What are some examples of Tobit analysis in Stata for data analysis?

Authored by

stats writer

June 29, 2024

Tobit analysis is a statistical method used to analyze data that contains censored or truncated values. In Stata, this technique can be applied to various types of data, such as income, expenditures, and length of stay in a hospital. For instance, Tobit analysis can be used to examine the factors that influence individuals' income levels, even when some individuals have zero or very low incomes. Additionally, it can be used to investigate the determinants of household expenditures, taking into account that some households may not spend any money on certain items. Another example is using Tobit analysis to understand the factors that affect the length of stay for patients in a hospital, considering that some patients may have a shorter or longer stay than others. Overall, Tobit analysis in Stata is a valuable tool for analyzing data with censored or truncated values, allowing researchers to gain valuable insights into various socioeconomic and healthcare-related phenomena.

Tobit Analysis | Stata Data Analysis Examples

Version info: Code for this page was tested in Stata 12.

The tobit model, also called a censored regression model, is designed to estimate linear relationships between variables when there is either left- or right-censoring in the dependent variable (also known as censoring from below and above, respectively). Censoring from above takes place when cases with a value at or above some threshold, all take on the value of that threshold, so that the true value might be equal to the threshold, but it might also be higher. In the case of censoring from below, values those that

fall at or below some threshold are censored.

Please Note: The purpose of this page is to show how to use various data analysis commands. It does not cover all aspects of the research process which researchers are expected to do. In particular, it does not cover data cleaning and checking, verification of assumptions, model diagnostics and potential follow-up analyses.

Examples of tobit regression

Example 1.

In the 1980s there was a federal law restricting speedometer readings to no more than 85 mph. So if you wanted to try and predict a vehicle's top-speed from a combination of horse-power and engine size, you would get a reading no higher than 85, regardless of how fast the vehicle was really traveling. This is a classic case of right-censoring (censoring from above) of the data. The only thing we are certain of is that those vehicles were traveling at least 85 mph.

Example 2. A research project is studying the level of lead in home drinking water as a function of the age of a house and family income. The water testing kit cannot detect lead concentrations below 5 parts per billion (ppb). The EPA considers levels above 15 ppb to be dangerous. These data are an example of left-censoring (censoring from below).

Example 3. Consider the situation in which we have a measure of academic aptitude (scaled 200-800) which we want to model using reading and math test scores, as well as, the type of program the student is enrolled in (academic, general, or vocational). The problem here is that students who answer all questions on the academic aptitude test correctly receive a score of 800, even though it is likely that these students are not "truly" equal in aptitude. The same is true of students who answer all of the questions incorrectly. All such

students would have a score of
200, although they may not all be of equal aptitude.

Description of the data

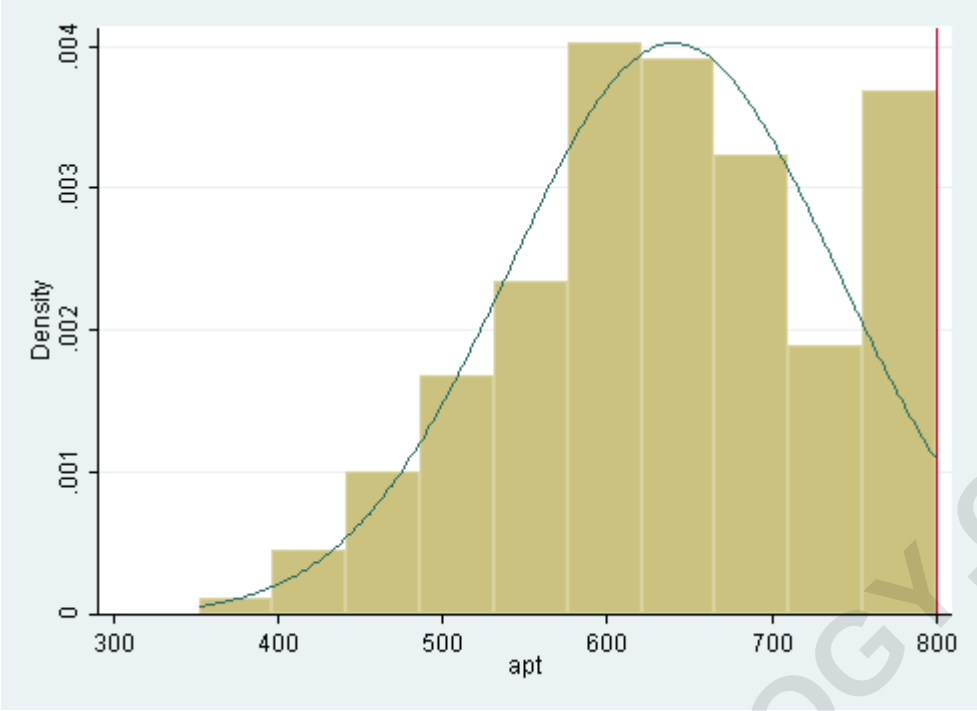
Let's pursue Example 3 from above.

We have a hypothetical data file, tobit.dta with 200 observations.
The academic aptitude variable is apt, the reading and math test scores are read
and math respectively. The variable prog is the type of program
the student is in, it is a categorical (nominal) variable that takes on three
values, academic (prog = 1), general (prog = 2), and vocational (prog
= 3).

Let's look at the data.
Note that in this dataset, the lowest
value of apt is 352. No students received a score of 200 (i.e. the lowest
score possible), meaning that even though censoring from below was possible, it

does not occur in the dataset.

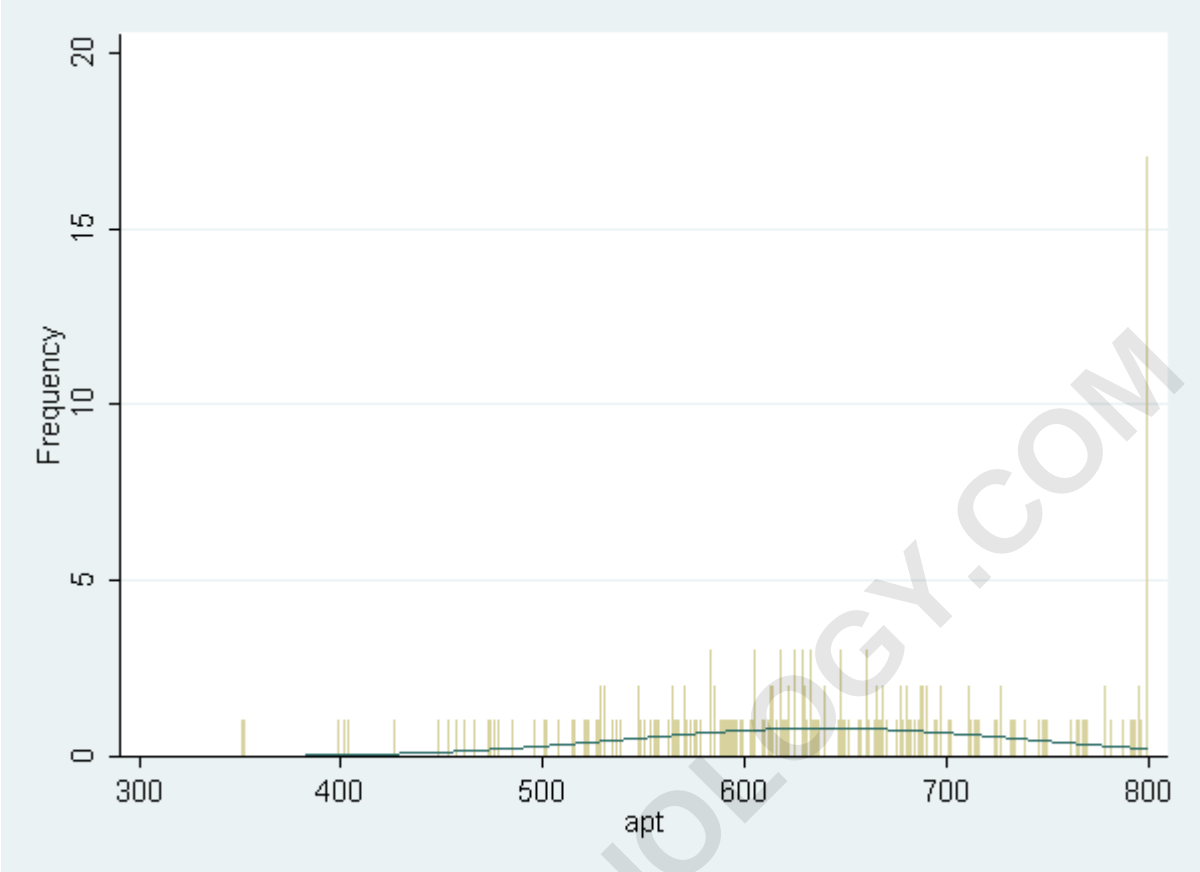
```
use https://stats.idre.ucla.edu/stat/stata/dae/tobit, clear
summarize apt read math

    Variable | Obs Mean Std. Dev. Min Max
-------------+--------------------------------------------------------
         apt | 200 640.035 99.21903 352 800
        read | 200 52.23 10.25294 28 76
        math | 200 52.645 9.368448 33 75

tabulate prog

   type of |
   program | Freq. Percent Cum.
-----------+-----------------------------------
  academic | 45 22.50 22.50
   general | 105 52.50 75.00
vocational | 50 25.00 100.00
-----------+-----------------------------------
     Total | 200 100.00

histogram apt, normal bin(10) xline(800)
```



Looking at the above histogram showing the distribution of apt, we can see the censoring in the data, that is, there are far more cases with scores of 750 to 800 than one would expect looking at the rest of the distribution. Below is an alternative histogram that further highlights the excess of cases where apt=800. In the histogram below, the discrete option produces a histogram where each unique value of apt has its own bar. The freq option causes the y-axis to

be labeled with the frequency for each value, rather than the density. Because
apt is continuous, most values of apt are unique in the dataset,
although close to the center of the distribution there are a few values of
apt that have two or three cases. The spike on the far right of the
histogram is the bar for cases where apt=800, the height of this bar
relative to all the others clearly shows the excess number of cases with this value.

```
histogram apt, discrete freq
```

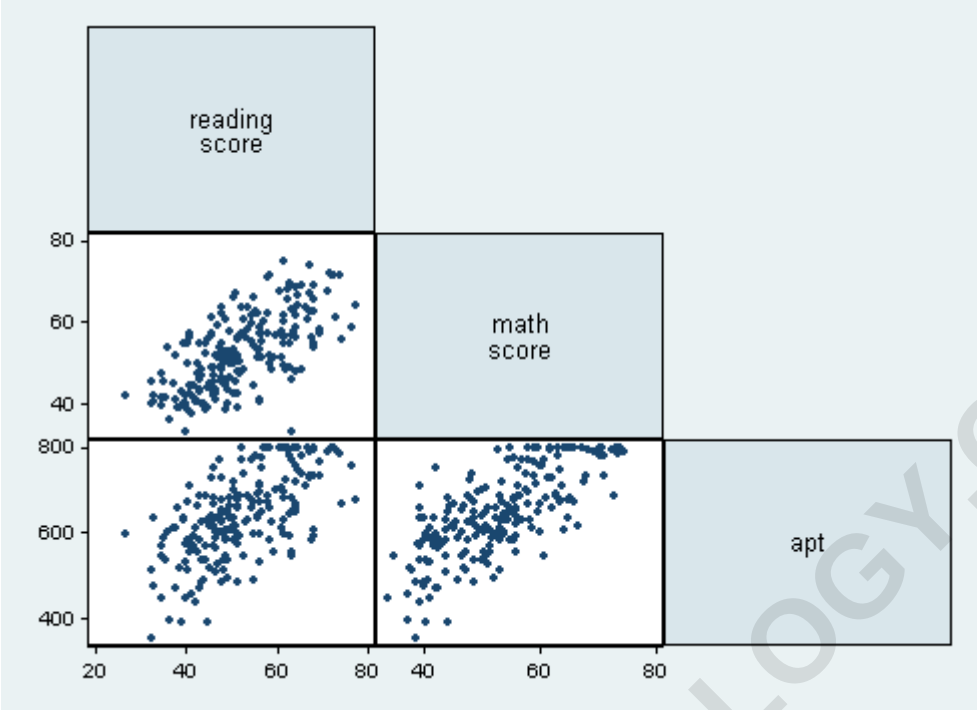


Next we'll explore the bivariate relationships in our dataset.

correlate read math apt
(obs=200)

| read math apt
----------+---------------------------
read | 1.0000
math | 0.6623 1.0000
apt | 0.6451 0.7333 1.0000

```
graph matrix read math apt, half jitter(2)
```



In the last row of the scatterplot matrix shown above, we see the scatterplots showing read and apt, as well as math and apt. Note the collection of cases at the top of each scatterplot due to the censoring in the distribution of apt.

Analysis methods you might consider

Below is a list of some analysis methods you may have encountered. Some of the methods listed are quite reasonable while

others have either
fallen out of favor or have limitations.

Tobit regression

Below we run the tobit model, using read, math, and prog to predict apt. The ul( ) option in the tobit command indicates the value at which the right-censoring
begins (i.e., the upper limit). There is also a ll( ) option to indicate the value of the left-censoring
(the lower limit) which was not needed in this example. The i. before
prog indicates that prog is a factor
variable (i.e.,
categorical variable), and that it should be included in the model as a series
of dummy variables. Note that this syntax was introduced in Stata 11.

```
tobit apt read math i.prog, ul(800)

Tobit regression Number of obs = 200
LR chi2(4) = 188.97
Prob > chi2 = 0.0000
```

Log likelihood = -1041.0629 Pseudo R2 = 0.0832

--

apt | Coef. Std. Err. t P>|t|

-------------+--

read | 2.697939 .618798 4.36 0.000 1.477582 3.918296

math | 5.914485 .7098063 8.33 0.000 4.514647 7.314323

|

prog |

2 | -12.71476 12.40629 -1.02 0.307 -37.18173 11.7522

3 | -46.1439 13.72401 -3.36 0.001 -73.2096 -19.07821

|

_cons | 209.566 32.77154 6.39 0.000 144.9359 274.1961

-------------+--

/sigma | 65.67672 3.481272 58.81116 72.54228

--

Obs. summary: 0 left-censored observations

183 uncensored observations

17 right-censored observations at apt>=800

We can test for an overall effect of prog using the test command. Below we see that the overall effect of prog is statistically significant.

test 2.prog 3.prog

(1) 2.prog = 0
(2) 3.prog = 0

F(2, 196) = 5.98
Prob > F = 0.0030

We can also test additional hypotheses about the differences in the
coefficients for different levels of prog. Below we
test that the coefficient for prog=2 is equal to the coefficient for
prog=3. In the output below we see that the coefficient for prog=2 is
significantly different than the coefficient for prog=3.

test 2.prog = 3.prog

(1) 2.prog - 3.prog = 0

F(1, 196) = 6.66
Prob > F = 0.0106

We may also wish to see measures of how well our

model fits. This can be
particularly useful when comparing competing models. One method of doing this is
to compare the predicted values based on the tobit model to the observed values in
the dataset. Below we use predict to generate predicted values of apt
based on the model. Next we correlate the observed values of apt with the
predicted values (yhat).

```
predict yhat
(option xb assumed; fitted values)

correlate apt yhat
(obs=200)

| apt yhat
-------------+------------------
apt | 1.0000
yhat | 0.7825 1.0000
```

The correlation between the predicted and observed values of apt is 0.7825. If we square this
value, we get the multiple squared correlation, this

indicates predicted values
share about 61% (0.7825^2 = 0.6123) of their variance with apt. Additionally,
we can use the user-written command
fitstat to produce a variety of fit statistics. You can find more information
on fitstat by typing search fitstat (see
How can I use the
search command to search for programs and get additional help? for more
information about using search).

fitstat

Measures of Fit for tobit of apt

Log-Lik Intercept Only: -1135.545 Log-Lik Full Model: -1041.063
D(193): 2082.126 LR(4): 188.965
Prob > LR: 0.000
McFadden's R2: 0.083 McFadden's Adj R2: 0.077
ML (Cox-Snell) R2: 0.611 Cragg-Uhler(Nagelkerke) R2: 0.611
McKelvey & Zavoina's R2: 0.616
Variance of y*: 11230.171 Variance of error: 4313.432

AIC: 10.481 AIC*n: 2096.126

BIC: 1059.550 BIC': -167.772

BIC used by Stata: 2113.916 AIC used by Stata: 2094.126

See Also

References

McDonald, J. F. and Moffitt, R. A. 1980. The Uses of Tobit Analysis. The Review of Economics and Statistics Vol 62(2): 318-321.