Outliers

[NOTE: These notes draw heavily from several sources, including Fox's Regression Diagnostics; Pindyck and Rubinfeld; Statistics for Social Data Analysis, by George Bohrnstedt and David Knoke, 1982; Norusis's SPSS 11 chapter 22 on "Analyzing residuals;" Hamilton's chapter on "Robust regression." I'm hitting highlights here, but the readings include lots of other good suggestions and details.]

Description of the problem. One problem with least squares occurs when there are one or more large deviations, i.e. cases whose values differ substantially from the other observations. The slope and intercept of the least squares line is very sensitive to data points which lie far from the true regression line. These points are called *outliers*, i.e. extreme values of observed variables that can distort estimates of regression coefficients.

Detecting the problem

- Scatterplots, frequencies can reveal atypical cases
- Can also look for cases with very large residuals.
- Suspicious correlations sometimes indicate the presence of outliers.
- SPSS has some good routines for detecting outliers.
    - There is always the FREQUENCIES routine, of course.
    - The PLOTS command can do scatterplots of 2 variables.
    - The EXAMINE procedure includes an option for printing out the cases with the 5 lowest and 5 highest values.
    - The REGRESSION command can print out scatterplots (particularly good is *ZRESID by *ZPRED, which is a plot of the standardized residuals by the standardized predicted values). In addition, the regression procedure will produce output on CASEWISE DIAGNOSTICS, which indicate which cases are extreme outliers. This is particularly useful in that you see which cases stand out even after all IVs have been controlled for.
- Stata counterparts to the above include
    - The `tab1` and `table` commands
    - The `scatter` command (also `graph7` will work, and seems to be quicker albeit old-fashioned; Stata redid its graphics in Stata 8 but `graph7` will let you use the old graphics)
    - The `extremes` command. This is an add-on module written by Nick Cox
    - There are several plotting routines, including `rvfplot` (residuals versus fitted)
    - The `predict` command has several options that can help you identify outliers

- Stata also has lots of other routines, many of them graphics-oriented, for detecting outliers. I won't go through many of them, but I'll include links on the course web page that give examples
- Probably the most critical difference between SPSS and Stata is that Stata includes additional routines (e.g., `rreg`, `qreg`) for addressing the problem of outliers, which we will discuss below.

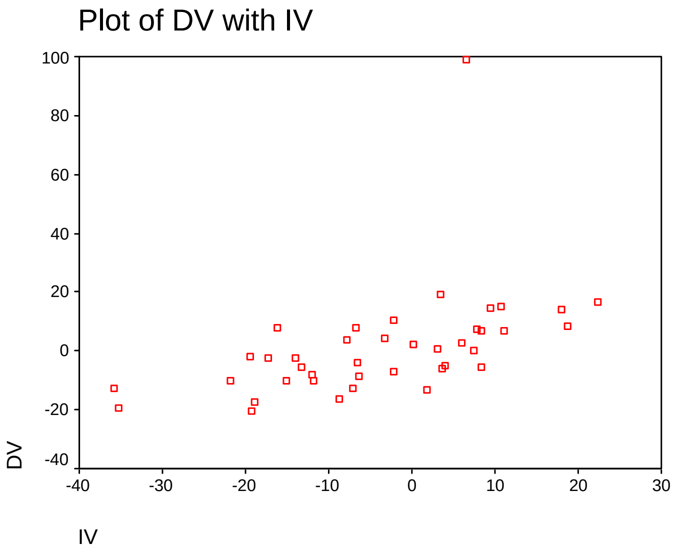
SPSS Example. Following is a hypothetical example of 40 cases. I constructed the data set so the DV and IV would have a correlation of about .7. I then changed one of the DV values into an extreme outlier. Note how the first three analyses (PLOT, EXAMINE, and REGRESSION) all provide means of detecting the outlier. Then, see how the results change once the outlier is deleted and the regression is rerun.

```
Get File = 'D:\Soc593\Outlier.sav'.
```

* This program shows some of the ways SPSS can be used to identify outliers.
* Do a scatterplot of vars to visually ID cases.
* Note that one case is way out of line with the rest.

```
Plot plot = dv with iv.
```

Plot



* Use examine procedure to id cases with extreme values on X or Y.
* However, note that these need not be outliers on a regression line.
* Note that Case 9 has a very extreme, and also very suspicious, value for DV.

```
EXAMINE
  VARIABLES=dv iv
  /PLOT NONE
  /STATISTICS EXTREME
  /MISSING LISTWISE
  /NOTOTAL.
```

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
DV	40	100.0%	0	.0%	40	100.0%
IV	40	100.0%	0	.0%	40	100.0%

Extreme Values

			Case Number	Value
DV	Highest	1	9	99.00
		2	22	19.36
		3	13	16.84
		4	36	15.23
		5	32	14.52
	Lowest	1	5	-20.45
		2	7	-19.62
		3	8	-17.22
		4	38	-16.34
		5	16	-13.34
IV	Highest	1	13	22.45
		2	35	18.77
		3	23	17.97
		4	10	11.15
		5	36	10.65
	Lowest	1	21	-35.75
		2	7	-35.24
		3	28	-21.74
		4	30	-19.53
		5	5	-19.23

```
* Run regression with outlier in.
* Outlier will also show up in the plot.
* /Casewise prints out stats that help to ID extreme outliers, if any.
* /Scatterplot graphically helps to ID extreme outliers.
* SPSS Regression has many other options for analyzing residuals
* that may sometimes be useful.
```

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT dv
  /METHOD=ENTER iv
  /Casewise defaults dfbeta
  /SCATTERPLOT=(*ZRESID ,*ZPRED ) .
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
DV	1.2328	18.99052	40
IV	-3.7314	13.59168	40

Correlations

		DV	IV
Pearson Correlation	DV	1.000	.479
	IV	.479	1.000
Sig. (1-tailed)	DV	.	.001
	IV	.001	.
N	DV	40	40
	IV	40	40

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	IV[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: DV

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.479[a]	.229	.209	16.89297

a. Predictors: (Constant), IV

b. Dependent Variable: DV

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3220.796	1	3220.796	11.286	.002[a]
	Residual	10844.154	38	285.372		
	Total	14064.950	39			

a. Predictors: (Constant), IV

b. Dependent Variable: DV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.728	2.772		1.345	.187
	IV	.669	.199	.479	3.360	.002

a. Dependent Variable: DV

Casewise Diagnostics[a]

Case Number	Std. Residual	DV	Predicted Value	Residual	DFBETA	
					(Constant)	IV
9	5.379	99.00	8.1398	90.8602	2.872	.136

a. Dependent Variable: DV

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-20.1733	18.7376	1.2328	9.08760	40
Std. Predicted Value	-2.356	1.926	.000	1.000	40
Standard Error of Predicted Value	2.67222	6.90899	3.63283	1.04815	40
Adjusted Predicted Value	-21.6857	18.9965	1.2077	9.22506	40
Residual	-18.3091	90.8602	.0000	16.67499	40
Std. Residual	-1.084	5.379	.000	.987	40
Stud. Residual	-1.100	5.489	.001	1.008	40
Deleted Residual	-18.8626	94.6275	.0251	17.37645	40
Stud. Deleted Residual	-1.103	11.900	.162	1.961	40
Mahal. Distance	.001	5.549	.975	1.314	40
Cook's Distance	.000	.625	.021	.098	40
Centered Leverage Value	.000	.142	.025	.034	40

a. Dependent Variable: DV

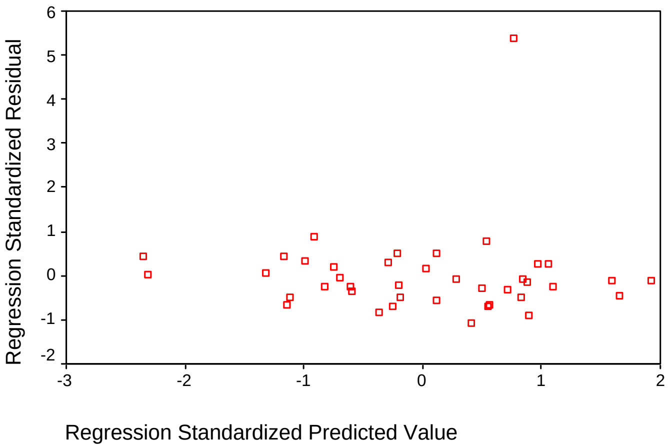
To explain a few of these statistics: DFBETA shows how much a coefficient would change if that case were dropped from the data. In this case, it shows that the effect of IV would drop by .136 if case 9 were dropped. [CAUTION: To make things confusing, Stata uses the term dfbeta to refer to what SPSS would call standardized dfbetas. There are other instances where Stata and SPSS use different naming conventions. With a standardized dfbeta, values of 1 or larger are generally considered important; it has also been suggested that standardized dfbetas > 2/Sqrt(N) should be checked.]

Std residual and Stud Residual are slightly different ways of standardizing the residuals; values of 3 or greater (or -3 or less) may be problematic. Values of Cook's distance that are greater than 4/N (in this case, 4/40 = .10) may be problematic. From the above, we see that we have some very large standardized residuals and a large value for Cook's distance; further, we see that case 9 in particular is a problem (the fact that it was printed in the casewise diagnostics means it has a standardized residual of at least 3.)

Charts

Scatterplot

Dependent Variable: DV



* Get rid of the outlying case.

```
USE ALL.
COMPUTE filter_$=(dv < 99).
VARIABLE LABEL filter_$ 'dv < 99 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
```

* Rerun the regression without the outlier. Note changes in
* the correlation and in the coefficients. Now that the weird case is gone,
* the slope goes down because the regression line doesn't need to try to
* reach the outlier. Also note that "Casewise diagnostics"
* does not show up anymore because there are no extreme outliers left.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT dv
  /METHOD=ENTER iv
  /Casewise
 /SCATTERPLOT=(*ZRESID ,*ZPRED ) .
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
DV	-1.2741	10.58975	39
IV	-3.9963	13.66436	39

Correlations

		DV	IV
Pearson Correlation	DV	1.000	.688
	IV	.688	1.000
Sig. (1-tailed)	DV	.	.000
	IV	.000	.
N	DV	39	39
	IV	39	39

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	IV[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: DV

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.688[a]	.473	.459	7.79169

a. Predictors: (Constant), IV

b. Dependent Variable: DV

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2015.146	1	2015.146	33.193	.000[a]
	Residual	2246.284	37	60.710		
	Total	4261.430	38			

a. Predictors: (Constant), IV

b. Dependent Variable: DV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.856	1.301		.658	.515
	IV	.533	.093	.688	5.761	.000

a. Dependent Variable: DV

Residuals Statistics[a]

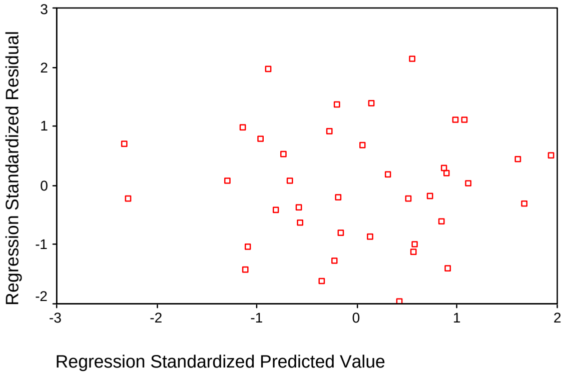
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-18.1951	12.8196	-1.2741	7.28218	39
Std. Predicted Value	-2.324	1.935	.000	1.000	39
Standard Error of Predicted Value	1.24920	3.19102	1.69670	.49064	39
Adjusted Predicted Value	-19.3137	12.2492	-1.3109	7.30290	39
Residual	-15.1845	16.6508	.0000	7.68848	39
Std. Residual	-1.949	2.137	.000	.987	39
Stud. Residual	-1.979	2.174	.002	1.008	39
Deleted Residual	-15.6618	17.2284	.0368	8.02534	39
Stud. Deleted Residual	-2.065	2.296	.004	1.029	39
Mahal. Distance	.002	5.399	.974	1.306	39
Cook's Distance	.000	.099	.022	.025	39
Centered Leverage Value	.000	.142	.026	.034	39

a. Dependent Variable: DV

Charts

Scatterplot

Dependent Variable: DV

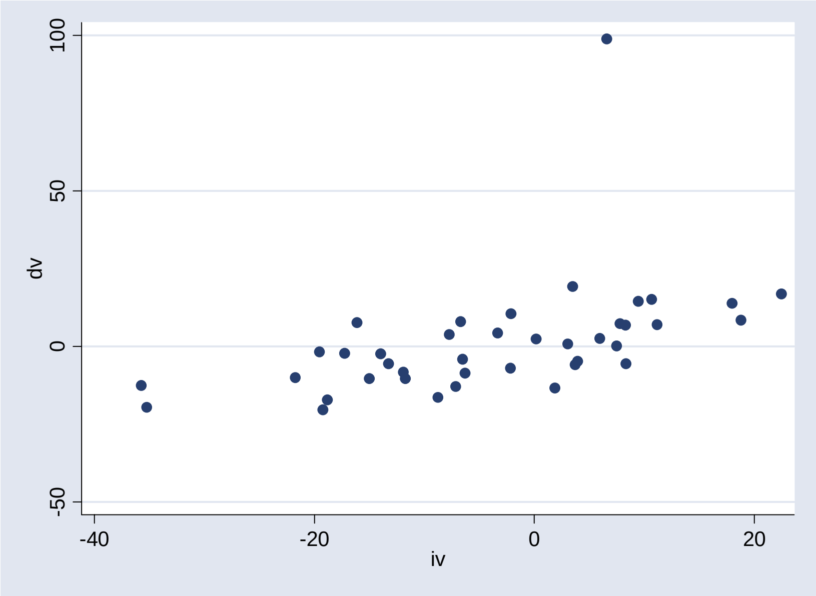


Having dropped the problematic case, we see that all is pretty much well now.

Stata Example

As is often the case, we can get very similar output with Stata; but, instead of a few big commands with several options, we execute several smaller commands instead. First, we can use the `scatter` command to plot the dv and the iv.

```
. scatter dv iv
```



Again, we see that extreme outlier. But, if our eyes had missed it, the `extremes` command (similar to SPSS Examine) identifies the cases with the most extreme high and low values.

```
. extremes dv iv

  +------------------------------+
  | obs:          dv          iv |
  |------------------------------|
  |   5.   -20.44946   -19.22762 |
  |   7.   -19.62192   -35.24309 |
  |   8.   -17.21676   -18.83887 |
  |  38.   -16.34352   -8.757764 |
  |  16.   -13.33637    1.862242 |
  +------------------------------+

  +----------------------------+
  |  32.   14.51918   9.434002 |
  |  36.   15.22688   10.65133 |
  |  13.   16.84143   22.44931 |
  |  22.   19.36333   3.484266 |
  |   9.         99   6.599043 |
  +----------------------------+
```

Notice the format of the command and the layout of the output are a little different than SPSS. You could just specify one variable, and it would give you the extreme values for it. If you specify two or more variables, it will give you the extreme values of the first variable, and the values of the other variables for those same cases. This can be useful for determining if the extreme values really are that extreme, given the values of the other variables. Again we see that case 9 seems very different from the rest of the cases and has a very suspicious value of 99. To check for extreme values on iv,

```
. extremes iv dv

  +-------------------------------+
  | obs:          iv          dv |
  |-------------------------------|
  |  21.   -35.74697   -12.64403 |
  |   7.   -35.24309   -19.62192 |
  |  28.   -21.73665   -10.07544 |
  |  30.     -19.526   -1.810622 |
  |   5.   -19.22762   -20.44946 |
  +-------------------------------+

  +-----------------------------+
  |  36.   10.65133   15.22688 |
  |  10.   11.14967   7.092674 |
  |  23.   17.97275   13.88848 |
  |  35.   18.77257   8.434381 |
  |  13.   22.44931   16.84143 |
  +-----------------------------+
```

Once again, we'll run the correlational and regression analysis:

```
. corr dv iv, means
(obs=40)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
          dv |     1.232763     18.99052    -20.44946           99
          iv |    -3.731381     13.59168    -35.74697     22.44931


             |       dv       iv
-------------+------------------
          dv |   1.0000
          iv |   0.4785   1.0000

. reg dv iv

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =   11.29
       Model |  3220.79618     1  3220.79618           Prob > F      =  0.0018
    Residual |  10844.1543    38  285.372482           R-squared     =  0.2290
-------------+------------------------------           Adj R-squared =  0.2087
       Total |  14064.9505    39  360.639757           Root MSE      =  16.893

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .6686151   .1990218     3.36   0.002     .2657166    1.071514
       _cons |   3.727621   2.772329     1.34   0.187    -1.884665    9.339907
------------------------------------------------------------------------------
```

Now, the residual statistics require a little more work than they do with SPSS. Basically, we use the `predict` command to compute the measures we want, and then run the summary statistics on them. I'll show a few examples; typing `help regress` will show you other options. Some of these are the same as SPSS, although sometimes SPSS and Stata use different names for the same thing. (e.g. dfbeta = SPSS's SDBETA; rstandard = SPSS's Student residual; rstudent = SPSS's Stud. Deleted residual).

[CAUTION: In general, `predict` calculates the requested statistic for all observations possible, whether they were used in fitting the model or not. This can be quite handy at times, as we will see later in the course. But, if your regression was not run on all the cases, e.g. you were analyzing a subsample, you might want to modify the following commands to something like `predict stdresid if e(sample), rstandard`. The `if` parameter will limit the computations to the cases used by the previous regression.]

```
. * Predicted value - predict gives this by default
. predict yhat
(option xb assumed; fitted values)

. * Standardized predicted value -- use the zscore command
. zscore yhat
z_yhat created with 0 missing values

. * Get the residual
. predict resid, residual

. * Standardized residuals -- values more extreme than 3 may be a problem
. predict stdresid, rstandard

. * Studentized residual
. predict rstudent, rstudent

. * DFBetas -- SPSS calls these SDBETAS -- values larger than 1 are a problem
. dfbeta
                             DFiv:  DFbeta(iv)
. * Get Cook's Distance measure -- values greater than 4/N may cause concern
. predict cooksd, cooksd

. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          dv |        40    1.232763    18.99052  -20.44946         99
          iv |        40   -3.731381    13.59168  -35.74697   22.44931
        yhat |        40    1.232763    9.087602  -20.17334   18.73757
      z_yhat |        40    9.02e-18           1  -2.355528   1.926229
       resid |        40    4.36e-08    16.67499  -18.30911   90.86016
-------------+--------------------------------------------------------
    stdresid |        40    .0007058      1.0076   -1.10009   5.488952
    rstudent |        40    .1617825    1.960709  -1.103228   11.90047
        DFiv |        40    .0190428    .2449757  -.1998377   1.478083
      cooksd |        40    .0208666      .09811   .0000153   .6246131
```

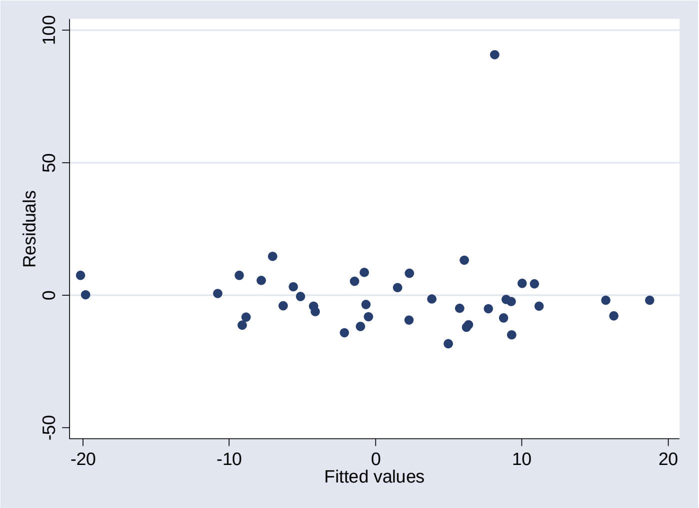
The last four s are especially revealing. They show us that at least one standardized and studentized residual is much larger than 3, at least one of the dfbetas is larger that one (which means that deletion of that case would cause a substantial change in the parameter estimates), and that at least one Cook's distance is much larger than 4/N = .10. Again using the `extremes` command, it is pretty obvious case 9 is our biggest problem.

```
. extremes  stdresid rstudent DFiv cooksd

  +------------------------------------------------------+
  | obs:    stdresid    rstudent         DFiv     cooksd |
  |------------------------------------------------------|
  |  16.    -1.10009   -1.103228   -.0737939   .0182921 |
  |  11.   -.9054657   -.9032696   -.1315701   .0194316 |
  |  38.   -.8537651   -.8506546    .0511074   .0106943 |
  |  39.   -.7236397    -.719026   -.0640732   .0088459 |
  |  25.   -.7062296    -.701494     .028504   .0068167 |
  +------------------------------------------------------+

  +----------------------------------------------------+
  |  29.   .4976577    .492674    .0094423   .0032218 |
  |  40.    .519945   .5148928   -.0183424   .0036419 |
  |  22.   .8006765   .7968211    .0688563    .010674 |
  |   1.   .8894223   .8869218   -.1324265   .0191859 |
  |   9.   5.488952   11.90047    1.478083   .6246131 |
  +----------------------------------------------------+
```

NOTE: To be thorough, we should probably run the `extremes` command specifying the other residual measures first as well; the cases that are the most extreme on stdresid won't necessarily be the most extreme on other measures, although there tends to be a lot of overlap.

The `rvfplot` (residuals versus fitted) command run after regress gives a plot of the predicted values versus the residuals (this isn't standardized like the earlier SPSS plot but it looks the same except that the axes are scaled differently). Ideally, this should look like a random scatter of points. In this case, it does not.

```
. rvfplot
```



We can now basically repeat the process without case 9, like we did in SPSS. We could give a command like

```
. drop in 9
```

which would drop case 9. Or, we could do something like

```
. recode dv(99=0)(else=1), gen(touse)
```

and then give commands like

```
. reg dv iv if touse
```

Since repeating the analysis without the dropped case is a fairly straightforward process, and the outcomes are pretty much the same as in SPSS, I won't bother going though it. (One caution: Stata doesn't let you overwrite existing variables easily, so you probably want to `drop` the old variables first before going through the predict commands, or else use different variable names).

Dealing with outliers (Both Stata and SPSS)

- First, check to make sure there are no coding errors. Has an extra zero been added to the outlying case?
- Make sure missing data coding is correct. For example, if you have a variable whose coding runs from 0 to 7 with an MD code of 99, and you have failed to tell SPSS that 99 is an MD code (or have not recoded 99 to . in Stata), the regression estimates will be way, way off. I've seen this produce extremely high correlations, when both the IVs and DVs were not being properly treated as missing.
- Run the regression both with and without the outlying cases. If the results are substantially different, this should be noted. You should either explain why some cases were deleted, or present both sets of analyses.
- Large outliers might be accounted for by adding more explanatory variables. Naturally, you prefer to explain the values of cases, rather than just discard them.
- Remember, though, that outliers may represent important information about the relationship between variables. Don't throw the outlier away without examining it first. Maybe you will catch a coding error. Perhaps you can explain why this case doesn't really fall into the population of interest. Or, perhaps you can add IVs which will explain why this case's values differ so much from the rest.

Dealing with outliers (Stata) – Robust Regression Techniques

One advantage of Stata over SPSS is that it includes so-called robust regression routines that are better able to handle outliers. (We would, of course, still want to do all the things described above, but if the outliers do appear to be legitimate, these techniques can help.) These routines work best when it is the DV that has outliers rather than the IVs. As Hamilton notes (Statistics With Stata, Version 8, p. 239):

> OLS tends to track outliers, fitting them at the expense of the rest of the sample. Over the long run, this leads to greater sample-to-sample variation or inefficiency when samples often contain outliers. Robust regression methods aim to achieve almost the efficiency of OLS with ideal data and substantially better than OLS efficiency in non-ideal (for example, nonnormal errors) situations….[The Stata routines] `rreg` and `qreg` resist the pull of outliers, giving them better than OLS efficiency in the face of nonnormal, heavy-tailed error distributions.

To show how this works, first, let us repeat our regression results with all 40 cases, including the outlier:

```
. reg dv iv

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =   11.29
       Model |  3220.79618     1  3220.79618           Prob > F      =  0.0018
    Residual |  10844.1543    38  285.372482           R-squared     =  0.2290
-------------+------------------------------           Adj R-squared =  0.2087
       Total |  14064.9505    39  360.639757           Root MSE      =  16.893

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .6686151   .1990218     3.36   0.002     .2657166    1.071514
       _cons |   3.727621   2.772329     1.34   0.187    -1.884665    9.339907
------------------------------------------------------------------------------
```

With the problematic case dropped,

```
. reg dv iv if dv!=99

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  1,    37) =   33.19
       Model |  2015.14589     1  2015.14589           Prob > F      =  0.0000
    Residual |  2246.28396    37  60.7103773           R-squared     =  0.4729
-------------+------------------------------           Adj R-squared =  0.4586
       Total |  4261.42986    38  112.142891           Root MSE      =  7.7917

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .5329324   .0925018     5.76   0.000     .3455059    .7203589
       _cons |   .8556491   1.301279     0.66   0.515    -1.780992     3.49229
------------------------------------------------------------------------------
```

Now, we'll see what happens when we run Stata's `rreg` (robust regression) routine with all 40 cases:

```
. rreg dv iv, nolog

Robust regression estimates                            Number of obs =      40
                                                       F(  1,    38) =   28.83
                                                       Prob > F      =  0.0000

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .5352655   .0996925     5.37   0.000     .3334486    .7370824
       _cons |   .7893176   1.388694     0.57   0.573    -2.021946    3.600581
------------------------------------------------------------------------------
```

Note that we get estimates that are very similar to what we got when we used the regress command and dropped case 9.

`rreg` is a bit hard to explain. Basically, it goes through an iterative procedure (as Hamilton notes, it uses “iteratively reweighted least squares with Huber and biweight functions tuned for 95% Gaussian efficiency”), where the more extreme an outlier is, the less heavily it gets weighted in the regression calculations. Very extreme cases get dropped altogether. In this problem, `rreg` basically dropped case 9 altogether, which is why its final results looked so similar to the results we got when we ran a regression with case 9 excluded.

Another alternative is `qreg`, which stands for quantile regression (you’ll also hear it referred to as Least Absolute Value Models or minimum L1-norm models). The most common form of quantile regression is median regression, where the goal is to estimate the median (rather than the mean) of the dependent variable, conditional on the values of the independent variables. Put another way, median regression finds a line through the data that minimizes the sum of the absolute residuals rather than the sum of the squares of the residuals as in ordinary regression (hence the term Least Absolute Value as opposed to Least Squares) . Medians are less affected by outliers than means are, so `qreg` can do better than `regress` when there are extreme outliers.

```
. qreg dv iv, nolog

Median regression                                    Number of obs =        40
  Raw sum of deviations 444.7982 (about -2.2611923)
  Min sum of deviations 335.8461                     Pseudo R2     =    0.2449

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .6079845   .0891699     6.82   0.000     .4274695    .7884994
       _cons |   1.805331   1.326887     1.36   0.182    -.8808107    4.491472
------------------------------------------------------------------------------
```

As to which routine is better, and when? Well, that is a good question. In this particular case, `rreg` seems better, but all it basically did was drop the extreme case, which we could have done by ourselves. Had 99 been a legitimate code, `qreg` might have seemed the more appropriate choice. Hamilton also argues that when the IVs have outliers, `rreg` tends to do better because it tends to just drop such observations rather than try to fit them.

Again, I would stress that either of these routines should only be used after you have checked out other issues, e.g. are there coding errors in the data, is missing data being handled properly, would the addition of some other variable to the model make the outliers not be outliers any more? But, if none of these solve the problem, here is some advice that was offered on Statalist when I asked about this on Jan. 30, 2004.

> Nicholas Cox (he wrote a couple of messages and hopefully I have combined them correctly): This raises the old classical trope, beaten almost to death by the late Sir Isaiah Berlin in many of his essays on intellectual history, that the fox knows many things, but the hedgehog knows one big thing.
>
> When attacked, the hedgehog has just one means of defence, although it is usually effective. -qreg- is a hedgehog. The fox has many different tricks. -rreg- is a fox. Its mixed strategy is an attempt to be smart in different ways.
>
> My experience loosely matches Richard's, certainly in terms of wanting to think that -qreg- is as good because of the much greater ease in explaining it. At the same time, [if you have] “well-behaved” data + a “few” outliers (n ~ 1) it is sensible to use robust regression as a check on standard [regression]. [But if you have long-tailed data] you are possibly working on inappropriate scales and should wonder about reaching for a transformation or, in some frameworks, a different link function.
>
> Also from Cox, as to whether outliers are more likely to be “real” or just coding mistakes: I think it depends, partly, on the kind of data you deal with. In fact these tribal differences among groups of statistical users are one of the persistently interesting features of Statalist.
>
> In geography (that's my field) the big cities, countries, rivers, storms, etc. really are big and they really are important, and my advice to students and colleagues hinges on the idea that most outliers are likely to be genuine and important. Often this means taking logarithms! Also, there's usually a story behind each outlier and extra information somewhere.
>
> As many people may know, the Antarctic ozone hole was only spotted belatedly because an outlier-trapping program led to some very low values being overlooked, until someone had a closer look at the real data.
>
> In some other fields it may be that most outliers are mistakes and/or that in terms of advancing science it's better strategy to ignore them. The person who reports watching 180 hours of television a week is likely to be confused about something or other; and short of re-interviewing or some smart way of finding out that they really meant 18.0 or 108, the only possible thing may be to omit that data point.

> Michael Blasnik: One difference between qreg and rreg is that they attempt to estimate different versions of the central tendency -- qreg estimates the median while rreg comes closer (in theory) to estimating a robust mean. The difference may be negligible in essentially symmetrical distributions, but for skewed distributions where the mean and median are not expected to be equal, one would expect their estimates to deviate systematically. If you really want to model the mean but are concerned about outliers, then rreg may be a better choice than qreg. If you want to model the median (or think the underlying distribution is fairly symmetrical), then qreg may be preferred.
>
> I usually look at both and then try to figure out any substantive differences in results, but I'm generally partial to the coefficient estimates from rreg (I often deal with skewed distributions where the median is noticeably lower than the mean). On the other hand, I sometimes find rreg's std errors estimates questionable.