The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal (2005)
5, Number 3, pp. 395–404

Stings in the tails: Detecting and dealing with censored data

Ronán M. Conroy
Royal College of Surgeons in Ireland
rconroy@rcsi.ie

Abstract. Variables often show evidence of clustering at extreme values and of graininess, that is, of a limited number of distinct values. Scores on two subscales of a quality-of-life measure, traditionally analyzed with OLS regression or ANOVA models, provide examples. Ignoring or failing to detect such features of the data will result in poor estimates of effect size.

Keywords: st0090, censored data, diagnostic plots, intreg

1 Spotting the problem

1.1 Diagnostic plots

The SF-36 is probably the most widely used measure of patient health status in the world, having been translated into more than 40 languages. It has a number of subscales that reflect different aspects of well-being and function. It is frequently analyzed using ordinary least squares (OLS) models, such as t-tests, regressions, and analysis of variance.

The data used in this paper were collected as part of a study of quality of life in people with chronic lung disease. The investigators measured severity of disease in two ways: disease severity, which is based on clinical assessment by the doctor, and symptom severity, which reflects each person's rating of their difficulties in breathing. The question is simple: does your quality of life depend on how unwell the doctor thinks you are or on how unwell you feel, or both? For simplicity, the clinical severity and symptom severity ratings are reduced here to binary variables, reflecting presence or absence of severe disease and severe symptoms, respectively.

A little regression is tempting. Here is one using disease severity and symptom severity to predict pain scores, one of the quality-of-life subscales.

```
. regress sfpain symptom_severity disease_severity
```

Source	SS	df	MS
Model	5775.08058	2	2887.54029
Residual	132197.078	136	972.037336
Total	137972.158	138	999.798248

Number of obs	= 139
F(2, 136)	= 2.97
Prob > F	= 0.0546
R-squared	= 0.0419
Adj R-squared	= 0.0278
Root MSE	= 31.178

sfpain	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
symptom_se~y	14.17748	5.967738	2.38	0.019	2.375918	25.97905
disease_se~y	-4.380711	5.496686	-0.80	0.427	-15.25074	6.48932
_cons	65.45231	5.302044	12.34	0.000	54.9672	75.93743

Mean pain scores are 14 points higher in those with severe symptoms (95% confidence interval 2.4 to 26.0) in patients who have the same clinical severity of disease. In patients who have the same symptom severity, pain scores are a little lower (4.4 points) in those with severe disease, as diagnosed by their doctor, but the 95% confidence interval overlaps zero (−15.3 to 6.5). Essentially, for two people with the same disease severity, as rated by their doctor, the person with the more severe rating of their own symptoms will have a significantly higher pain score. But for two people with the same severity of symptoms, the difference between severe disease and less severe disease has no appreciable impact on pain scores. This conclusion is simple and quick but wrong.

Why do I say so? Graphing the distribution of pain scores shows an odd picture. Figure 1 shows two quantile–normal plots for the subscales obtained with `qnorm`. I use an aspect ratio of 1 in diagnostic plots. The square shape makes it easier to assess how and where the data deviate from the line of perfect agreement, which is the reference line on the graph.

```
. qnorm sfpain, aspect(1) saving(qnorm1, replace)
. qnorm sfsocial, aspect(1) saving(qnorm2, replace)
. graph combine "qnorm1.gph" "qnorm2.gph"
```

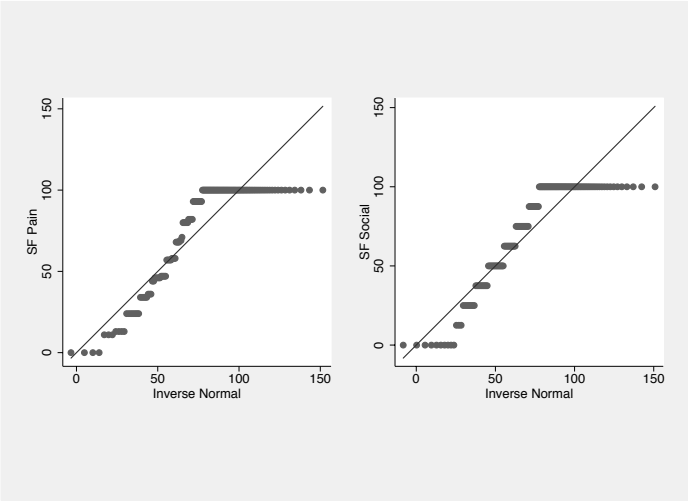

Figure 1: Test ceiling problems are shown by quantile–normal plots for pain and social-function problems

The pain scores, on the left, show a cluster of patients who score top marks. In fact, 46% of the patients scored 100. This is known as a *test ceiling* problem. The pain scale is unable to distinguish between people with high levels of pain, and they are all scored as 100. A more-nuanced pain scale would be able to distinguish between these people by including more items that differentiated severe pain from very severe pain. The scores of 0 also provoke doubt. Are those people utterly pain-free, or does the scale just not capture variation at that end either?

In comparison, note that the social-function subscale, on the right, has clear evidence of people piling up against both the top and the bottom of the distribution. It certainly fails to capture variation at either extreme of the range.

In the case of the pain scores, the ceiling effect will tend to understate the pain levels in groups who experience more pain since more of these people will be "off the scale". Figure 2 is a dot plot of the pain scores of the two symptom severity groups. The medians are shown as lines of plus symbols.

```
. dotplot sfpain, by(symptom_severity) median center bounded nogroup aspect(1)
```

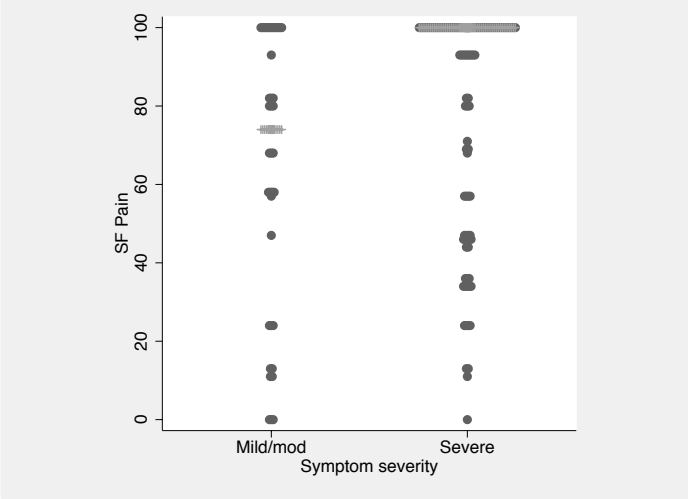

Figure 2: Pain scores by symptom severity level

1.2 Censored data

The patients with severe symptoms have scores that pile up against the scale's maximum value, to the extent that the median coincides with the highest value. (The medians are shown as grey lines, and it took me a moment to locate the median of the severe group, too.)

What can we do at this point? We could compare the pain scores of the two groups of patients using a Wilcoxon–Mann–Whitney test, but I am always reluctant to abandon regression models. They have two important advantages over the nonparametric approach. First, a regression model can assess how far quality of life is influenced by disease severity, symptom severity, or both, which is the important question. Second, differences between groups can be expressed in real-life units. This is not much of an advantage here because quality of life is measured on an arbitrary 100-point scale, but similar distributions can be observed in a number of other areas of research. In many of these, expressing effect sizes in real-life units is a great advantage.

Censored data arise in quite a number of settings:

- Tests cannot read beyond a certain range. In biological assays, there are upper and lower limits of detection. I have frequently discovered these in data analysis. In some cases, the researchers were unaware of the limits. Just because researchers do not tell you about limitations of measurement, it does not mean that they do not exist.

- Limits can be problematic for other reasons. Microbe counts also suffer from censoring. The number of colonies on a plate can become uncountably big, whereas no colonies forming on a plate does not mean that the water the sample came from is sterile. It just means that the 100 milliliters in the sample bottle did not produce any colonies.

- Floor and ceiling problems are also evident in some psychological and educational tests, including examinations. This happens when a test is too hard or too easy so that many people score zero or full marks. In percent terms, these marks should not be treated as real 0s or 100s. Rather, each such person's true score lay outside the score range of the test. In the present case, the measures of pain and social adjustment seem unable to capture the full range of patient experiences. Admittedly, as a group they are diverse and quite ill.

- Values may not be known precisely but may be known to fall into a particular range. This often happens when researchers ask sensitive questions about age or income, say, by asking respondents to tick which category they belong. This case may seem different, but it is equivalent. In the case of a floor or ceiling problem, the true score is known to be in the interval bounded on one side by the lowest or highest score on the test and on the other by a value that may be either definable in theory (microbe counts cannot be negative) or not (we have no way of knowing the maximum pain a person can experience if we measure it on the scale used with the patients).

- Finally, survival data are frequently censored. A person can be under observation for only a part of the time course being modeled. The most common case is that a person has been followed for a known period and has still not experienced the event of interest. Such data are known as right-censored. But people can also enter the study after the beginning of the process being studied. Since analysis of

failure-time data is a subject that has a Stata manual to itself, I simply mention it here for completeness.

What can be done to model censored data? One of the most useful commands is `intreg`.

2 Setting up data for intreg

2.1 What intreg expects

Alone among Stata's regression procedures, `intreg` needs two variables to represent the values of the response variable. Together they represent the range in which the value falls. In our case, we know that pain scores represented by 100 are actually anywhere between 100 and plus infinity, while values of 0 could really be either 0 or negative.

The two variables will represent the lower and upper points of the interval in which we know the measurement lies. In our case, we cannot define a maximum or minimum value beyond which pain simply cannot go. Where this is the case, we give the variable a missing value to indicate that the interval boundary is unknown. (In fact, it means that the interval is unbounded.)

```
. gen l_sfpain = sfpain if sfpain > 0 & !missing(sfpain)
. gen u_sfpain = sfpain if sfpain < 100 & !missing(sfpain)
```

These two variables define the upper and lower boundaries of each pain score. For scores that fall within the range of the test, the score is considered as a point value: I will return to that assumption later. Then the lower boundary, the variable `l_sfpain`, will be missing when the pain score is 0. The variable `u_sfpain` will be missing when the pain score is 100.

Note that Stata considers that a numeric missing value is greater than any other numeric value, so I have had to ensure that both the upper and lower boundaries are missing in cases where the pain score is missing. Stata will only treat data as missing in `intreg` if both upper and lower boundary variables contain missing values.

2.2 Getting results

Now we can run `intreg` as a normal regression command. Note that in place of the single-response variable, `intreg` needs two.

```
. intreg l_sfpain u_sfpain symptom_severity disease_severity
  (output omitted)
Interval regression                             Number of obs   =        139
                                                LR chi2(2)      =       6.61
Log likelihood = -434.77732                     Prob > chi2     =     0.0367
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
symptom_se~y	27.52809	11.01611	2.50	0.012	5.936912	49.11927
disease_se~y	-8.903424	10.24419	-0.87	0.385	-28.98166	11.17482
_cons	75.6211	9.658709	7.83	0.000	56.69037	94.55182
/lnsigma	3.975445	.094836	41.92	0.000	3.78957	4.16132
sigma	53.27382	5.052276			44.23736	64.15616

```
  Observation summary:         4  left-censored observations
                              71     uncensored observations
                              64 right-censored observations
                               0       interval observations
```

Of the 139 patients, 64 had pain scores at the top of the scale, while only 4 had scores at the bottom. The analysis comes to the same substantive conclusion as the original regression, but the effect sizes are now quite different. The original difference in quality-of-life scores between those with severe and those with mild or moderate disease was 14 points. It doubles in size when we account for censoring. While the analysis has not changed our conclusions, the original regression understated the difference between the two groups. This can be very important in another area in which I do a lot of work: analyzing bacteria counts in water.

3 Grainy data

Another look at the quantile–normal plot for social functioning shows not only strong evidence of floor and ceiling effects, but also a very stepped appearance. Variables assuming a limited number of discrete values are sometimes called 'grainy'. A tabulation makes the problem clear:

```
. tab sfsocial
```

SF Social	Freq.	Percent	Cum.
0	10	7.19	7.19
12.5	3	2.16	9.35
25	7	5.04	14.39
37.5	9	6.47	20.86
50	14	10.07	30.94
62.5	12	8.63	39.57
75	14	10.07	49.64
87.5	11	7.91	57.55
100	59	42.45	100.00
Total	139	100.00	

It seems that social function was originally measured as scores from 0 to 8, which were multiplied by 12.5 to scale them to percents. Arguably, the scores should be treated as intervals rather than as point values.

```
. gen l_sfsocial=sfsocial-6.25 if sfsocial > 0 & !missing(sfsocial)
(10 missing values generated)
. gen u_sfsocial=sfsocial+6.25 if sfsocial < 100
(59 missing values generated)
. list sfsocial l_sfsocial u_sfsocial in 1/10, clean
       sfsocial   l_sfso~l   u_sfso~l
 1.        62.5      56.25      68.75
 2.           0          .       6.25
 3.          50      43.75      56.25
 4.        62.5      56.25      68.75
 5.          50      43.75      56.25
 6.        37.5      31.25      43.75
 7.        87.5      81.25      93.75
 8.         100      93.75          .
 9.          25      18.75      31.25
10.           0          .       6.25
```

The original values are first, followed by the upper and lower boundary values. Values of 0 are now treated as having an upper boundary of 6.25 (halfway to the next possible score) and an unknown lower value, while values of 100 are treated as bounded by 93.75 below and an unknown value above. All other values have upper and lower bounds that differ by 6.25 from the original score.

What effect does this have on descriptive statistics?

```
. summarize sfsocial
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    sfsocial |       139    71.31295     32.4656         0        100
. intreg l_sfsocial u_sfsocial
  (output omitted)
Interval regression                             Number of obs   =        139
                                                LR chi2(0)      =       0.00
Log likelihood = -256.75801                     Prob > chi2     =          .
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   83.55192    5.04277    16.57   0.000     73.66827    93.43556
-------------+----------------------------------------------------------------
    /lnsigma |   3.944525   .0967705    40.76   0.000     3.754858    4.134191
-------------+----------------------------------------------------------------
       sigma |   51.65178   4.998368                      42.72815    62.43907
------------------------------------------------------------------------------
  Observation summary:        10  left-censored observations
                               0     uncensored observations
                              59 right-censored observations
                              70       interval observations
```

Note that I ran a regression without any predictors. Actually, the intercept is a predictor but not a variable. Hence when there are no predictors, the intercept is a

constant minimizing the sum of squared prediction errors, namely the mean. (The constant that minimizes the sum of absolute errors is the median, which is thus the basis for quantile regression.) So any OLS regression run without predictor variables will have a constant term that is the mean of the data. For example,

```
. regress sfsocial
```

Source	SS	df	MS		Number of obs =	139
					F(0, 138) =	0.00
Model	0	0	.		Prob > F =	.
Residual	145454.137	138	1054.01548		R-squared =	0.0000
					Adj R-squared =	0.0000
Total	145454.137	138	1054.01548		Root MSE =	32.466

sfsocial	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons	71.31295	2.753697	25.90	0.000	65.86806	76.75784

The coefficient is exactly what we got for the mean using `summarize` but much lower than the value we get from the interval regression model.

As you might expect, the use of interval regression has the same effect it had for pain scores: the difference between the two patient groups defined by symptom severity gets bigger compared with an OLS regression. We might be pleased that interval regression was unmasking a difference which had been attenuated by the failure of the measuring scale to capture the full extent of variation in the predicted variable. But is this just wishful thinking?

4 Assumptions and checking conclusions

4.1 Assumptions

Interval regression assumes that the data come from a normal distribution and that the behavior of the data in the tails can be inferred from its behavior in its observable range. In many cases, this is a fairly reasonable assumption. However, there may be other reasons why you might have a cluster of data at the minimum or maximum value, and it is worth thinking about these.

There will be clusters of zeros in a number of cases in which interval regression is not appropriate because the zeros represent real zeros, rather than the lower limit of detection. Common cases include zero-inflated count data, such as the number of times a person has visited the doctor in the past year. Some people just do not visit the doctor; of those who do, not everyone visits in any given year. Furthermore, visits have to occur in multiples of one, so the error structure of the data will be far from normal. This is a case for zero-inflated models for binomial or Poisson data (see the `zip` command).

Another case is where the data reflect a two-stage process. For example, water may become contaminated; if it does, it will have a microbe count. The factors that

determine whether it becomes contaminated may not be the factors that determine the extent of the contamination. This may be a case for Heckman models, which are also well beyond the scope of this article.

More importantly, the observable range of the data may not fall into a normal distribution. This is certainly the case with microbe counts, but these can be (and usually are) transformed using logarithms before analysis, as there are good biological reasons for measuring concentrations on a log scale. In the case of nonnormal distributions, there is often a scientific rationale for an appropriate transformation. Indeed, when there is not, the analyst can be on shaky ground.

4.2 Checking assumptions

I always run ordered logit regression models in parallel with interval regression models. The coefficients will differ, naturally, but I like to see that the substantive conclusions are the same.

```
. ologit sfpain symptom_severity disease_severity
  (output omitted)
Ordered logistic regression                     Number of obs   =        139
                                                LR chi2(2)      =       5.71
                                                Prob > chi2     =     0.0575
Log likelihood = -297.68788                     Pseudo R2       =     0.0095
```

sfpain	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
symptom_se~y	.7958587	.3479314	2.29	0.022	.1139256	1.477792
disease_se~y	-.2789116	.3230477	-0.86	0.388	-.9120735	.3542504

(*output omitted*)

This model gives a similar result to the interval regression: the person's experience of their symptoms is what affects their experience of pain, not the doctor's rating of disease severity. Note that although the coefficients themselves have no immediate interpretation in ordered logistic regression, the signs are interpretable and agree with the interval regression results.

I could have done the analysis using ordered logistic regression, but I would have lost information about the size of the effects. In this case, it would not have been critical, but there are many cases in which the effect size and its confidence interval are very important pieces of information. In such cases, interval regression gives the analyst a way to estimate them.

5 Conclusion

- Never take data on trust: the observed values are not observed until you look at them. The SF-36 scales that provided the data used in this paper have been analyzed for years using standard OLS models. On inspection, they reveal distri-

butional properties conveniently ignored by generations of researchers. Of course, a number of papers have pointed out the problem, but these have been ignored until recently.

- The usual way of analyzing data can as easily be wrong as right. The argument that 'everyone else uses ANOVA' is informative but not decisive.
- When data cluster at one or both extremes, suspect censoring.
- When data take on a restricted number of values, consider using interval regression.
- There is a strong assumption of underlying normality in interval regression. Whenever data are counts, zero-inflated models for counted data may be more appropriate.
- It is worth the effort of trying to build a regression model rather than running for nonparametric tests at the first hint of trouble. Regression models allow adjustment for confounding factors and, with a little extra effort, allow estimation of effect sizes and their confidence intervals.
- It is good practice to run a model with less restrictive assumptions to make sure that the substantive conclusions of a model are supported. Ordered logistic regression can help to validate interval regression models, just as robust regression can be used to validate OLS regression.
- Finally, remember that you can use interval regression without any predictor variables to estimate the mean of the data and its confidence interval, just as you can run ordinary regression to estimate the mean of uncensored data. (See Conroy [2002] for more examples of this.)

6 References

Conroy, R. M. 2002. Choosing an appropriate real-life measure of effect size: the case of a continuous predictor and a binary outcome. *Stata Journal* 2(3): 290–295.

About the Author

Ronán Conroy studied piano with Marie Jones and graduated in music from Trinity College, Dublin. He teaches at the Royal College of Surgeons in Ireland. His research interests run from cardiovascular risk prediction to low-technology health solutions for developing countries. He can offer no adequate explanation for his apparent career.