What is the procedure for conducting a Tobit regression analysis and what can be learned from the SAS annotated output?

Authored by

stats writer

June 30, 2024

A Tobit regression analysis is a statistical method used to analyze data with censored or limited dependent variables. This procedure is commonly used in economics, social sciences, and other fields to understand the relationship between these variables and their potential predictors.

The procedure for conducting a Tobit regression analysis involves the following steps:

1. Data Preparation: The first step is to collect and organize the data that will be used in the analysis. This includes the dependent variable, independent variables, and any other relevant variables.

2. Model Specification: The next step is to specify the Tobit regression model, which involves selecting the appropriate distribution for the dependent variable and determining the functional form and variables to be included in the model.

3. Estimation: Once the model is specified, the parameters are estimated using maximum likelihood estimation techniques. This involves finding the values of the model coefficients that best fit the data.

4. Interpretation of Results: The output from the Tobit regression analysis will include estimates of the model coefficients, their significance, and measures of goodness of fit. These results can be used to understand the impact of the independent variables on the dependent variable.

The SAS annotated output is a helpful tool for interpreting the results of the Tobit regression analysis. It provides a detailed explanation of the output, including the estimated coefficients, their standard errors, significance levels, and other relevant statistics. This allows researchers to understand the significance and direction of the relationship between the variables, as well as the overall fit of the model.

In summary, the procedure for conducting a Tobit regression analysis involves data preparation, model specification, estimation, and interpretation of results. The SAS annotated output can provide valuable insights into the relationships between variables and help researchers draw meaningful conclusions from the analysis.

Tobit Regression | SAS Annotated Output

This page shows an example of tobit regression analysis in SAS with footnotes explaining the output. The data in this example were

gathered on undergraduates
applying to graduate school and include undergraduate GPAs, the reputation of
the school of the undergraduate (a topnotch indicator), the students' GRE score,
and whether or not the student was admitted to graduate school.

The range of possible GRE scores is 200 to 800. This means that our outcome
variable is both left censored and right-censored. In other words, if two
students score 800, they are equal according to our scale but might not truly
be equal in aptitude. (In other words, we have a ceiling effect.) The same is
true of two students scoring 200 (a floor effect). Tobit regression generates a
model that predicts a censored outcome variable.

If we are interested in predicting a student's GRE score using their
undergraduate GPA and the reputation of their undergraduate institution, we

should first consider GRE as an outcome variable. We can use the dataset https://stats.idre.ucla.edu/wp-content/uploads/2016/02/tobit-1.sas7bdat .

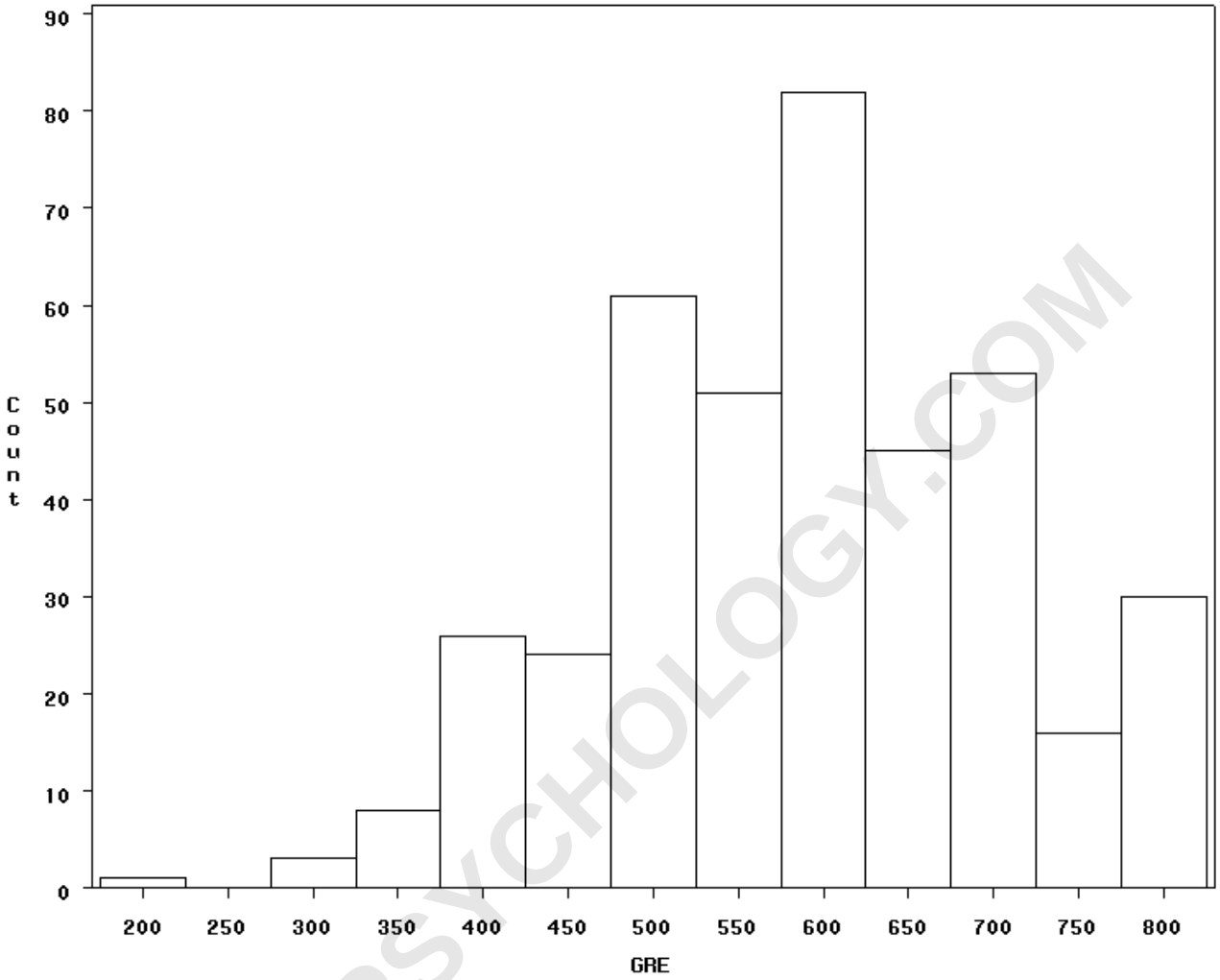
```
data tobit;
set "C:temptobit";
run;

proc means data = tobit;
var gre;
run;
```

The MEANS Procedure

Analysis Variable : GRE

N	Mean	Std Dev	Minimum	Maximum
400	587.7000000	115.5165364	220.0000000	800.0000000

```
proc univariate data = tobit noprint;
histogram gre / vscale = count;
run;
```



To generate a tobit model in SAS, we must first add a variable to our data indicating which observations are censored.

```
data tobit; set tobit;
censor = (gre >= 800 or gre <= 200);
```

```
run;

proc freq data = tobit;
table censor;
run;
```

The FREQ Procedure

censor	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	375	93.75	375	93.75
1	25	6.25	400	100.00

In the output above, we see that 25 of our observations are censored while
375 are not. Next, we program our model in SAS. This can be done with proc
lifereg. We specify our model, indicating that our response is censored.

```
proc lifereg data = tobit;
model gre*censor(1) = gpa topnotch /d=normal;
run;
```

The LIFEREG Procedure

Model Information

Data Set WORK.TOBIT
Dependent Variable GRE
Censoring Variable censor
Censoring Value(s) 1
Number of Observations 400
Noncensored Values 375
Right Censored Values 25
Left Censored Values 0
Interval Censored Values 0
Name of Distribution Normal
Log Likelihood -2331.431433

Number of Observations Read 400
Number of Observations Used 400

Algorithm converged.

Type III Analysis of Effects

Wald
Effect DF Chi-Square Pr > ChiSq

GPA 1 53.6486 <.0001

TOPNOTCH 1 8.7718 0.0031

Analysis of Parameter Estimates

Standard 95% Confidence Chi-

Parameter DF Estimate Error Limits Square Pr > ChiSq

Intercept 1 205.8515 51.2408 105.4213 306.2818 16.14 <.0001

GPA 1 111.3085 15.1967 81.5235 141.0934 53.65 <.0001

TOPNOTCH 1 46.6577 15.7536 15.7813 77.5342 8.77 0.0031

Scale 1 111.4882 4.1438 103.6554 119.9129

Tobit Regression Output

The LIFEREG Procedure

Model Information

Data Seta WORK.TOBIT

Dependent Variableb GRE

Censoring Variablec censor

Censoring Value(s)d 1

Number of Observationse 400

Noncensored Valuesf 375

Right Censored Valuesg 25

Left Censored Valuesh 0

Interval Censored Valuesi 0

Name of Distributionj Normal

Log Likelihoodk -2331.431433

Number of Observations Read 400

Number of Observations Used 400

Algorithm converged.l

Type III Analysis of Effectsm

Wald

Effect DF Chi-Square Pr > ChiSq

GPA 1 53.6486 <.0001

TOPNOTCH 1 8.7718 0.0031

Analysis of Parameter Estimates

Standard 95% Confidence Chi-

ParameternDFoEstimatep Errorq Limitsr Squares Pr>ChiSqt

Intercept 1 205.8515 51.2408 105.4213 306.2818 16.14

<.0001

GPA 1 111.3085 15.1967 81.5235 141.0934 53.65 <.0001

TOPNOTCH 1 46.6577 15.7536 15.7813 77.5342 8.77 0.0031

Scaleu 1 111.4882 4.1438 103.6554 119.9129

a. Dataset - This indicates the dataset used in the analysis. If a
dataset is not specified in the model command, SAS uses the most recently
created/modified dataset by default.

b. Dependent Variable - This is the response variable predicted by the
model. We are using a tobit model because this response variable is censored: the GRE scores are scaled from 200 to 800 and cannot be measured outside of this range
(although the phenomenon underlying the scores, in this case aptitude, is not
bounded).

c. Censoring Variable - This is the variable we defined in preparation

for running the tobit model. Values in our dataset will be considered censored
based on the corresponding value of our censoring variable.

d. Censoring Value(s) - These are the values of the censoring variable
that indicate a censored value in the dependent variable. In this example, the
observations where censor = 1 are censored. We indicated this in
the proc lifereg command with the 1 in parenthesis after censor.

e. Number of Observations - This is the number of observations from
the dataset used in the model. If an observation is missing data in the outcome or
any of the predictor variables, then it is excluded from the analysis.

f. Noncensored Values - This is the number of observations in the
model that were not censored. In this example, there were 375 observations in

the dataset with 200 < gre < 800.

g. Right Censored Values - This is the number of observations in the
model that were right censored. In this example, there were 25 observations in
the dataset with gre >= 800.

h. Left Censored Values - This is the number of observations in the
model that were left censored. In this example, there were zero observations in
the dataset with gre <= 200.

i. Interval Censored Values - This is the number of observations in
the model that were interval censored (where the outcome variable fell in an
interval that was censored). This type of censoring was not used in this model,
and so there were zero observations in the dataset in this category.

j.
Name of Distribution

- This indicates the distribution assumed for the errors terms of the model.
In a tobit model, this distribution is normal.

k.
Log Likelihood
- This is the log likelihood of the fitted model. It is used in the Likelihood
Ratio Chi-Square test of whether all predictors' regression coefficients in the
model are simultaneously zero.

l.
Algorithm Converged.
- This indicates that the SAS convergence criterion for the iterating steps used
in maximizing the likelihood was met. The default criterion in SAS is the
relative gradient convergence criterion, and the default precision is 10-8.

m.
Type III Analysis of Effects
- This is an analysis of the model variables using Type III sum of squares.

**That is, the effect of a model variable does not depend on the order in which
the variable is specified in the model.**

**n.
Parameter
- This lists the model parameters. Our model includes an intercept
and the specified predictor variables.**

**o.
DF
- These are the degrees of freedom associated with each of the model parameters.**

**p. Estimate - These are the regression coefficients. These
coefficients are interpreted as you would interpret coefficients from an OLS regression: the
expected GRE score changes by Estimate for each unit increase in the
corresponding predictor.**

Intercept- If all of the predictor variables in the model are

evaluated at zero, the predicted GRE score would be Intercept = 205.8515. For subjects from non-topnotch undergraduate institutions (topnotch evaluated at zero) with zero gpa, the predicted GRE score would be 205.8515. This may seem very low, considering the mean GRE score is 587.7, but note that evaluating gpa at zero is out of the range of plausible values for gpa.

gpa - If a subject were to increase his gpa by one point, his expected GRE score would increase by 111.3085 points while holding all other variables in the model constant. Thus, the higher a student's gpa, the higher the predicted GRE score.

topnotch - If a subject attended a topnotch institution for her undergraduate education, her expected GRE score would be 46.65774 points higher than a subject with the same grade point average

**who attended a
non-topnotch institution. Thus, subjects from topnotch undergraduate
institutions have higher predicted GRE scores than subjects from
non-topnotch undergraduate institutions if grade point averages are held
constant.**

**q. Standard Error - These are the standard errors of the individual
regression coefficients.**

**r. 95% Confidence Limits - This is the Confidence Interval (CI) for an
individual coefficient given that the other predictors are in the model.
For a given predictor with a level of 95% confidence, we'd say that we are 95%
confident that the "true" coefficient lies between the lower and upper limit of
the interval. The CI is equivalent to the t test statistic: if the CI
includes zero, we'd fail to reject the null hypothesis that**

a particular
regression coefficient is zero given the other predictors are in the model with
alpha level of zero. An advantage of a CI is that it is illustrative; it
provides a range where the "true" parameter may lie.

s. Chi-Square - This is the Chi-Square test statistic corresponding to the hypothesis that the given parameter's estimate is equal to zero.

t. Pr>ChiSq - This is the probability the Chi-Square test statistic (or a
more extreme test statistic) would be observed under the null hypothesis that a
particular predictor's regression coefficient is zero, given that the rest of
the predictors are in the model. For a given alpha level, P>ChiSq
determines whether or not the null hypothesis can be rejected. If P>ChiSq
is less than alpha, then the null hypothesis can be rejected and the parameter
estimate is considered statistically significant at that

alpha level.

Intercept - The Chi-Square test statistic for the intercept,
Intercept is 16.14 with an associated p-value of < 0.001. If we set our alpha level at 0.05, we would reject the null
hypothesis and conclude that the Intercept has been found to be statistically
different from zero given gpa and topnotch are in the model
and evaluated at zero.

gpa - The Chi-Square test statistic for the predictor gpa is 53.65 with an associated p-value of <0.001. If we
set our alpha level to 0.05, we would reject the null hypothesis and
conclude that the regression coefficient for gpa has been found to be
statistically different from zero given topnotch is in the model.

topnotch -The Chi-Square test statistic for the predictor topnotch is 8.77 with an associated p-value of
0.0031. If we set our alpha level to 0.05, we would reject

the null
hypothesis and conclude that the regression coefficient for topnotch
has been found to be statistically different from zero given gpa is
in the model.

u. Scale - This is the estimated standard error of the regression.
This value, 111.4882, is comparable to the root mean squared error that would be
obtained in an OLS regression.