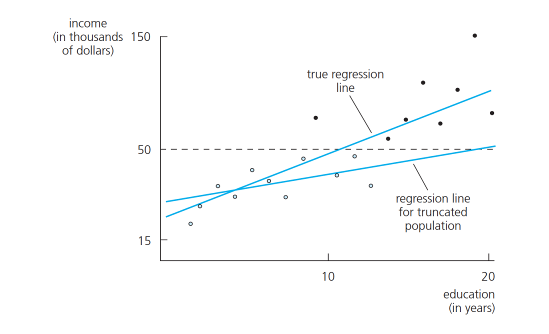
- This gives 0.589. When we multiply the coefficients by this adjustment, the coefficients are much closer, however, in general, they are still larger in magnitude.

- For example, an extra year of education increases hours per year by 29 (roughly half an hour per week) according to OLS. But according to the Tobit model 0.589*80.65=47.5 (close to one hour per week).

Truncation

- The key issue here, is that we don't have a random sample of our population. Suppose we are conducting a survey, and for some reason we cannot survey people who earn more than $50 000, but we want to generalise our results to the whole population... this could be a problem:

FIGURE 17.4 **A true, or population, regression line and the incorrect regression line for the truncated population with observed incomes below $50,000.**



(Credit: Wooldridge 2016)

Truncation

- To solve this issue, we start where we normally do:

$$y = X\beta + u$$

  where $u|X \sim N(0, \sigma^2)$, i.e. our standard linear model, so $y|X \sim N(X\beta, \sigma^2)$. Our model is not the problem, it is our data!

- To estimate β we need the distribution of y given X and that $y \leq c$. (Note that we can just as easily incorporate $y \geq c$)

- Let $f_{X\beta,\sigma^2}(y)$ denote the normal density with mean $X\beta$ and variance σ^2, let $F_{X\beta,\sigma^2}(y)$ be the analogous CDF. Then we can write the distribution of y given that $y \leq c$ and X as

$$\frac{f_{X\beta,\sigma^2}(y)}{F_{X\beta,\sigma^2}(c)}.$$

Truncation

- This expression makes sense when thought of in relation to Bayes theorem. On the top we have the population density of y and on the bottom we have the probability that y is observed.

- It remains only to set up the log likelihood function and maximise it to obtain our β coefficients.

- This was pretty straightforward.

- To do this in R, you can use the 'truncreg' package which has a trunreg() function. You just have to specify the truncation threshold and which direction the truncation is... above (right) or below (left).

When is OLS Fine?

- Take a look back at the plot with truncation according to y.
- What happens if the truncation is due to x, e.g. we don't observe anyone who has an education above 15 years.
- Everything is fine!
- Let's add a bit more formality to this idea...

When is OLS Fine?

- Let $s_i = 1$ if we observe the observation, and 0 otherwise.
- We can think of our model as being from

$$s_i y_i = s_i X_i \beta + s_i u_i,$$

  where we now actually have a random sample.
- For us to use OLS on this model to recover β, the key condition is $E\left[su|sX\right] = 0$.
- If s is a function of only X then $sE\left[u|sX\right] = 0$, which holds under $E\left[u|X\right] = 0$.

When is OLS Fine?

- Equally, if s is determined randomly, we reach the same conclusion and can just use OLS.
- But when $s_i = 1$ if $y_i \leq c$, i.e. $u_i \leq c - X_i\beta$. So s and u are intrinsically linked, and so will not be uncorrelated even conditional on X. Hence $E\left[su|sX\right] \neq 0$ and OLS is biased.
- This is not the only way that s and u can be related. In the next slide we consider incidental truncation and Heckman's sample selection (2-step) estimator. (James Heckman)

Sample Selection (Incidental Truncation)

- Suppose we have the following model

$$\begin{aligned} y &= X\beta + u, \text{ where } E(u|x) = 0 \\ s &= I[Z\gamma + v \geq 0], \end{aligned}$$

  where $s = 1$ if we have access to the observation.

- The **selection equation** depends on observables Z which satisfies $E[u|x,z] = 0$ and where $X \subset Z$. For ease, also assume $v \sim Normal$.

- If u and v are related (very likely - think wage equation and intelligence), we have a selection problem...

Sample Selection

- We want to look at the expectation of y conditional on Z, X, and $s = 1$. To help see what's going on though, we first notice

$$E\left[u|v, Z, s\right] = E\left[u|v\right] = \rho v$$

  for some parameter ρ (if u and v are jointly normal and zero mean), and using the fact that (u, v) is independent of Z and u is independent of s conditional on v.

- We also need to make use of a generalisation of the law of total expectations which says

$$E\left[A|B\right] = E\left[E\left[A|B, C\right]|B\right].$$

Sample Selection

- Using these two results, we can write

$$\begin{aligned} E\left[y|Z,s=1\right] &= X\beta + E\left[u|Z,s=1\right] \\ &= X\beta + E\left[E(u|Z,v,s=1)|Z,s=1\right] \\ &= X\beta + \rho E\left[v|Z,s=1\right] \\ &= X\beta + \rho\lambda(Z\gamma), \end{aligned}$$

  where $\lambda(\cdot)$ is the inverse Mills ratio again.

- The final equality comes from using the same type of argument we used on slide 13.

Sample Selection

- So, we again have this similar idea of an omitted inverse Mills ratio term in our model. What this also shows is that we can estimate β, our parameter of interest, even if we only have data on a subsample, providing we account for $\lambda(Z\gamma)$.
- Of course, we do not know γ. So we need to first estimate this, then create $\lambda(Z\gamma)$, and finally run the regression of y on X and $\lambda(Z\gamma)$.

Heckman 2-step Estimator

1. Using the **entire sample** estimate a probit model for s on Z and obtain $\hat{\gamma}$. Compute $\hat{\lambda}_i = \lambda(Z_i \hat{\gamma})$.
2. Using the selected sample, i.e. where $s = 1$, run the regression of y_i on X_i and $\hat{\lambda}_i$.

- Using this method, $\hat{\beta}$ is consistent and asymptotically normally distributed.
- **Bonus**: We can also test for whether sample section is a problem by using a t-test on the coefficient on $\hat{\lambda}_i$.

Heckman 2-step Estimator - Issue

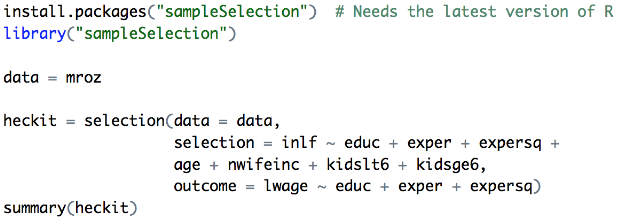
- Recall, we said that $X \subset Z$, this turns out to be quite important. That is, we need something that affects selection but **not** the outcome. In many cases this can be quite difficult to find.

- The reason is this: suppose $X = Z$, although $\lambda(X\gamma)$ is a nonlinear function of X, it is often well approximated by a linear function. In which case, in our final regression, X_i and $\hat{\lambda}_i$ are very highly correlated and our results become unconvincing.

- Intuitively, if we don't have something which affects selection but not the outcome, there is no way to distinguish the effect of selection on y from the effect of X on y.

Example: Women's Wages (mroz)

- We use the same dataset as before, but we now want to know the causal effects on women's wages, accounting for the fact that we only see the wages of women who participate in formal employment.



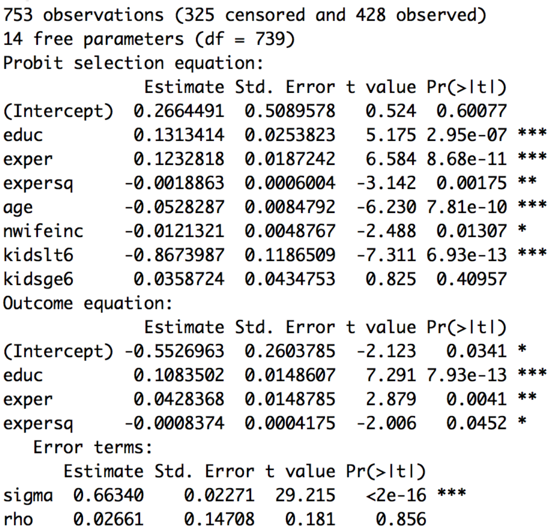
```
install.packages("sampleSelection")  # Needs the latest version of R
library("sampleSelection")

data = mroz

heckit = selection(data = data,
                   selection = inlf ~ educ + exper + expersq +
                   age + nwifeinc + kidslt6 + kidsge6,
                   outcome = lwage ~ educ + exper + expersq)
summary(heckit)
```

- Notice that we assume *age*, *nwifeinc*, *kidslt6*, and *kidsge6* don't affect a woman's wage.

Example: Women's Wages (Heckit)

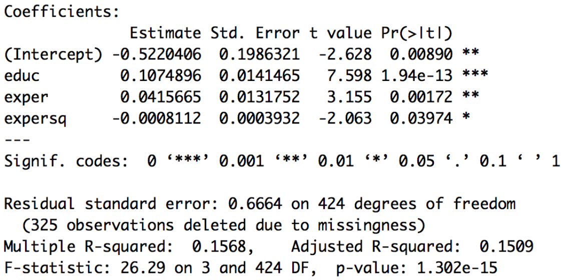


```
753 observations (325 censored and 428 observed)
14 free parameters (df = 739)
Probit selection equation:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2664491  0.5089578   0.524  0.60077
educ         0.1313414  0.0253823   5.175 2.95e-07 ***
exper        0.1232818  0.0187242   6.584 8.68e-11 ***
expersq     -0.0018863  0.0006004  -3.142  0.00175 **
age         -0.0528287  0.0084792  -6.230 7.81e-10 ***
nwifeinc    -0.0121321  0.0048767  -2.488  0.01307 *
kidslt6     -0.8673987  0.1186509  -7.311 6.93e-13 ***
kidsge6      0.0358724  0.0434753   0.825  0.40957
Outcome equation:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.5526963  0.2603785  -2.123   0.0341 *
educ         0.1083502  0.0148607   7.291 7.93e-13 ***
exper        0.0428368  0.0148785   2.879   0.0041 **
expersq     -0.0008374  0.0004175  -2.006   0.0452 *
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma  0.66340    0.02271  29.215   <2e-16 ***
rho    0.02661    0.14708   0.181    0.856
```

Example: Women's Wages

- Interestingly, we don't seem to have any sample selection issue. We can see this from the ρ coefficient being insignificant.
- This is means that there is no difference in the relationship between wages and education and experience for women in the formal labour market and those who are not.
- So we could have just run an OLS regression on the subsample of women who have a wage...

Example: Women's Wages (OLS)



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.5220406  0.1986321  -2.628  0.00890 **
educ         0.1074896  0.0141465   7.598 1.94e-13 ***
exper        0.0415665  0.0131752   3.155  0.00172 **
expersq     -0.0008112  0.0003932  -2.063  0.03974 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6664 on 424 degrees of freedom
  (325 observations deleted due to missingness)
Multiple R-squared:  0.1568,    Adjusted R-squared:  0.1509
F-statistic: 26.29 on 3 and 424 DF,  p-value: 1.302e-15
```

- Notice that the coefficients are very similar - this is exactly what we thought would happen.
- Note, technically the standard errors from the Heckit model are not valid as they don't account for error from the first stage estimation.

Summary

- We have learnt how to estimate three new models:

1. Censored regression
2. Truncated regression
3. Sample selection

- Each of these models is defined by the type of data we have.
- While the first two models are estimated using maximum likelihood, the third model requires two stages (MLE for probit, then regression).

Examples to Test Yourself - Answers

- Beer demand - censored from below at 0. Everyone with zero or negative demand for beer is lumped together at 0.
- Criminal sentencing - sample selection. We only observe y (sentence) if x (level of guilt) is above some threshold.
- Police records - truncated from below at 0. Anyone with less than 1 arrest is not in the data at all.
- Convictions - now there's no truncation issue as people in the data may have 0 or more convictions. (You could maybe argue for a censored model if you're thinking about 'taste for committing crime').