What is the SAS annotated output for truncated regression?

Authored by

stats writer

June 30, 2024

The SAS annotated output for truncated regression is a statistical analysis method that allows for the examination of relationships between variables when the dependent variable is censored or truncated, meaning that it is limited in its range of values. The output provides a detailed summary of the regression model, including the estimated coefficients, standard errors, and p-values for each variable, as well as the model's goodness of fit measures. It also includes a diagnostic plot to assess the model's assumptions and any influential data points. The annotated output allows for a thorough understanding and interpretation of the results, aiding in the decision-making process for the given dataset.

Truncated Regression | SAS Annotated Output

This page shows an example of truncated regression analysis in SAS with footnotes explaining the output. A truncated regression model predicts an outcome variable restricted to a truncated sample of its distribution. For example, if we wish to predict the age of licensed motorists from driving habits, our outcome variable is truncated at 16 (the legal driving age in the U.S.). While the population of ages extends below 16, our sample of the population does not. It is important to note the difference between truncated and censored data. In the case of censored data, there are limitations to the measurement scale that prevent us

from knowing the true value of the dependent variable despite having some
measurement of it. Consider the speedometer in a car. The speedometer may
measure speeds up to 120 miles per hour, but all speeds equal to or greater than
120 mph will be read as 120 mph. Thus, if the speedometer measures the speed to
be 120 mph, the car could be traveling 120 mph or any greater speed-we have no
way of knowing. Censored data suggest limits on the measurement scale of the
outcome variable, while truncated data suggest limits on the outcome variable in
the sample of interest.

In this example, we will look at data from a study of students in a special GATE (gifted
and talented education) program,
https://stats.idre.ucla.edu/wp-content/uploads/2016/02/truncated.sas7bdat. We wish to model achievement (achiv) as
a function of gender, language skills and math skills (female, langscore and

mathscore in the dataset). A major concern is that students require a minimum achievement score of 40 to enter the special program.
Thus, the sample is truncated at an achievement score of 40.

First, we will examine the data. We are interested in checking the range of values of our
outcome variable, so
we will include a histogram of achiv. For our other variables, we simply
want a general sense of the values. For this, we can look at the summary
statistics from proc means and a frequency of the categorical variable
female.

```
data truncated;
set "D:datatrunctated";
run;

proc means data = truncated;
run;
```

The MEANS Procedure

Variable N Mean Std Dev Minimum Maximum

ID 178 103.6235955 57.0895709 3.0000000 200.0000000

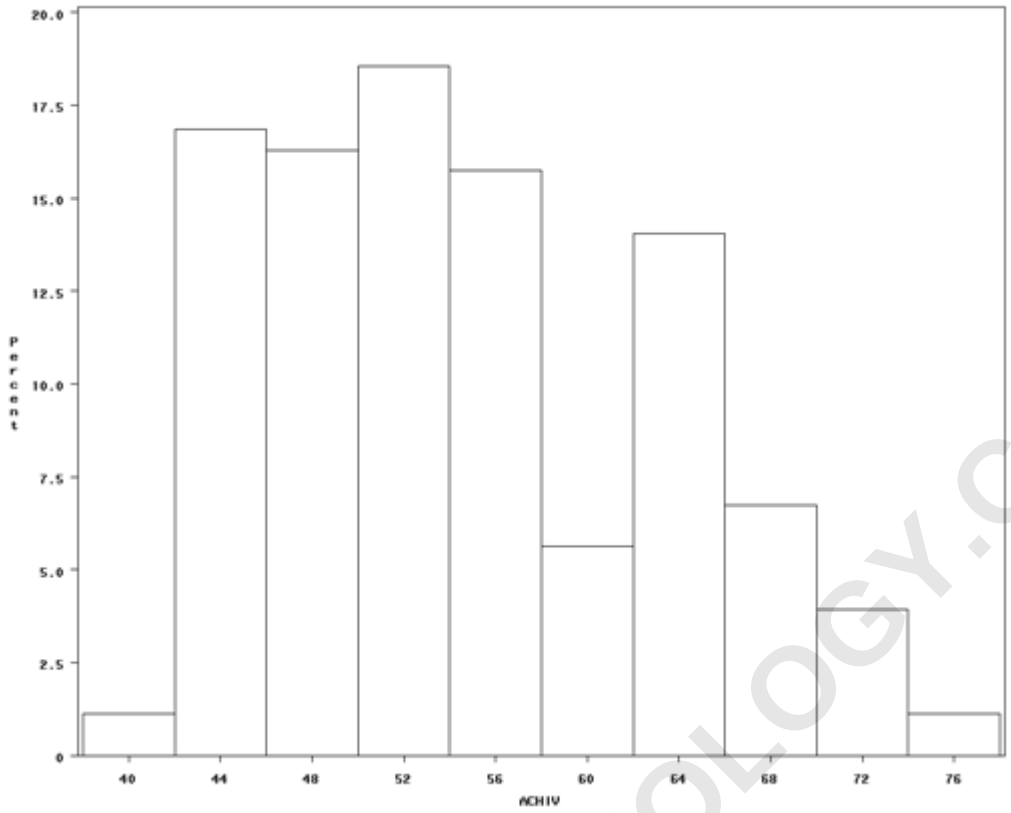
ACHIV 178 54.2359551 8.9632299 41.0000000 76.0000000

FEMALE 178 0.5505618 0.4988401 0 1.0000000

LANGSCORE 178 5.4011236 0.8944896 3.0999999 6.6999998

MATHSCORE 178 5.3028090 0.9483515 3.0999999 7.4000001

```
proc univariate data = truncated;
var achiv;
histogram achiv;
run;
```



proc freq data = truncated;

table female;

run;

The FREQ Procedure

Cumulative Cumulative

FEMALE Frequency Percent Frequency Percent

----

0 80 44.94 80 44.94

1 98 55.06 178 100.00

Now, we can generate a truncated regression model in SAS
using proc qlim. We first indicate the outcome and predictors in the
model statement. We then indicate in the endogenous statement that our outcome variable,
achiv, is truncated with a lower bound of 40. If our data also had
an upper bound, we would include it in this line as well.

```
proc qlim data = truncated;
model achiv = female langscore mathscore;
endogenous achiv ~ truncated(lb=40);
run;
```

The QLIM Procedure

Summary Statistics of Continuous Responses

Variable	Mean	Standard Error	Type	N Obs Lower Bound	N Obs Upper Bound	Lower Bound	Upper Bound
achiv	54.23596	8.963230	Truncated			40	

Model Fit Summary
Number of Endogenous Variables 1

Endogenous Variable achiv

Number of Observations 178

Log Likelihood -574.53056

Maximum Absolute Gradient 2.72145E-6

Number of Iterations 12

AIC 1159

Schwarz Criterion 1175

Algorithm converged.

Parameter Estimates

Standard Approx

Parameter Estimate Error t Value Pr > |t|

Intercept -0.293996 6.204858 -0.05 0.9622

FEMALE -2.290930 1.490333 -1.54 0.1242

LANGSCORE 5.064697 1.037769 4.88 <.0001

MATHSCORE 5.004053 0.955571 5.24 <.0001

_Sigma 7.739052 0.547644 14.13 <.0001

Truncated Regression Output

The QLIM Procedure

Summary Statistics of Continuous Responses

N Obs N Obs

Standard Lower Upper Lower Upper
Variablea Meanb Errorc Typed Bounde Boundf Boundg Boundh
achiv 54.23596 8.963230 Truncated 40

Model Fit Summary
Number of Endogenous Variables 1
Endogenous Variable achiv
Number of Observations 178
Log Likelihoodi -574.53056
Maximum Absolute Gradientj 2.72145E-6
Number of Iterationsk 12
AICl 1159
Schwarz Criterionm 1175

Algorithm converged.

Parameter Estimates
Standard Approx
Parameter Estimaten Erroro t Valuep Pr > |t|q
Intercept -0.293996 6.204858 -0.05 0.9622
FEMALE -2.290930 1.490333 -1.54 0.1242
LANGSCORE 5.064697 1.037769 4.88 <.0001
MATHSCORE 5.004053 0.955571 5.24 <.0001

_Sigmar 7.739052 0.547644 14.13 <.0001

a. Variable - This is the outcome variable predicted in the
regression. In this example, achiv is the truncated outcome variable.

b. Mean - This is the mean of the outcome variable. In this
example, the mean of achiv is 54.23596.

c. Standard Error - This is the standard error of our outcome
variable. It is equal to 8.9632299, the standard deviation we saw in the proc
means output earlier.

d. Type - This describes the type of endogenous variable being
modeled. Procqlim allows for both truncated and censored
outcome variables. In this example, our outcome is truncated.

e. Lower Bound - This indicates the lower limit specified

for the
outcome variable. In this example, the lower limit is 40.

f. Upper Bound - This indicates the upper limit specified for
the outcome variable. In this example, we did not specify an upper limit.

g. N Obs Lower Bound - This indicates how many observations in the
model had outcome variable values below the lower limit indicated in the
function call. In this example, it is the number of observations where achiv < 40. The minimum value of achiv listed
in the data summary was 41, so there were zero observations truncated from below.

h. N Obs Upper Bound - This indicates how many observations in the
model had outcome variable values above the upper limit indicated on the
endogenous statement. In this example, we did not specify an upper limit, so there were

zero observations truncated from above.

i. 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.

j.
Maximum Absolute Gradient
- This is the absolute value of the gradient seen in the last iteration.
The default convergence criterion used by proc qlim is an absolute gradient of 0.00001.
Thus, when the absolute gradient falls below 0.00001, the model has
converged. This value is the first absolute gradient less than 0.00001. If you
wish to see additional output regarding the iteration history, add the itprint
option
to the proc qlim statement.

k. Number of Iterations - This is the number of iterations

required by
SAS for the model to converge. Truncated regression uses maximum
likelihood estimation, which is an iterative procedure. The first
iteration is the "null" or "empty" model; that is, a model with no predictors.
At the next iteration, the specified predictors are included in the model. In
this example, the predictors are female, langscore and mathscore.At each iteration, the log likelihood increases because the goal is to
maximize the log likelihood. When the difference between successive iterations
is very small, the model is said to have "converged" and the iterating stops. For more information on this process, see

Regression Models for Categorical and Limited Dependent Variables by J.
Scott Long (page 52-61).

I.
AIC

- This is the Akaike Information Criterion. It is a measure of model fit that is calculated as AIC = -2 Log L +
2p, where p is the number of parameters estimated in the model. In this
example, p=5; three predictors, one intercept, and _Sigma (see
superscript r).
AIC
is used for the comparison of models from different samples or non-nested models. Ultimately, the model with the smallest
AIC
is considered the best.

m.
Schwarz Criterion
- This is the Schwarz Criterion. It is defined as - 2 Log L + p*log(Σ
fi),
where fi's
are the frequency values of the
ith
observation, and p was defined previously. Like

AIC,
SC
penalizes for the number of predictors in the model and the smallest
SC
is most desirable.

n. Estimate - These are the estimated regression coefficients.
They are interpreted in the same manner as OLS regression coefficients: for a one unit
increase in the predictor variable, the expected value of the outcome variable
changes by the regression coefficient, given the other predictor variables in
the model are held constant.

Intercept - Sometimes called the constant, this is the regression estimate when all
predictor variables in the
model are evaluated at zero. For a male student (the variable
female
evaluated at zero) with langscore and

mathscore of zero, the
predicted achievement score is -0.293996. Note that evaluating
langscore
and mathscore at zero is out of the range of plausible test scores.

female - The expected achievement score for a female student is
2.290930 units lower than the expected achievement score for a male student
while holding all other variables in the model constant. In other words, if two
students, one female and one male, had identical language and math scores, the
predicted achievement score of the male would be 2.290930 units higher than the
predicted achievement score of the female student.

langscore - This is the estimated regression estimate for a one
unit increase in langscore, given the other variables are held constant

in the model. If a student were to increase her langscore by one point, her predicted achievement score would increase by 5.064697 units, while holding the other variables in the model constant. Thus, the students with higher language scores will have higher predicted achievement scores than students with lower language scores, holding the other variables constant.

mathscore - This is the estimated regression estimate for a one unit increase in mathscore, given the other variables are held constant in the model. If a student were to increase her mathscore by one point, her predicted achievement score would increase by 5.004053 units, while holding the other variables in the model constant. Thus, the students with higher math scores will have higher predicted achievement scores than students with lower

math scores, holding the other variables constant.

o. Standard Error - These are the standard errors of the individual
regression coefficients. They are used in the calculation of the
t
test statistic, superscript p.

p. t Value - The test statistic t is the ratio of the
Coef.
to the Std. Err. of the respective predictor. The t value follows a
t-distribution which is used to test against a two-sided alternative hypothesis that the Estimate is not equal to zero.

q. Approx Pr > |t| - This is the probability the t 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>|t|

determines whether or not the null hypothesis can be rejected. If
P>|t|
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
t test statistic for Intercept,
is (-0.293996/6.204858) = -0.05 with an associated p-value of 0.9622. If we set
our alpha level at 0.05, we would fail to reject the null hypothesis and
conclude that Intercept has not been found to be statistically different from
zero given female,
langscore
and
mathscore are in the model
and evaluated at zero.

female - The
t test statistic for the predictor

female
is (-2.290930/1.490333) = -1.54 with an associated p-value of 0.1242. If we set
our alpha level to 0.05, we would fail to reject the null hypothesis and
conclude that the regression coefficient for
female has not been found to
be statistically different from zero given
langscore and
mathscore
are in the model.

langscore - The
t test statistic for the predictor

langscore is (5.064697/1.037769) = 4.88 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
langscore has been found
to be statistically different from zero given
female and

mathscore
are in the model.

mathscore - The
t test statistic for the predictor

mathscore is (5.004053/0.955571) = 5.24 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
mathscore has been found
to be statistically different from zero given
female and
langscore
are in the model.

r. _Sigma - This is the estimated standard error of the regression. In
this example, the value, 7.739052, is comparable to the root mean squared error
that would be obtained in an OLS regression. If we ran an OLS regression
with the same outcome and predictors, our RMSE

would be 6.8549. This is indicative of how much the outcome varies from the predicted value.

_Sigma approximates this quantity for truncated regression.