What is the concept of Truncated Regression and how is it applied in Stata Data Analysis?

Authored by

stats writer

June 29, 2024

Truncated regression is a statistical technique used to analyze data with limited or restricted values. It is applied in Stata data analysis to handle cases where the dependent variable is censored or truncated, meaning it has values that are only observed within a certain range. This technique allows for the estimation of regression parameters by taking into account the truncated nature of the data. It is commonly used in various fields such as economics, finance, and social sciences to better understand relationships between variables and make more accurate predictions. In Stata, truncated regression can be performed using the "truncreg" command, which takes into account the truncation point and produces results that account for the truncated nature of the data. Overall, the concept of truncated regression is a valuable tool in data analysis for handling complex data sets with restricted values.

Truncated Regression | Stata Data Analysis Examples

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

Truncated regression is used to model dependent variables for which some of the observations are not included in the analysis because of the value of the dependent variable.

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 or
potential follow-up analyses.

Examples of truncated regression

Example 1.

A study of students in a special GATE (gifted and talented education) program
wishes to model achievement as a function of language skills and the type of
program in which the student is currently enrolled. A major concern is
that students are required to have a minimum achievement score of 40 to enter
the special program. Thus, the sample is truncated at an achievement score
of 40.

Example 2. A researcher has data for a sample of Americans whose income is
above the poverty line. Hence, the lower part of the distribution of
income is truncated. If the researcher had a sample of Americans whose

income was at or below the poverty line, then the upper part of the income
distribution would be truncated. In other words, truncation is a result of
sampling only part of the distribution of the outcome variable.

Description of the data

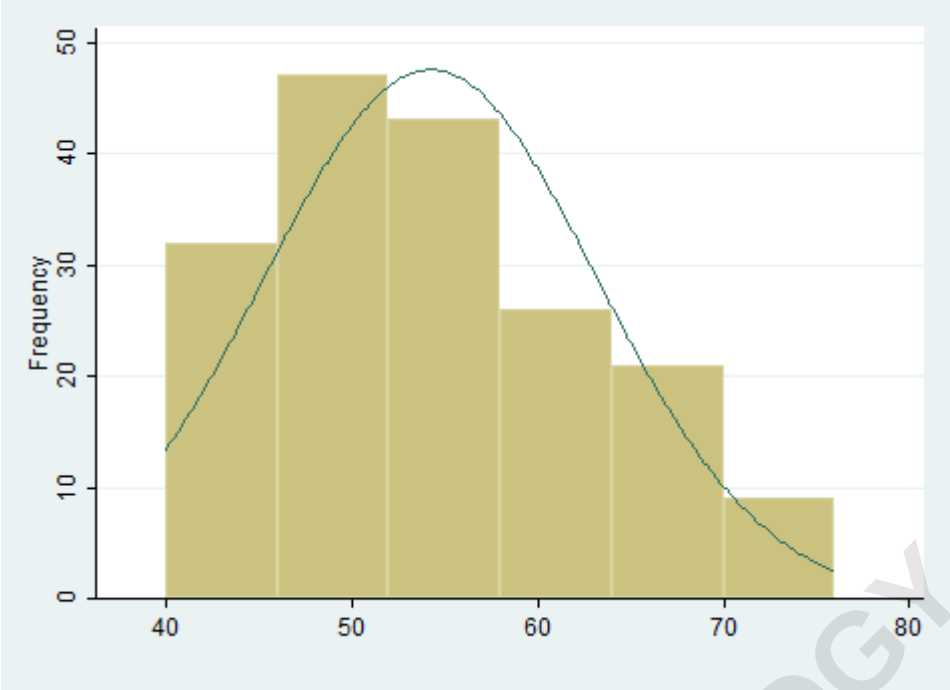
Let's pursue Example 1 from above.

We have a hypothetical data file, truncreg.dta, with 178 observations. The
outcome variable is called achiv, and the language test score
variable is called langscore. The variable prog is a categorical predictor variable with
three levels indicating the type of program in which the students were
enrolled.

Let's look at the data. It is always a good idea to start with descriptive
statistics.

use https://stats.idre.ucla.edu/stat/stata/dae/truncreg,

```
clear

summarize achiv langscore

Variable | Obs Mean Std. Dev. Min Max
-------------+--------------------------------------------------------
achiv | 178 54.23596 8.96323 41 76
langscore | 178 54.01124 8.944896 31 67

tabstat achiv, by(prog) stats(n mean sd)

Summary for variables: achiv
by categories of: prog (type of program)

prog | N mean sd
---------+----------------------------
general | 40 51.575 7.97074
academic | 101 56.89109 9.018759
vocation | 37 49.86486 7.276912
---------+----------------------------
Total | 178 54.23596 8.96323
----------------------------------------

histogram achiv, bin(6) start(40) freq normal
```



tabulate prog

type of |
program | Freq. Percent Cum.
----------+----------------------------------
general | 40 22.47 22.47
academic | 101 56.74 79.21
vocation | 37 20.79 100.00
----------+----------------------------------
Total | 178 100.00

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.

Truncated regression

Below we use the truncreg command to estimate a truncated regression
model. The i. before prog indicates that it is a factor variable (i.e., categorical variable), and that it should be included in the
model as a series of indicator variables. The ll() option in the truncreg command indicates the value at which the left truncation
take place. There is also a ul() option to indicate the value of
the right truncation, which was not needed in this example.

```
truncreg achiv langscore i.prog, ll(40)

(note: 0 obs. truncated)

Fitting full model:

Iteration 0: log likelihood = -598.11669
Iteration 1: log likelihood = -591.68374
```

```
Iteration 2: log likelihood = -591.31208
Iteration 3: log likelihood = -591.30981
Iteration 4: log likelihood = -591.30981

Truncated regression
Limit: lower = 40 Number of obs = 178
upper = +inf Wald chi2(3) = 54.76
Log likelihood = -591.30981 Prob > chi2 = 0.0000

------------------------------------------------------------------------------
achiv | Coef. Std. Err. z P>|z|
-------------+----------------------------------------------------------------
langscore | .7125775 .1144719 6.22 0.000 .4882168
.9369383
|
prog |
2 | 4.065219 2.054938 1.98 0.048 .0376131 8.092824
3 | -1.135863 2.669961 -0.43 0.671 -6.368891 4.097165
|
_cons | 11.30152 6.772731 1.67 0.095 -1.97279 24.57583
-------------+----------------------------------------------------------------
/sigma | 8.755315 .666803 13.13 0.000 7.448405 10.06222
------------------------------------------------------------------------------
```

test 2.prog 3.prog

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

chi2(2) = 7.19
Prob > chi2 = 0.0274

The two degree-of-freedom chi-square test indicates that prog is a
statistically significant predictor of achiv.

We can use the margins
command to obtain the expected cell means. Note that these are different
from the means we obtained with the tabstat command above.

margins prog

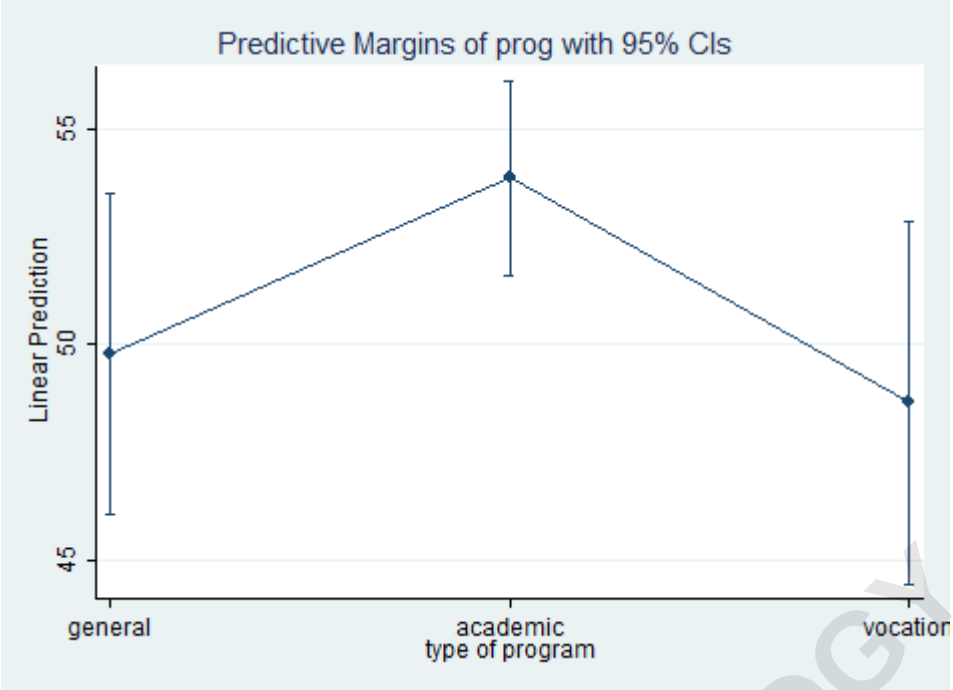
Predictive margins Number of obs = 178
Model VCE : OIM

Expression : Linear prediction, predict()

| Delta-method
| Margin Std. Err. z P>|z|
----------+----------
prog |
1 | 49.78871 1.897166 26.24 0.000 46.07034 53.50709
2 | 53.85393 1.150041 46.83 0.000 51.59989 56.10797
3 | 48.65285 2.140489 22.73 0.000 44.45757 52.84813

In the table above, we can see that the expected mean of avchiv for
the first level of prog is approximately 49.79; the expected mean for
level 2 of prog is 53.85; the expected mean for the third level of
prog is 48.65.

marginsplot



If you would like to compare truncated regression models, you can issue the estat ic command to get the log likelihood, AIC and BIC values.

estat ic

Model | Obs ll(null) ll(model) df AIC BIC

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

. | 178 . -591.3098 5 1192.62 1208.529

Note: N=Obs used in calculating BIC; see BIC note

The truncreg output includes neither an R2 nor a pseudo-R2. You can compute
a rough estimate of the degree of association by correlating achiv with the predicted
value and squaring the result.

```
predict p
correlate p achiv
(obs=178)

| p achiv
-------------+------------------
p | 1.0000
achiv | 0.5524 1.0000

display r(rho)^2
.30519203
```

The calculated value of .31 is rough estimate of the R2 you would find in an OLS
regression. The squared correlation between the observed and predicted
academic aptitude values is about 0.31, indicating that these predictors
accounted for over 30% of the variability in the outcome

variable.

Things to consider

See also

References