PPOL 503-03, PPOL 503-04, Fall 2016
Course Notes #12: Tobit, Censored and Truncated Regression Models

---

I. Tobit Regression Models

- Example: Suppose we have cross-sectional data on car purchases by individuals in a given year. We want to estimate the impact of income on the amount of money spent on a car in the year. Car buyers have positive expenditures, which can reasonably be treated as continuous random variables, but non-buyers spent $0. Thus the distribution of car expenditures is a combination of a discrete distribution (at zero) and a continuous distribution.

- We could drop the observations of people who did not purchase a car and regress the amount spent on each individual's car against income. However, this would neglect those who did not purchase a car at all. These people also hold a valuation of a car (which depends on income), and it is misleading to omit them from the analysis. Such an OLS estimation would thus lead to biased and inconsistent coefficient estimates.

- It is also misleading to fit a line (using OLS) through the data when many of the observations are equal to zero. This also leads to biased and inconsistent estimates (though the magnitude of the bias will depend on what proportion of the data are at zero).

- Other examples in the literature:

    - Household purchases of durable goods
    - Hours of female labor force participation
    - The number of arrests after release from prison
    - The number of tickets demanded for events at a certain arena
    - Charitable giving by households

    In each case, conventional regression methods fail to account for the qualitative difference between the limit (zero) observations and the non-limit (continuous) observations.

- We can consider the observed response, Y, in terms of an underlying latent variable, Y^*:

$$Y^* = \beta_0 + \beta_1 X + u, \text{ where } u|X \sim \text{Normal } (0,\sigma^2)$$
$$Y = \max (0, Y^*)$$

- The latent variable equation satisfies the classical linear model assumptions. The problem is that Y cannot be below 0.

- As with logit and probit models, we use MLE to estimate the Tobit model. Recall that MLE chooses coefficient estimates that maximize the likelihood of the sample data set being observed. For the values of $Y > 0$, we use the equation for the normal probability distribution function. The normal PDF for $Y > 0$ given X is the same as the normal PDF for Y^* given X. We know the equation for this because we assumed above that u|x has a normal distribution with mean 0 and variance σ^2. For the values of $Y = 0$, the probability distribution is equal to $1 - \Phi[(\beta_0 + \beta_1 X)/ \sigma]$, where Φ is the standard normal *cumulative* distribution function. SAS and Stata can easily compute this Tobit MLE.

- The Tobit coefficient estimates are the estimated impacts of the independent variable on the expected value of the latent variable, Y^*. This is not the same as the estimated impact of the independent variable on the expected value of the observed variable, Y. Unless the latent variable is what is of interest, you can't just interpret the coefficient estimates as marginal effects.

- The change in the expected value of Y given a change in X (holding all else equal) is equal to the Tobit coefficient β_1 multiplied by an adjustment factor (which is between 0 and 1) of $\Phi[(\beta_0 + \beta_1 X)/ \sigma]$. For dummy variables, this is called the marginal effect. For continuous variables, we calculate the elasticity. The elasticity is calculate by multiplying the ratio of the mean of X_i divided by the mean of Y.

- To obtain the Tobit slope coefficients, McDonald and Mofitt (1980) have shown, that the CDF of the standardized Tobit index is approximately equal to the proportion of non-limit observations.

- Example: Labor supply determinants for married women:

```
> We seek to estimate the determinants of female labor
> force participation (hours).  Independent variables:
> Husband's earnings ($1000s), education (educ), experience (exper),
> experience squared, age, number of children less than 6 (kidslt6),
> and number of kids between 6 and 18 (kidsge6).
> */
>
> gen exper2=exper*exper;

. /*
> First we estimate the OLS equation,
> in which 43.6% of the observations have hours=0.
> */
>
> regress hours nwifeinc educ exper exper2 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     3.329284    54.19297
       exper |   65.67251   9.962983     6.59   0.000     46.11365    85.23138
      exper2 |  -.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
```

Interpretation of OLS results (does not account for censoring)

Educ: each additional year of education increases annual hours of work by about 29 hours.

Age: each additional year of age reduces annual hours by about 30.5 hours.

Kidslt6: Women with one or more children under age 6 work on average 442 fewer hours than women without children under age 6.

```
. /*Next we estimate the tobit model.*/
>
> tobit hours nwifeinc educ exper exper2 age kidslt6 kidsge6, ll;

Tobit estimates                                   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   -.0603725
        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
      exper2 |  -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     88.88531    1841.725
-------------+----------------------------------------------------------------
         _se |   1122.022   41.57903           (Ancillary parameter)
------------------------------------------------------------------------------

  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
```

To obtain the Tobit slope coefficients, multiply each coefficient by .57—the proportion of non-limit observations.

Educ: (80.65 x .57) = 45.97 each additional year of schooling increases annual hours by 46 hours.

Age: (-54.40 x .57)= -31. Each additional year of age is associated with 31 fewer annual hours of work.

Kidslt6: (-894.02 x .57) = -509.6 Women with one or more children under age 6 work on average 510 fewer hours than women without children under age 6.

Comparing the Tobit results to the OLS results illustrates how the OLS results (which ignore the censoring at zero) are biased towards zero. Since the educ variable has a positive coefficient, the OLS results understate the effect of educ on hours of work. Since the coefficient on kidslt6 is negative, the OLS results are biased towards zero and thus understate the effect of having kidslt6 on hours of work.

- One note: Tobit relies heavily on the assumption of the normality of the disturbance term with mean zero and constant variance. If these assumptions fail, then the Tobit model may be meaningless. For moderate departures from these assumptions, Tobit can provide good estimates of partial effects.

II. CENSORED REGRESSION MODELS

- Sometimes we don't have complete values for the dependent variable. Instead, the dependent variable is *censored*, typically due to survey design or institutional constraints.
- In the case of data censoring, we do randomly sample units from the population. The censoring problem is that while we always observe the explanatory variables for each randomly drawn unit, we observe the outcome for *y* only when it is not censored above or below a given threshold.

- Examples:
  - An earnings model in which some respondents' income are recorded as "in excess of $100,000." (*Right-censored)*
  - An earnings model in which some respondents' income are recorded as "below $20,000." (*Left-censored)*

- We can consider the observed response, Y, in terms of an underlying latent variable, Y^*:

$$Y^* = \beta_0 + \beta_1 X + u, \text{ where } u|X,c \sim \text{Normal } (0,\sigma^2)$$
$$Y = \min (c, Y^*)$$

- Analyzing censored data is similar to the Tobit analysis, it just uses a different MLE equation for right-censored versus left-censored and for different values of c.

- The good news is that, unlike Tobit models, the estimated regression coefficients yield marginal effects. **The censored models essentially adjust for the missing dependent variable values, whereas the Tobit models effectively incorporate the corner-solutions into the functional form.**

- Book Example: The dependent variable is the time in months until an inmate in a North Carolina prison is arrested from prison—*durat*, Time in months until an inmate is arrested after being released from prison. (Note for this example, the data is right censored and the censoring times differ for each censored observation.) Of the 1,445 inmates, 893 had not been arrested during the period they were followed, so these observations are censored.

The censoring times differed among inmates ranging from 70 to 81 months.

```
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
black           byte   %9.0g                  =1 if black
alcohol         byte   %9.0g                  =1 if alcohol problems
drugs           byte   %9.0g                  =1 if drug history
super           byte   %9.0g                  =1 if release supervised
married         byte   %9.0g                  =1 if married when incarc.
felon           byte   %9.0g                  =1 if felony sentence
workprg         byte   %9.0g                  =1 if in N.C. pris. work prg.
property        byte   %9.0g                  =1 if property crime
person          byte   %9.0g                  =1 if crime against person
priors          byte   %9.0g                  # prior convictions
educ            byte   %9.0g                  years of schooling
rules           byte   %9.0g                  # rules violations in prison
age             int    %9.0g                  in months
tserved         int    %9.0g                  time served, rounded to months
follow          byte   %9.0g                  length follow period, months
durat           byte   %9.0g                  max(time until return, follow)
cens            byte   %9.0g                  =1 if duration right censored
ldurat          float  %9.0g                  log(durat)
------------------------------------------------------------------------------
. cnreg ldurat workprg priors tserved felon alcohol drugs black married educ ag
> e, censored(cens)

Censored normal regression                      Number of obs   =       1445
                                                LR chi2(10)     =     166.74
                                                Prob > chi2     =     0.0000
Log likelihood =  -1597.059                     Pseudo R2       =     0.0496
------------------------------------------------------------------------------
      ldurat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     workprg |  -.0625715   .1200369    -0.52   0.602    -.2980382    .1728951
      priors |  -.1372529   .0214587    -6.40   0.000    -.1793466   -.0951592
     tserved |  -.0193305   .0029779    -6.49   0.000    -.0251721    -.013489
       felon |   .4439947   .1450865     3.06   0.002     .1593903    .7285991
     alcohol |  -.6349092   .1442166    -4.40   0.000    -.9178072   -.3520113
       drugs |  -.2981602   .1327355    -2.25   0.025    -.5585367   -.0377837
       black |  -.5427179   .1174428    -4.62   0.000    -.7730958     -.31234
     married |   .3406837   .1398431     2.44   0.015      .066365    .6150024
        educ |   .0229196   .0253974     0.90   0.367    -.0269004    .0727395
         age |   .0039103   .0006062     6.45   0.000     .0027211    .0050994
       _cons |   4.099386    .347535    11.80   0.000     3.417655    4.781117
-------------+----------------------------------------------------------------
         _se |    1.81047   .0623022           (Ancillary parameter)
------------------------------------------------------------------------------

  Obs. summary:        552     uncensored observations
                       893 right-censored observations
```

Interpretation of Results

Each of the regression coefficients, when multiplied by 100, gives the estimated percentage change in expected duration given a ceteris paribus increase of one unit in the explanatory variable.

1. *priors* (number of prior convictions): an inmate with one more prior conviction has a duration until next arrest that is almost 14% less.

2. *tserved* (total months spent in prison): a year of time served reduces duration by about [100 x 12 x .019] = 22.8%.

3. *workprg* (participation in a work program) does not have a significant effect on recidivism durations.

4. *felony*(=1 if man is serving time for a felony) A man serving time for a felony has an estimated expected duration that is almost 56% longer (exp (.444) – 1= .56) than a man serving time for a nonfelony.

5. Black men have substantially shorter expected durations until the next arrest –about 42% [exp (-.543) – 1 = .42].

In this application is critical to account for the censoring because almost 62% of the durations are censored. If we apply OLS to the entire sample and treat the censored observations as if they were uncensored, the coefficient estimates are markedly different. The coefficient on *priors* becomes -.059 (se = .009) and that on the alcohol variables is -.262 (se = .060). The coefficients are all biased towards zero.

III. TRUNCATED REGRESSION MODELS

- A truncated regression model is similar to censored regression model, except that we do not observe *any* information about a certain segment of the population beyond the censoring point. In other words, for some group of individuals we have no data on the dependent and the independent variables, so we don't have a random sample.

- Examples:
    - An analysis of the negative income tax experiment in which only families that had income less than 1.5 times the 1967 poverty line are included in the data.
    - An analysis of test determinants using a data set containing people who scored below the thirtieth percentile on the Armed Forces Qualification Test.

- Applying OLS to a truncated sample leads to biased and inconsistent estimates. In general, a sample truncated from above produces estimators biased towards zero.

- Suppose that the true model is $Y^* = \beta_0 + \beta_1 X + u$, where $u|X \sim \text{Normal}(0,\sigma^2)$. However, the sample only includes observations in which $Y^* <= c$. As before, we estimate truncated regression models using MLE. The probability of observing Y^*, given that it is less than or equal to c is equal to the probability of observing Y^* divided by the probability that Y^* is less than or equal to c. Truncated regression models use MLE and adjust for the part of the distribution of u that is "cut-off" or truncated at c.

- As an example, let's re-examine the female labor participation data used earlier. Only now we'll trim the data so that we only have information on women who are in the labor force. (Note: It is a bad idea to delete useful information. This is only for teaching purposes.) **Note how OLS is biased towards zero.**

```
> We seek to estimate the determinants of female labor
> force participation (hours).  Independent variables:
> Husband's earnings ($1000s), education (educ), experience (exper),
> experience squared, age, number of children less than 6 (kidslt6),
> and number of kids between 6 and 18 (kidsge6). In this example, we do not
have data on women who work zero hours.
> gen exper2=exper*exper;
```

```
> First we estimate the OLS equation, for working women.
>
> regress hours nwifeinc educ exper exper2 age kidslt6 kidsge6 if hours>0;

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  7,   420) =    9.79
       Model |  36102842.8     7  5157548.98           Prob > F      =  0.0000
    Residual |   221208177   420  526686.136           R-squared     =  0.1403
-------------+------------------------------           Adj R-squared =  0.1260
       Total |   257311020   427   602601.92           Root MSE      =  725.73

------------------------------------------------------------------------------
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |   .4438515   3.613498     0.12   0.902    -6.658941    7.546644
        educ |  -22.78841   16.43448    -1.39   0.166    -55.09248    9.515669
       exper |   47.00509   14.55649     3.23   0.001     18.39244    75.61774
      exper2 |  -.5136442   .4373576    -1.17   0.241    -1.373327    .3460382
         age |  -19.66352   5.894026    -3.34   0.001    -31.24899   -8.078058
     kidslt6 |  -305.7209   96.45007    -3.17   0.002    -495.3058   -116.1359
     kidsge6 |  -72.36673   30.36099    -2.38   0.018    -132.0451   -12.68832
       _cons |   2056.643   346.4843     5.94   0.000     1375.583    2737.702
------------------------------------------------------------------------------

. /*Next we estimate the MLE truncated regression model.*/
>
> truncreg hours nwifeinc educ exper exper2 age kidslt6 kidsge6, ll(0);
(note: 325 obs. truncated)

Fitting full model:

Iteration 0:   log likelihood = -3401.0465
Iteration 1:   log likelihood = -3390.7888
Iteration 2:   log likelihood = -3390.6478
Iteration 3:   log likelihood = -3390.6476

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]
-------------+----------------------------------------------------------------
eq1          |
    nwifeinc |     .15344   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
      exper2 |  -.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        |
       _cons |    850.766   43.80097    19.42   0.000     764.9177    936.6143
------------------------------------------------------------------------------
```

- Whether OLS is preferred to the truncated regression model depends on the purpose of the estimation. If we are interested in the impacts on a wife's working hours conditional on the sub-sample of labor market participation, then OLS is appropriate. (Alternatively, Stata can do the MLE truncation regression model and then adjust the coefficient estimates to compute the marginal effects for the sub-sample of labor market participants. This should be close to the OLS estimates.) However, if we are interested in the impacts on a wife's working hours regardless of market or non-market labor status, then OLS will be biased and inconsistent.