# What is Interval Regression and how can it be used for Stata Data Analysis?

Authored by

**stats writer**

June 29, 2024

Interval regression is a statistical modeling technique used to analyze data in which the dependent variable is censored or truncated, meaning that its true value is not fully observed. This method is commonly used in Stata data analysis to handle data sets with censoring or truncation, such as survival data or limited dependent variables. It estimates the relationship between the dependent variable and a set of independent variables by taking into account the interval of possible values for the dependent variable. This allows for more accurate and meaningful interpretation of the results, as well as a more comprehensive understanding of the relationship between variables. Overall, interval regression is a valuable tool in Stata data analysis for handling complex data sets and obtaining reliable and informative results.

## Interval Regression | Stata Data Analysis Examples

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

## Interval regression is used to model outcomes that have interval censoring.
## In other words, you know the ordered category into which each observation falls,
## but you do not know the exact value of the observation. Interval
## regression is a generalization of censored regression.

## 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 interval regression

Example 1. We wish to model annual income using years of education and marital status. However, we do not have access to the precise values for income. Rather, we only have data on the income ranges: $100,000. Note
that the extreme values of the categories on either end of the range are either left-censored
or right-censored. The other categories are interval censored, that is, each interval is both
left- and right-censored. Analyses of this type require a generalization of censored regression
known as interval regression.

Example 2. We wish to predict GPA from teacher ratings of 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.**

**less than 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 3.9**

**4.0 or greater**

**Again, we have a situation with both interval censoring and left- and right-censoring. We do not know the exact value of GPA for each student; we only know the interval in which their GPA falls.**

**Example 3. We wish to predict GPA from teacher ratings of effort, writing test scores and the type of program in which the student was enrolled (vocational, general or academic). 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**

**This is a slight variation of Example 2. In this example, there is only interval censoring.**

### Description of the data

**Let's pursue Example 3 from above.**

**We have a hypothetical data file, intreg_data.dta with 30 observations. The GPA score is represented by two values, the lower interval score (lgpa) and the upper interval score (ugpa). The writing test scores, the teacher rating**
**and the type of program (a nominal variable which has three levels) are write, rating and type, respectively.**

**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/intreg_data, clear**
**list lgpa ugpa, clean**

**lgpa ugpa**
**1. 2.5 3**
**2. 3.4 3.8**
**3. 2.5 3**
**4. 0 2**
**5. 3 3.4**
**6. 3.4 3.8**
**7. 3.8 4**
**8. 2 2.5**
**9. 3 3.4**
**10. 3.4 3.8**
**11. 2 2.5**
**12. 2 2.5**
**13. 2 2.5**
**14. 2.5 3**
**15. 2.5 3**
**16. 2.5 3**
**17. 3.4 3.8**
**18. 2.5 3**
**19. 2 2.5**
**20. 3 3.4**

21. 3.4 3.8
22. 3.8 4
23. 2 2.5
24. 3 3.4
25. 3.4 3.8
26. 2 2.5
27. 2 2.5
28. 2 2.5
29. 2.5 3
30. 2.5 3

Note that there are two GPA responses for each observation, lgpa for the lower end of the interval and ugpa for the upper end.

```
summarize lgpa ugpa write rating

Variable | Obs Mean Std. Dev. Min Max
-------------+--------------------------------------------------------
lgpa | 30 2.6 .7754865 0 3.8
ugpa | 30 3.096667 .5708332 2 4
write | 30 113.8333 49.94278 50 205
rating | 30 57.53333 8.303441 48 72
```

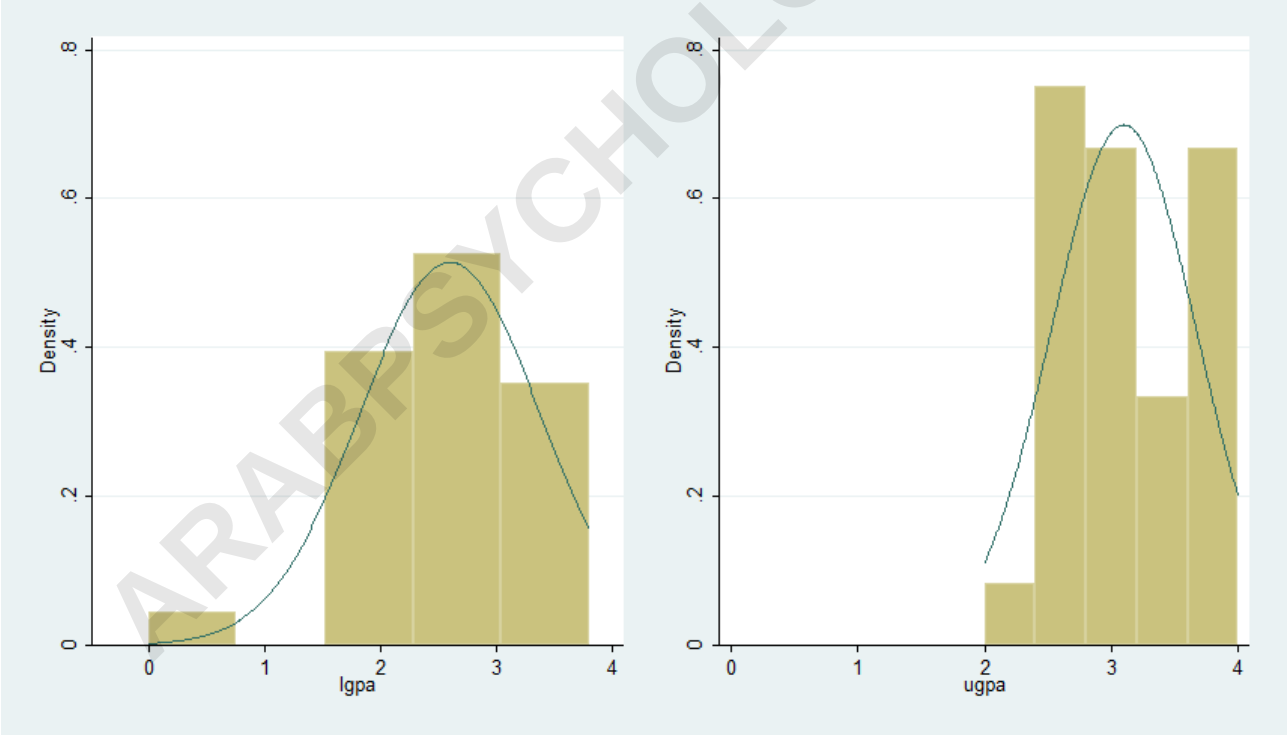
```
tabstat lgpa ugpa, by(type) stats(n mean sd)

Summary statistics: N, mean, sd
by categories of: type

type | lgpa ugpa
----------+--------------------
vocational | 8 8
| 1.75 2.4375
| .7071068 .1767767
----------+--------------------
general | 10 10
| 2.78 3.24
| .3852849 .3373096
----------+--------------------
academic | 12 12
| 3.016667 3.416667
| .6336522 .5474458
----------+--------------------
Total | 30 30
| 2.6 3.096667
| .7754865 .5708332
------------------------------
```

**Graphing these data can be rather tricky. Just to get an idea of what the distribution of GPA is, we will do separate histograms for lgpa and ugpa. We will also correlate the variables in the dataset.**

```
histogram ugpa, normal xlabel(0(1)4) name(hugpa)
histogram lgpa, normal xlabel(0(1)4) name(hlgpa)
graph combine hlgpa hugpa, ycommon xsize(7)
```






```
correlate lgpa ugpa write rating
(obs=30)
```

| lgpa ugpa write rating

---

lgpa | 1.0000

ugpa | 0.9488 1.0000

write | 0.6206 0.6572 1.0000

rating | 0.5355 0.5904 0.4763 1.0000

## 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.

## Interval regression

We will use the intreg command to run the interval regression
analysis. The intreg command requires two outcome variables, the
lower limit of the interval and the upper limit of the interval. The i. before
type 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. Note that this syntax was introduced**
**in Stata 11.**

**intreg lgpa ugpa write rating i.type**

**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 = -35.224403**
**Iteration 1: log likelihood = -33.142851**
**Iteration 2: log likelihood = -33.128906**
**Iteration 3: log likelihood = -33.128905**

**Interval regression Number of obs = 30**
**LR chi2(4) = 37.24**
**Log likelihood = -33.128905 Prob > chi2 = 0.0000**

**------------------------------------------------------------------------------**

**| Coef. Std. Err. z P>|z|**

**----------+----------**

**write | .0052847 .0016921 3.12 0.002 .0019683 .0086012**

**rating | .0133143 .0091197 1.46 0.144 -.0045601 .0311886**

**|**

**type |**

**2 | .374853 .19275 1.94 0.052 -.00293 .7526359**

**3 | .7097467 .1668399 4.25 0.000 .3827466 1.036747**

**|**

**_cons | 1.103863 .4452887 2.48 0.013 .2311137 1.976613**

**----------+----------**

**/lnsigma | -1.237263 .1596419 -7.75 0.000 -1.550155 -.9243703**

**----------+----------**

**sigma | .2901775 .0463245 .2122151 .3967812**

**----------**

**Observation summary: 0 left-censored observations**

**0 uncensored observations**

**0 right-censored observations**

**30 interval observations**

**contrast type**

**Contrasts of marginal linear predictions**

**Margins : asbalanced**

**---------------------------------------------**

**| df chi2 P>chi2**

**------------+--------------------------------**

**model |**

**type | 2 18.71 0.0001**

**---------------------------------------------**

**The two degree-of-freedom chi-square test indicates that type is a**
**statistically significant predictor of lgpa and ugpa.**

**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, because they**
**are adjusted for write and rating also.**

**margins type**

**Predictive margins Number of obs = 30**

**Model VCE : OIM**

## Expression : Linear prediction, predict()

---

| Delta-method
| Margin Std. Err. z P>|z|
---+---
type |
1 | 2.471456 .13236 18.67 0.000 2.212035 2.730877
2 | 2.846309 .118957 23.93 0.000 2.613157 3.07946
3 | 3.181203 .0969802 32.80 0.000 2.991125 3.37128

---

The expected mean GPA for students in program type 1 (vocational) is 2.47; the
expected mean GPA for students in program type 3 (academic) is 3.18.

If you would like to
compare interval 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**

**----------+----------**

**. | 30 -51.74729 -33.12891 6 78.25781 86.66499**

**----------**

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

**The intreg command does not compute an R2 or pseudo-R2. You can compute an approximate measure of fit by calculating the R2 between the predicted and observed values.**

**predict p**

**correlate lgpa ugpa p**

**(obs=30)**

**| lgpa ugpa p**

**----------+---------**

**lgpa | 1.0000**

**ugpa | 0.9488 1.0000**

**p | 0.7494 0.8430 1.0000**

**display .7494^2**

**.56160036**

**display .8430^2**

**.710649**

**The calculated values of approximately .56 and .71 are probably close to what you would find in an OLS regression if you had actual GPA scores. You can also make use of the Long and Freese utility command fitstat (search spostado)**
**(see How can I use**
**the search command to search for programs and get additional help? for more**
**information about using search),**
**which provides a number of pseudo-R2s in addition to other measures of fit. The**
**Cox-Snell pseudo-R2, in which the ratio of the likelihoods reflects**
**the improvement of the full model over the intercept-only model,**
**is close to our approximate estimates above.**

**fitstat**

**Measures of Fit for intreg of lgpa ugpa**

**Log-Lik Intercept Only: -51.747 Log-Lik Full Model: -33.129**

**D(23): 66.258 LR(4): 37.237**

**Prob > LR: 0.000**

**McFadden's R2: 0.360 McFadden's Adj R2: 0.225**

**ML (Cox-Snell) R2: 0.711 Cragg-Uhler(Nagelkerke) R2: 0.734**

**McKelvey & Zavoina's R2: 0.760**

**Variance of y*: 0.351 Variance of error: 0.084**

**AIC: 2.675 AIC*n: 80.258**

**BIC: -11.970 BIC': -23.632**

**BIC used by Stata: 86.665 AIC used by Stata: 78.258**

## Things to consider

## See also

## References