How can I conduct an interval regression analysis in Stata and interpret the results using annotated output?

Authored by

stats writer

June 30, 2024

Interval regression analysis is a statistical method used to analyze data with censored or truncated values. In Stata, conducting an interval regression analysis involves using the "intreg" command, which requires specifying the dependent variable, independent variables, and censoring variables. Once the analysis is run, the output will include various statistics such as coefficient estimates, standard errors, and p-values. To interpret the results, one can use annotated output, which provides a detailed explanation of each statistic and its significance. This allows for a better understanding of the relationship between the variables and can aid in making informed decisions based on the analysis. Overall, conducting an interval regression analysis in Stata and utilizing annotated output can provide valuable insights and aid in making data-driven decisions.

Interval Regression | Stata Annotated Output

**This page shows an example of interval regression analysis with footnotes
explaining the output in Stata. Suppose you are interested in predicting an outcome for
which the exact values are unobserved, but an interval containing the exact
value is observed. For instance, you may wish to predict income with education
and gender, but you can only observe income brackets. You could consider the
income brackets to be an ordered categorical outcome and model income
using an ordered logistic model. However, an ordered logistic model would
predict the likelihood that a person falls into a given bracket, but not the**

person's income. Interval regression models predict the value of outcome variable.
Thus, you could predict income using education and gender despite not observing
exact income values.

In this example, we will look at a dataset in which we wish to predict GPA from
teacher ratings of students' effort and from reading and writing test scores. The measure
of GPA is a self-report response to the following item:

Select the category that best represents your overall gpa.
0.0 to 2.0
2.0 to 2.5
2.5 to 3.0
3.0 to 3.4
3.4 to 3.8
3.8 to 4.0

Note that the intervals listed above are not of equal size (the lowest GPA
interval spans 2 points, while the highest GPA interval

spans 0.2 point).
This is not problematic for interval regression. The intervals appearing in the
data can also overlap, which we might see if we combined this dataset with
another where the GPA intervals were slightly different.

Our outcome variable will be GPA. We do not know exact GPAs, but we do know
the interval in which the GPA falls. Let us first examine our dataset. The
interval containing our outcome variable's value is described using two
variables-the interval's lower bound (lgpa) and the interval's upper bound (ugpa).
In a sense, our outcome variable is split into two variables. Note that this is the format required for interval regression in Stata. If
your intervals are not defined by a lower bound variable and an upper bound
variable, you must reformat your data before proceeding.

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

```
clear

list in 1/10, clean

id lgpa ugpa write rating read
1. 1 2.5 3 175 54 150
2. 2 3.4 3.8 125 68 250
3. 3 2.5 3 70 48 150
4. 4 0 2 50 52 50
5. 5 3 3.4 70 49 250
6. 6 3.4 3.8 205 53.5 150
7. 7 3.8 4 180 72 250
8. 8 2 2.5 50 50 250
9. 9 3 3.4 155 57.5 150
10. 10 3.4 3.8 105 69 250

tab lgpa ugpa

| ugpa
lgpa | 2 2.5 3 3.4 3.8 4 | Total
-----------+----------------------------------------------------------
+----------
0 | 1 0 0 0 0 0 | 1
```

2 | 0 9 0 0 0 0 | 9
2.5 | 0 0 8 0 0 0 | 8
3 | 0 0 0 4 0 0 | 4
3.4 | 0 0 0 0 6 0 | 6
3.8 | 0 0 0 0 0 2 | 2
----------+----------
+----------
Total | 1 9 8 4 6 2 | 30

summarize write rating read

Variable | Obs Mean Std. Dev. Min Max
----------+----------
write | 30 113.8333 49.94278 50 205
rating | 30 57.53333 8.303441 48 72
read | 30 171.6667 94.39767 50 350

Now, we can generate our interval model. In Stata, we use the intreg
command, first specifying the lower bound interval variable, then the upper
bound interval variable, and then the predictors. In this example, we are predicting

GPA with three predictors: write, rating and read.

intreg lgpa ugpa write rating read

Fitting constant-only model:

Iteration 0: log likelihood = -52.129849
Iteration 1: log likelihood = -51.74803
Iteration 2: log likelihood = -51.747288
Iteration 3: log likelihood = -51.747288

Fitting full model:

Iteration 0: log likelihood = -38.212102
Iteration 1: log likelihood = -36.680551
Iteration 2: log likelihood = -36.662189
Iteration 3: log likelihood = -36.662185
Iteration 4: log likelihood = -36.662185

Interval regression Number of obs = 30
LR chi2(3) = 30.17
Log likelihood = -36.662185 Prob > chi2 = 0.0000

--
| Coef. Std. Err. z P>|z|
-------------+--

write | .0052829 .0015363 3.44 0.001 .0022718 .0082939
rating | .016789 .009751 1.72 0.085 -.0023226 .0359005
read | .002329 .0008046 2.89 0.004 .000752 .003906
_cons | .9133711 .4794007 1.91 0.057 -.026237 1.852979
-------------+--
/lnsigma | -1.090882 .1516747 -7.19 0.000 -1.388159 -.7936051
-------------+--
sigma | .3359201 .0509506 .2495343 .4522116

Observation summary: 0 left-censored observations
0 uncensored observations
0 right-censored observations
30 interval observations

Interval Regression Output

Fitting constant-only modela:

Iteration 0: log likelihood = -52.129849
Iteration 1: log likelihood = -51.74803
Iteration 2: log likelihood = -51.747288
Iteration 3: log likelihood = -51.747288

Fitting full modelb:

Iteration 0: log likelihood = -38.212102
Iteration 1: log likelihood = -36.680551
Iteration 2: log likelihood = -36.662189
Iteration 3: log likelihood = -36.662185
Iteration 4: log likelihood = -36.662185

Interval regression Number of obsd = 30
LR chi2(3)e = 30.17
Log likelihoodc = -36.662185 Prob > chi2f = 0.0000

--
| Coef.g Std. Err.h zi P>|z|j k
-------------+--
write | .0052829 .0015363 3.44 0.001 .0022718 .0082939
rating | .016789 .009751 1.72 0.085 -.0023226 .0359005
read | .002329 .0008046 2.89 0.004 .000752 .003906
_cons | .9133711 .4794007 1.91 0.057 -.026237 1.852979
-------------+--
/lnsigmal| -1.090882 .1516747 -7.19 0.000 -1.388159 -.7936051
-------------+--
sigmam| .3359201 .0509506 .2495343 .4522116

Observation summaryn: 0 left-censored observations
0 uncensored observations
0 right-censored observations
30 interval observations

a.

Fitting constant-only model

- This is the iteration history for fitting the constant only model. This model does not include any predictors and is simply estimating the mean predicted value of the outcome variable. Because the observed values for the outcome variable are intervals, not exact values, the mean predicted value is not simply the mean of the observed values. Instead, the predicted mean is arrived at iteratively by maximizing the log likelihood of the data given a mean predicted value.

b.

Fitting full model

- This is the iteration history for fitting the model including the specified predictors.

c.

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.

d.

Number of obs

- This is the number of observations in the dataset for which all of the predictor variables and at least one of the outcome interval variables is non-missing. In interval regression, one of the interval bounds may be missing.

If the upper bound of an interval is missing, then the interval is treated as . If both the lower bound and

upper bound are missing, then the observation is not included in the model.

e.

LR chi2(3)

- This is the Likelihood Ratio (LR) Chi-Square test that at least one of the
predictors' regression coefficient is not equal to zero. The number in the
parentheses indicates the degrees of freedom of the Chi-Square distribution used
to test the LR Chi-Square statistic and is defined by the number of predictors
in the model (3).

f.

Prob > chi2

- This is the probability of getting a LR test
statistic as extreme as, or more so, than the observed statistic under the null
hypothesis; the null hypothesis is that all of the regression coefficients
are simultaneously equal to zero. In other words, this is the

probability of obtaining this chi-square statistic (30.17) or one more extreme if there is in fact
no effect of the predictor variables. This p-value is compared to a specified
alpha level, our willingness to accept a type I error, which is typically set at
0.05 or 0.01. The small p-value from the LR test, <0.0001, would lead us to
conclude that at least one of the regression coefficients in the model is not
equal to zero. The parameter of the chi-square distribution used to test the
null hypothesis is defined by the degrees of freedom in the prior line,
chi2(3).

g.
Coef.
- These are the 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.

write - This is the estimated regression
estimate for a one unit increase in writing test score, given the other
variables are held constant in the model. If a student were to increase her
writing test score by one point, her predicted GPA would increase by 0.0052829
unit, while holding the other variables in the model constant. Thus, the
students with higher writing test scores will have higher predicted GPAs than
students with lower writing test scores, holding other variables constant.

rating - This is the estimated regression
estimate for a one unit increase in teachers' ratings of students' effort, given the other
variables are held constant in the model. If a student were to increase her
rating by one point, her predicted GPA would increase

by 0.016789 unit, while
holding the other variables in the model constant. Thus, the students with
higher effort ratings will have higher predicted GPAs than students with lower
effort ratings, holding other variables constant.

read - This is the estimated regression
estimate for a one unit increase in reading test score, given the other
variables are held constant in the model. If a student were to increase her
reading test score by one point, the predicted GPA would increase by 0.002329
unit, while holding the other variables in the model constant. Thus, the
students with higher reading test scores will have higher predicted GPAs than
students with lower reading test scores, holding other variables constant.

_cons - This is the regression estimate when
all variables in the model are evaluated at zero. For a student with a writing

test, reading test, and effort rating of zero, the predicted GPA is 0.9133711. Note that
evaluating write, read and rating at zero is out of the range of
plausible test scores and ratings.

h.
Std. Err.
- These are the standard errors of the individual regression coefficients. They
are used in both the calculation of the z test statistic, superscript
i, and the
confidence interval of the regression coefficient, superscript k.

i.
z
- The test statistic z is the ratio of the Coef. to the Std. Err. of the respective predictor. The z value follows a standard normal distribution which is used to test against a two-sided alternative hypothesis that the Coef. is not equal to zero.

j.

P>|z|

- This is the probability the z 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>|z| determines whether or not the null hypothesis

can be rejected. If P>|z|

is less than alpha, then the null hypothesis can be rejected and the parameter

estimate is considered statistically significant at that alpha level.

write - The z test

statistic for the predictor write is (0.0052829/0.0015363) = 3.44 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

write has

been

found to be statistically different from zero given rating

and read
are in the model.

rating - The z test
statistic for the predictor rating is (0.016789/0.009751) = 1.72 with an
associated p-value of 0.085. If we set our alpha level to 0.05, we would
fail to reject the null hypothesis and conclude that the regression coefficient for
rating has not been
found to be statistically different from zero given write and read
are in the model.

read - The z test
statistic for the predictor read is (0.002329/0.0008046) = 2.89 with an
associated p-value of 0.004. If we set our alpha level to 0.05, we would
reject the null hypothesis and conclude that the regression coefficient for
read has been
found to be statistically different from zero given write

**and rating
are in the model.**

**_cons - The z test
statistic for the intercept, _cons, is (0.9133711/0.4794007) = 1.91 with
an associated p-value of 0.057. If we set our alpha level at 0.05, we would
fail to reject the null hypothesis and conclude that _cons has not been found to be
statistically different from zero given write, rating and read are in the model and evaluated at zero.**

k.

**- 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. It is calculated as the Coef. (zα/2)*(Std.Err.),
where zα/2 is a critical value on the standard normal**

distribution.
The CI is equivalent to the z 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. An advantage of a CI is
that it is illustrative; it provides a range where the "true" parameter may
lie.

l.
/lnsigma
- This is the log of the estimated standard error. See superscript m.

m.
sigma
- This is the estimated standard error of the regression. This value, 0.3359201,
is comparable to the root mean squared error that would be obtained in an OLS
regression of the actual outcome values on the same set of predictors.

Generally, the smaller the intervals, the closer this value will be to the RMSE
of an OLS regression. This can be explored by looking at an OLS
regression, then creating intervals of different sizes around the outcome
variable and examining the results of interval regressions using the different
intervals.

n.
Observation summary
- This is a breakdown of how many of the observations were uncensored,
right-censored, left-censored, or both left- and right-censored.

Left-censored observations
are those observations where the lower bound of the interval is missing, and
therefore considered to be negative infinity.

Uncensored
observations are those observations where the lower bound of the interval is

equal to the upper bound of the interval.

Right-censored observations
are those observations where the upper bound of the interval is missing, and
therefore considered to be infinity.

Interval observations
are those observations where both the lower bound and upper bound are
non-missing and not equal.