Robust Regression Modeling with STATA lecture notes

Robert A. Yaffee, Ph.D.
Statistics, Social Science, and Mapping Group
Academic Computing Services
Office:
75 Third Avenue, Level C-3
Phone: 212-998-3402
Email: yaffee@nyu.edu
What does Robust mean?

1. Definitions differ in scope and content. In the most general construction: Robust models pertains to stable and reliable models.

2. Strictly speaking:
   Threats to stability and reliability include
   influential outliers
   Influential outliers played havoc with statistical estimation. Since 1960, many robust techniques of estimation have developed that have been resistant to the effects of such outliers.
   
   SAS Proc Robustreg in Version 9 deals with these.
   S-Plus robust library in Stata rreg, prais, and arima models

3. Broadly speaking: Heteroskedasticity
   Heteroskedastically consistent variance estimators
   Stata regress y x1 x2, robust

4. Non-normal residuals
   1. Nonparametric Regression models
      Stata qreg, rreg
   2. Bootstrapped Regression
      1. bstrap
      2. bsqreg
Outline

1. Regression modeling preliminaries
   1. Tests for misspecification
      1. Outlier influence
      2. Testing for normality
      3. Testing for heteroskedasticity
      4. Autocorrelation of residuals
   2. Robust Techniques
      1. Robust Regression
      2. Median or quantile regression
      3. Regression with robust standard errors
      4. Robust autoregression models
   3. Validation and cross-validation
      1. Resampling
      2. Sample splitting
   4. Comparison of STATA with SPLUS and SAS
Preliminary Testing: Prior to linear regression modeling, use a matrix graph to confirm linearity of relationships

graph y x1 x2, matrix
The independent variables appear to be linearly related with $y$

We try to keep the models simple. If the relationships are linear then we model them with linear models. If the relationships are nonlinear, then we model them with nonlinear or nonparametric models.
Theory of Regression Analysis

What is linear regression Analysis?
Finding the relationship between a dependent and an independent variable.

\[ Y = a + bx + e \]

Graphically, this can be done with a simple Cartesian graph.
The Multiple Regression Formula

\[ Y = a + bx + e \]

- \( Y \) is the dependent variable
- \( a \) is the intercept
- \( b \) is the regression coefficient
- \( x \) is the predictor variable
Graphical Decomposition of Effects

Decomposition of Effects

\[ y_i - \bar{y} = \text{Total Effect} \]

\[ y_i - \hat{y}_i = \text{error} \]

\[ \hat{y} - \bar{y} = \text{regression effect} \]
Derivation of the Intercept

\[ y = a + bx + e \]
\[ e = y - a - bx \]

\[ \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_i - b \sum_{i=1}^{n} x_i \]

Because by definition \( \sum_{i=1}^{n} e_i = 0 \)

\[ 0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_i - b \sum_{i=1}^{n} x_i \]

\[ \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \]

\[ na = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \]

\[ a = \bar{y} - b\bar{x} \]
Derivation of the Regression Coefficient

Given: \( y_i = a + b x_i + e_i \)
\( e_i = y_i - a - b x_i \)

\[
\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a - b x_i)
\]

\[
\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - b x_i)^2
\]

\[
\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b} = 2x_i \sum_{i=1}^{n} (y_i) - 2b \sum_{i=1}^{n} x_i x_i
\]

\[
0 = 2x_i \sum_{i=1}^{n} (y_i) - 2b \sum_{i=1}^{n} x_i x_i
\]

\[
b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}
\]
• If we recall that the formula for the correlation coefficient can be expressed as follows:
From which it can be seen that the regression coefficient $b$, is a function of $r$. 

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} (x_i^2) \sum_{i=1}^{n} (y_i^2)}}$$

Where

$x = x_i - \bar{x}$

$y = y_i - \bar{y}$

$$b_j = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

From which it can be seen that the regression coefficient $b$, is a function of $r$.

$$b_j = r * \frac{sd_y}{sd_x}$$
Extending the bivariate to the multivariate Case

2. Multivariate Case

Suppose we have two independent variables: $x_1$ and $x_2$

We wish to examine the change in $y$ for a unit change in $x_1$ while holding $x_2$ constant.

Instead of the bivariate $r$ we may use a partial $r$.

Under these circumstances, the formula for $b_{y|x_2}$ is...
\[\beta_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2} \times \frac{sd_y}{sd_x}\] (6)

\[\beta_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2} \times \frac{sd_y}{sd_x}\] (7)

It is also easy to extend the bivariate intercept to the multivariate case as follows.

\[a = \bar{Y} - b_1 \bar{x_1} - b_2 \bar{x_2}\] (8)
Linear Multiple Regression

- Suppose that we have the following data set.

<table>
<thead>
<tr>
<th></th>
<th>y</th>
<th>x1</th>
<th>x2</th>
<th>x1 x2</th>
<th>x1sq</th>
<th>x1 cubed</th>
<th>x2sq</th>
<th>x2 cubed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>137.2</td>
<td>38.4</td>
<td>16</td>
<td>614.4</td>
<td>1474.56</td>
<td>56623.11</td>
<td>256</td>
<td>4096</td>
</tr>
<tr>
<td>2</td>
<td>146.4</td>
<td>41.3</td>
<td>16.5</td>
<td>691.45</td>
<td>1706.69</td>
<td>70444.99</td>
<td>272.25</td>
<td>4492.125</td>
</tr>
<tr>
<td>3</td>
<td>145.3</td>
<td>42.9</td>
<td>15.8</td>
<td>677.82</td>
<td>1840.41</td>
<td>78653.59</td>
<td>249.64</td>
<td>3944.312</td>
</tr>
<tr>
<td>4</td>
<td>166.5</td>
<td>52.3</td>
<td>16</td>
<td>836.8</td>
<td>2735.29</td>
<td>143055.7</td>
<td>256</td>
<td>4096</td>
</tr>
<tr>
<td>5</td>
<td>163.2</td>
<td>52</td>
<td>17.2</td>
<td>894.4</td>
<td>2704</td>
<td>140608</td>
<td>295.84</td>
<td>5088.449</td>
</tr>
<tr>
<td>6</td>
<td>164.4</td>
<td>45.2</td>
<td>16.8</td>
<td>759.36</td>
<td>2043.04</td>
<td>92345.41</td>
<td>282.24</td>
<td>4741.631</td>
</tr>
<tr>
<td>7</td>
<td>114</td>
<td>51.7</td>
<td>16.3</td>
<td>842.71</td>
<td>2672.89</td>
<td>138188.4</td>
<td>265.69</td>
<td>4330.747</td>
</tr>
<tr>
<td>8</td>
<td>161.1</td>
<td>52.5</td>
<td>17.8</td>
<td>994.4999</td>
<td>2756.25</td>
<td>144703.1</td>
<td>316.84</td>
<td>5639.751</td>
</tr>
<tr>
<td>9</td>
<td>181.6</td>
<td>46.9</td>
<td>17.3</td>
<td>811.37</td>
<td>2199.61</td>
<td>103161.7</td>
<td>299.29</td>
<td>5177.716</td>
</tr>
<tr>
<td>10</td>
<td>207.5</td>
<td>66.1</td>
<td>18.2</td>
<td>1203.02</td>
<td>4369.21</td>
<td>288804.8</td>
<td>331.24</td>
<td>6028.569</td>
</tr>
<tr>
<td>11</td>
<td>252.8</td>
<td>49.5</td>
<td>15.9</td>
<td>787.06</td>
<td>2450.26</td>
<td>121287.4</td>
<td>252.81</td>
<td>4019.679</td>
</tr>
<tr>
<td>12</td>
<td>154.6</td>
<td>47.8</td>
<td>16.3</td>
<td>779.14</td>
<td>2284.84</td>
<td>109215.3</td>
<td>265.69</td>
<td>4330.747</td>
</tr>
<tr>
<td>13</td>
<td>145.4</td>
<td>48.9</td>
<td>16.6</td>
<td>811.7401</td>
<td>2391.21</td>
<td>116930.2</td>
<td>276.56</td>
<td>4574.296</td>
</tr>
<tr>
<td>14</td>
<td>174.4</td>
<td>68.5</td>
<td>16.7</td>
<td>1143.95</td>
<td>4692.25</td>
<td>321419.1</td>
<td>278.89</td>
<td>4657.464</td>
</tr>
<tr>
<td>15</td>
<td>191.1</td>
<td>72.8</td>
<td>17.1</td>
<td>1244.88</td>
<td>5299.84</td>
<td>385882.4</td>
<td>292.41</td>
<td>5000.211</td>
</tr>
<tr>
<td>16</td>
<td>224.1</td>
<td>82.7</td>
<td>19.1</td>
<td>1579.57</td>
<td>6899.29</td>
<td>56609.3</td>
<td>364.81</td>
<td>6967.872</td>
</tr>
<tr>
<td>17</td>
<td>209.7</td>
<td>85.7</td>
<td>18.4</td>
<td>1576.08</td>
<td>7344.49</td>
<td>62942.8</td>
<td>338.56</td>
<td>6229.503</td>
</tr>
<tr>
<td>18</td>
<td>241.9</td>
<td>87.9</td>
<td>18.3</td>
<td>1608.57</td>
<td>7726.41</td>
<td>67915.1</td>
<td>334.89</td>
<td>6128.486</td>
</tr>
<tr>
<td>19</td>
<td>232</td>
<td>88.4</td>
<td>17.4</td>
<td>1538.16</td>
<td>7814.56</td>
<td>69080.7</td>
<td>302.76</td>
<td>5268.023</td>
</tr>
<tr>
<td>20</td>
<td>232.6</td>
<td>89.6</td>
<td>18.1</td>
<td>1621.76</td>
<td>8028.16</td>
<td>71923.1</td>
<td>327.61</td>
<td>5929.741</td>
</tr>
<tr>
<td>21</td>
<td>244.2</td>
<td>91.3</td>
<td>18.2</td>
<td>1661.66</td>
<td>8385.69</td>
<td>76104.6</td>
<td>331.24</td>
<td>6028.569</td>
</tr>
</tbody>
</table>
We now see that the significance levels reveal that $x_1$ and $x_2$ are both statistically significant. The $R^2$ and adjusted $R^2$ have not been significantly reduced, indicating that this model still fits well. Therefore, we leave the interaction term pruned from the model.

What are the assumptions of multiple linear regression analysis?
Regression modeling and the assumptions

1. What are the assumptions?
   1. linearity
   2. Heteroskedasticity
   3. No influential outliers in small samples
   4. No multicollinearity
   5. No autocorrelation of residuals
   6. Fixed independent variables-no measurement error
   7. Normality of residuals
Testing the model for mispecification and robustness

Linearity
  matrix graphs shown above
Multicollinearity
  vif
Misspecification tests
  heteroskedasticity tests
    rvfplot
    hettest
  residual autocorrelation tests
    corrgram
outlier detection
  tabulation of standardized residuals
  influence assessment
residual normality tests
  sktest
Specification tests (not covered in this lecture)
Misspecification tests

• We need to test the residuals for normality.
• We can save the residuals in STATA, by issuing a command that creates them, after we have run the regression command.
• The command to generate the residuals is
• predict resid, residuals
Generation of the regression residuals

```
predict resid, residuals
.list resid
```

```
 resid
  1.  .3539732
  2.  .6530005
  3.   3.78156
  4.  9.435602
  5. -4.666642
  6.  10.17057
  7. -15.00131
  8. -13.1132
  9.  20.21508
 10.  9.758573
 11.  .7449242
 12.  1.271493
 13. -12.33821
 14. -12.79413
 15. -6.084922
 16. -6.216054
 17. -18.42389
 18.  11.51263
 19.  9.314301
 20.  1.612983
 21.  9.803662
```
Generation of standardized residuals

• Predict rstd, rstandard

```
. predict rstd, rstandard
. list rstd

     rstd
1.    1.193339
2.   -.715803
3.    1.54602
4.    1.881816
5.  -.5420839
6.  -.9652568
7.  -.8337771
8.  -.4848421
9.  -.1.04553
10.    .436832
11.    .8842576
12.    .9685738
13.  -.4798729
14.  -.0174715
15.    .8092066
16.    .2993545
17.  -.6112104
18.  -.3151504
19.  -.2030219
20.    .4539774
21.  -.2.63822
```
Generation of studentized residuals

- Predict rstud, rstudent
Testing the Residuals for Normality

1. We use a Smirnov-Kolmogorov test.
2. The command for the test is: sktest resid

```
. sktest resid
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pr(Skewness)</th>
<th>Pr(Kurtosis)</th>
<th>adj chi2(2)</th>
<th>joint Prob&gt;chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>resid</td>
<td>0.837</td>
<td>0.370</td>
<td>0.91</td>
<td>0.6348</td>
</tr>
</tbody>
</table>

This tests the cumulative distribution of the residuals against that of the theoretical normal distribution with a chi-square test. To determine whether there is a statistically significant difference. The null hypothesis is that there is no difference. When the probability is less than .05, we must reject the null hypothesis and infer that the residuals are non-normally distributed.
Testing the Residuals for heteroskedasticity

1. We may graph the standardized or studentized residuals against the predicted scores to obtain a graphical indication of heteroskedasticity.

2. The Cook-Weisberg test is used to test the residuals for heteroskedasticity.
A Graphical test of heteroskedasticity: \texttt{rvfplot, border yline(0)}

This displays any problematic patterns that might suggest heteroskedasticity. But it doesn’t tell us which residuals are outliers.
Cook-Weisberg Test

\[ \text{Var}(e_i) = \sigma^2 \exp(zt) \]

where

\[ e_i = \text{error in regression model} \]

\[ z = x \hat{\beta} \text{ or variable list supplied by user} \]

The test is whether \( t = 0 \)

hettest estimates the model \( e_i^2 = \alpha + z_i t + \nu_i \)

it forms a score test \( S = \frac{\text{SS of model}}{2} \)

\( h0 : S_{df=p} \sim \chi^2 \) where \( p = \text{number of parameters} \)
Cook-Weisberg test syntax

1. The command for this test is: hettest resid

```
. hettest resid
Cook-Weisberg test for heteroskedasticity using variables specified
Ho: Constant variance
    chi2(1)    =    0.09
    Prob > chi2 =    0.7706
```

An insignificant result indicates lack of heteroskedasticity. That is, such a result indicates the presence of equal variance of the residuals along the predicted line. This condition is otherwise known as homoskedasticity.
Testing the residuals for Autocorrelation

1. One can use the command, `dwstat`, after the regression to obtain the Durbin-Watson d statistic to test for first-order autocorrelation.

2. There is a better way. Generate a casenum variable: `Gen casenum = _n`
Create a time dependent series

.list casenum

    casenum
  1.      1
  2.      2
  3.      3
  4.      4
  5.      5
  6.      6
  7.      7
  8.      8
  9.      9
 10.     10
 11.     11
 12.     12
 13.     13
 14.     14
 15.     15
 16.     16
 17.     17
 18.     18
 19.     19
 20.     20
 21.     21

.tsset casenum
time variable: casenum, 1 to 21
Run the Ljung-Box Q statistic which tests previous lags for autocorrelation and partial autocorrelation

The significance of the AC (Autocorrelation) and PAC (Partial autocorrelation) is shown in the Prob column. None of these residuals has any significant autocorrelation.

The STATA command is: corrgram resid

The image shows a table with columns labeled LAG, AC, PAC, Q, Prob>Q, [Autocorrelation], and [Partial Autocorrelation]. The table includes the following data:

<table>
<thead>
<tr>
<th>LAG</th>
<th>AC</th>
<th>PAC</th>
<th>Q</th>
<th>Prob&gt;Q</th>
<th>0</th>
<th>1</th>
<th>-1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1053</td>
<td>0.1103</td>
<td>0.26761</td>
<td>0.6049</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.1253</td>
<td>-0.1445</td>
<td>0.66703</td>
<td>0.7164</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.0749</td>
<td>0.1229</td>
<td>0.81748</td>
<td>0.8453</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.3449</td>
<td>-0.4979</td>
<td>4.1972</td>
<td>0.3800</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.2008</td>
<td>-0.1370</td>
<td>5.4149</td>
<td>0.3674</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.0637</td>
<td>-0.0657</td>
<td>5.5455</td>
<td>0.4760</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.1090</td>
<td>-0.1877</td>
<td>5.9551</td>
<td>0.5450</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.1116</td>
<td>-0.1975</td>
<td>6.4177</td>
<td>0.6005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
One can run Autoregression in the event of autocorrelation

This can be done with
newey y x1 x2 x3 lag(1) time
prais y x1 x2 x3
Outlier detection

• Outlier detection involves the determination whether the residual (error = predicted – actual) is an extreme negative or positive value.
• We may plot the residual versus the fitted plot to determine which errors are large, after running the regression.
• The command syntax was already demonstrated with the graph on page 16: rvfplot, border yline(0)
Create Standardized Residuals

- A standardized residual is one divided by its standard deviation.

\[
\text{resid}_{\text{standardized}} = \frac{\hat{y}_i - y_i}{s}
\]

where \( s = \text{std dev of residuals} \)
Standardized residuals

predict residstd, rstandard
list residstd
tabulate residstd

<table>
<thead>
<tr>
<th>Standardized residuals</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.805076</td>
<td>1</td>
<td>4.76</td>
<td>4.76</td>
</tr>
<tr>
<td>-1.426854</td>
<td>1</td>
<td>4.76</td>
<td>9.52</td>
</tr>
<tr>
<td>-1.399868</td>
<td>1</td>
<td>4.76</td>
<td>14.29</td>
</tr>
<tr>
<td>-1.228921</td>
<td>1</td>
<td>4.76</td>
<td>19.05</td>
</tr>
<tr>
<td>-1.163865</td>
<td>1</td>
<td>4.76</td>
<td>23.81</td>
</tr>
<tr>
<td>-0.6549696</td>
<td>1</td>
<td>4.76</td>
<td>28.57</td>
</tr>
<tr>
<td>-0.5814227</td>
<td>1</td>
<td>4.76</td>
<td>33.33</td>
</tr>
<tr>
<td>-0.4481118</td>
<td>1</td>
<td>4.76</td>
<td>38.10</td>
</tr>
<tr>
<td>0.0345273</td>
<td>1</td>
<td>4.76</td>
<td>42.86</td>
</tr>
<tr>
<td>0.0630611</td>
<td>1</td>
<td>4.76</td>
<td>47.62</td>
</tr>
<tr>
<td>0.0731241</td>
<td>1</td>
<td>4.76</td>
<td>52.38</td>
</tr>
<tr>
<td>0.1004419</td>
<td>1</td>
<td>4.76</td>
<td>57.14</td>
</tr>
<tr>
<td>0.1599981</td>
<td>1</td>
<td>4.76</td>
<td>61.90</td>
</tr>
<tr>
<td>0.372687</td>
<td>1</td>
<td>4.76</td>
<td>66.67</td>
</tr>
<tr>
<td>0.5229192</td>
<td>1</td>
<td>4.76</td>
<td>71.43</td>
</tr>
<tr>
<td>0.9662861</td>
<td>1</td>
<td>4.76</td>
<td>76.19</td>
</tr>
<tr>
<td>0.969979</td>
<td>1</td>
<td>4.76</td>
<td>80.96</td>
</tr>
<tr>
<td>0.9763196</td>
<td>1</td>
<td>4.76</td>
<td>85.71</td>
</tr>
<tr>
<td>0.9798224</td>
<td>1</td>
<td>4.76</td>
<td>90.48</td>
</tr>
<tr>
<td>1.13156</td>
<td>1</td>
<td>4.76</td>
<td>95.24</td>
</tr>
<tr>
<td>2.006145</td>
<td>1</td>
<td>4.76</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total 21 100.00
Limits of Standardized Residuals

If the standardized residuals have values in excess of 3.5 and -3.5, they are outliers.

If the absolute values are less than 3.5, as these are, then there are no outliers.

While outliers by themselves only distort mean prediction when the sample size is small enough, it is important to gauge the influence of outliers.
Outlier Influence

• Suppose we had a different data set with two outliers.
• We tabulate the standardized residuals and obtain the following output:
Outlier a does not distort the regression line but outlier b does.

Y = a + bx

Outlier a has bad leverage and outlier a does not.
In this data set, we have two outliers. One is negative and the other is positive.

<table>
<thead>
<tr>
<th>Standardized residuals</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6.658807</td>
<td>1</td>
<td>4.76</td>
<td>4.76</td>
</tr>
<tr>
<td>-1.816594</td>
<td>1</td>
<td>4.76</td>
<td>9.52</td>
</tr>
<tr>
<td>-1.068852</td>
<td>1</td>
<td>4.76</td>
<td>14.29</td>
</tr>
<tr>
<td>-.9231888</td>
<td>1</td>
<td>4.76</td>
<td>19.05</td>
</tr>
<tr>
<td>-.4742897</td>
<td>1</td>
<td>4.76</td>
<td>23.81</td>
</tr>
<tr>
<td>-.3774867</td>
<td>1</td>
<td>4.76</td>
<td>28.57</td>
</tr>
<tr>
<td>-.2336429</td>
<td>1</td>
<td>4.76</td>
<td>33.33</td>
</tr>
<tr>
<td>-.1958244</td>
<td>1</td>
<td>4.76</td>
<td>38.10</td>
</tr>
<tr>
<td>-.1440506</td>
<td>1</td>
<td>4.76</td>
<td>42.86</td>
</tr>
<tr>
<td>-.1213372</td>
<td>1</td>
<td>4.76</td>
<td>47.62</td>
</tr>
<tr>
<td>-.1061324</td>
<td>1</td>
<td>4.76</td>
<td>52.38</td>
</tr>
<tr>
<td>-.0915209</td>
<td>1</td>
<td>4.76</td>
<td>57.14</td>
</tr>
<tr>
<td>-.0190165</td>
<td>1</td>
<td>4.76</td>
<td>61.90</td>
</tr>
<tr>
<td>-.0183856</td>
<td>1</td>
<td>4.76</td>
<td>66.67</td>
</tr>
<tr>
<td>-.0131741</td>
<td>1</td>
<td>4.76</td>
<td>71.43</td>
</tr>
<tr>
<td>.0905697</td>
<td>1</td>
<td>4.76</td>
<td>76.19</td>
</tr>
<tr>
<td>.3354265</td>
<td>1</td>
<td>4.76</td>
<td>80.95</td>
</tr>
<tr>
<td>.3532699</td>
<td>1</td>
<td>4.76</td>
<td>85.71</td>
</tr>
<tr>
<td>.3931204</td>
<td>1</td>
<td>4.76</td>
<td>90.48</td>
</tr>
<tr>
<td>.5398999</td>
<td>1</td>
<td>4.76</td>
<td>95.24</td>
</tr>
<tr>
<td>4.038815</td>
<td>1</td>
<td>4.76</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
Studentized Residuals

• Alternatively, we could form studentized residuals. These are distributed as a t distribution with df=n-p-1, though they are not quite independent. Therefore, we can approximately determine if they are statistically significant or not.

• Belsley et al. (1980) recommended the use of studentized residuals.
Studentized Residual

\[ e_i^s = \frac{e_i}{\sqrt{s^2_{(i)}(1 - h_i)}} \]

where

\[ e_i^s = \text{studentized residual} \]
\[ s_{(i)} = \text{standard deviation where ith obs is deleted} \]
\[ h_i = \text{leverage statistic} \]

These are useful in estimating the statistical significance of a particular observation, of which a dummy variable indicator is formed. The t value of the studentized residual will indicate whether or not that observation is a significant outlier.
The command to generate studentized residuals, called rstudt is:
\text{predict rstudt, rstudent}
Influence of Outliers

1. Leverage is measured by the diagonal components of the hat matrix.

2. The hat matrix comes from the formula for the regression of $Y$.

$$\hat{Y} = X \beta = X'(X'X)^{-1}X'Y$$

where $X'(X'X)^{-1}X' =$ the hat matrix, $H$

Therefore,

$$\hat{Y} = HY$$
Leverage and the Hat matrix

1. The hat matrix transforms Y into the predicted scores.
2. The diagonals of the hat matrix indicate which values will be outliers or not.
3. The diagonals are therefore measures of leverage.
4. Leverage is bounded by two limits: 1/n and 1. The closer the leverage is to unity, the more leverage the value has.
5. The trace of the hat matrix = the number of variables in the model.
6. When the leverage > 2p/n then there is high leverage according to Belsley et al. (1980) cited in Long, J.F. Modern Methods of Data Analysis (p.262). For smaller samples, Vellman and Welsch (1981) suggested that 3p/n is the criterion.
Cook’s D

1. Another measure of influence.
2. This is a popular one. The formula for it is:

$$Cook's \ D_i = \left( \frac{1}{p} \right) \left( \frac{h_i}{1-h_i} \right) \left( \frac{e_i^2}{s^2(1-h_i)} \right)$$

Cook and Weisberg(1982) suggested that values of D that exceeded 50% of the F distribution (df = p, n-p) are large.
Using Cook’s D in STATA

• Predict cook, cooksd
• Finding the influential outliers
• List cook, if cook > 4/n
• Belsley suggests 4/(n-k-1) as a cutoff

```
predict cook, cooksd
list cook if cook > 4/21

<table>
<thead>
<tr>
<th>cook</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.</td>
</tr>
<tr>
<td>1.022088</td>
</tr>
<tr>
<td>16.</td>
</tr>
<tr>
<td>5.713506</td>
</tr>
</tbody>
</table>

list y cook if cook > 4/21

<table>
<thead>
<tr>
<th>y</th>
<th>cook</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.</td>
<td>388</td>
</tr>
<tr>
<td></td>
<td>1.022088</td>
</tr>
<tr>
<td>16.</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>5.713506</td>
</tr>
</tbody>
</table>
```
Graphical Exploration of Outlier Influence

- Graph `cook residstd, xlab ylab`

The two influential outliers can be found easily here in the upper right.
DFbeta

• One can use the DFbetas to ascertain the magnitude of influence that an observation has on a particular parameter estimate if that observation is deleted.

\[ DFbeta_j = \frac{b_j - b_{(i)j} u_j}{\sqrt{\sum u_j^2 (1 - h_j)}} \]

where \( u_j = \text{residuals of regression of } x \text{ on remaining } x_s \).
## Obtaining DFbetas in STATA

```
. regress y x1 x2
Number of obs = 21
F(  2,  18) = 89.64
Prob > F    = 0.0000
R-squared   = 0.9088
Adj R-squared = 0.8986
Root MSE    = 3.2865

Source | SS      | df | MS
--------|---------|----|----
Model   | 1880.44276 | 2  | 940.22138
Residual| 188.79334  | 18 | 10.4886297
Total   | 2069.2381  | 20 | 103.461905

`y` | Coef.  | Std. Err. |  t   | P>|t|  | [95% Conf. Interval] |
-----|--------|-----------|------|------|------------------------|
   x1 | 671154  | .126691   | 5.30 | 0.000 | .4049864               | .9373225  |
   x2 | 1.295851| .9574854  | -1.32| 0.206 | .5292931               | 2.06741   |
_cons| -50.35884| 5.138328  | -9.80| 0.000 | -61.15407              | -39.56361 |
```

. predict dfbx1, dfbeta(x1)
. predict dfbx2, dfbeta(x2)
. list id dfbx1 dfbx2

<table>
<thead>
<tr>
<th>id</th>
<th>dfbx1</th>
<th>dfbx2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.395095</td>
<td>0.1159482</td>
</tr>
<tr>
<td>2</td>
<td>-0.132735</td>
<td>-0.0392102</td>
</tr>
<tr>
<td>3</td>
<td>0.4057819</td>
<td>-0.0047129</td>
</tr>
<tr>
<td>4</td>
<td>-4.25597</td>
<td>-0.61115</td>
</tr>
<tr>
<td>5</td>
<td>-0.010055</td>
<td>-0.006179</td>
</tr>
<tr>
<td>6</td>
<td>0.109435</td>
<td>-0.1711058</td>
</tr>
<tr>
<td>7</td>
<td>0.2337774</td>
<td>-0.3354144</td>
</tr>
<tr>
<td>8</td>
<td>0.1608564</td>
<td>-0.2307903</td>
</tr>
<tr>
<td>9</td>
<td>0.3192272</td>
<td>-0.3511808</td>
</tr>
<tr>
<td>10</td>
<td>0.132594</td>
<td>-0.2043691</td>
</tr>
<tr>
<td>11</td>
<td>-0.132594</td>
<td>-2.043691</td>
</tr>
<tr>
<td>12</td>
<td>0.2485714</td>
<td>-0.361683</td>
</tr>
<tr>
<td>13</td>
<td>-0.0547803</td>
<td>0.0844714</td>
</tr>
<tr>
<td>14</td>
<td>-0.0350302</td>
<td>0.06301</td>
</tr>
<tr>
<td>15</td>
<td>0.0674575</td>
<td>-0.016317</td>
</tr>
<tr>
<td>16</td>
<td>-0.0218951</td>
<td>-0.0052961</td>
</tr>
<tr>
<td>17</td>
<td>0.0143292</td>
<td>0.0052403</td>
</tr>
<tr>
<td>18</td>
<td>-0.0143292</td>
<td>-0.0052403</td>
</tr>
<tr>
<td>19</td>
<td>0.0116224</td>
<td>-0.0072811</td>
</tr>
<tr>
<td>20</td>
<td>-0.0457675</td>
<td>0.0067213</td>
</tr>
<tr>
<td>21</td>
<td>-1.923631</td>
<td>1.701042</td>
</tr>
</tbody>
</table>
Robust statistical options when assumptions are violated

1. Nonlinearity
   1. Transformation to linearity
   2. Nonlinear regression

2. Influential Outliers
   1. Robust regression with robust weight functions
   2. rreg y x1 x2

3. Heteroskedasticity of residuals
   1. Regression with Huber/White/Sandwich variance-covariance estimators
   2. Regress y x1 x2, robust

4. Residual autocorrelation correction
   1. Autoregression with prais y x1 x2, robust
   2. newey-west regression

5. Nonnormality of residuals
   1. Quantile regression: qreg y x1 x2
   2. Bootstrapping the regression coefficients
Nonlinearity: Transformations to linearity

1. When the equation is not intrinsically nonlinear, the dependent variable or independent variable may be transformed to effect a linearization of the relationship.

2. Semi-log, translog, Box-Cox, or power transformations may be used for these purposes.

1. Boxcox regression permits determines the optimal parameters for many of these transformations.
Fix for Nonlinear functional form: Nonlinear Regression Analysis

Examples of 2 exponential growth curve models, the first of which we estimate with our data.

\[ \text{nl exp2 } y \ x \]

\textit{estimates} \quad Y = b_1 b_2^x

\[ \text{nl exp3 } y \ x \]

\textit{estimates} \quad y = b_0 + b_1 b_2^x
Nonlinear Regression in Stata

- `. nl exp2 y x`
- `(obs = 15)`

- Iteration 0: residual SS = 56.08297
- Iteration 1: residual SS = 49.46372
- Iteration 2: residual SS = 49.4593
- Iteration 3: residual SS = 49.4593

- Source SS df MS Number of obs = 15
- F( 2, 13) = 1585.01
- Model 12060.5407 2 6030.27035 Prob > F = 0.0000
- Residual 49.4592999 13 3.80456153 R-squared = 0.9959
- Adj R-squared = 0.9953
- Total 12110 15 807.333333 Root MSE = 1.950529
- Res. dev. = 60.46465
- 2-param. exp. growth curve, \( y = b_1 b_2^x \)

- \( y \quad \text{Coef.} \quad \text{Std. Err.} \quad \text{t} \quad \text{P}>t \quad [95\% \text{ Conf. Interval}] \)

- \( b_1 \quad 58.60656 \quad 1.472156 \quad 39.81 \quad 0.000 \quad 55.42616 \quad 61.78696 \)
- \( b_2 \quad .9611869 \quad .0016449 \quad 584.36 \quad 0.000 \quad .9576334 \quad .9647404 \)

- (SE's, P values, CI's, and correlations are asymptotic approximations)

- `.`
Heteroskedasticity correction

1. Prof. Halbert White showed that heteroskedasticity could be handled in a regression with a heteroskedasticity-consistent covariance matrix estimator (Davidson & McKinnon (1993), Estimation and Inference in Econometrics, Oxford U Press, p. 552).

2. This variance-covariance matrix under ordinary least squares is shown on the next page.
OLS Covariance Matrix Estimator

\[(X'X)^{-1}(X'\Sigma X)(X'X)^{-1}\]

where \(\Sigma = s_t^2 / (X'X)\)
White’s HAC estimator

1. White’s estimator is for large samples.

2. White’s heteroskedasticity-corrected variance and standard errors can be larger or smaller than the OLS variances and standard errors.
Heteroskedastically consistent covariance matrix “Sandwich” estimator (H. White)

\[ n^{-1} (X'X)^{-1} (n^{-1} X'\Omega X)(n^{-1} X'X)^{-1} \]

where \( \Omega = \frac{e_t^2}{1 - h_t^2} \)

However, there are different versions:

- \( HC_0 : \Omega = e_t^2 \)
- \( HC_1 : \Omega = \frac{n}{n-k} e_t^2 \)
- \( HC_2 : \Omega = \frac{e_t^2}{1 - h_t} \)
- \( HC_3 : \Omega = \frac{e_t^2}{(1 - h_t)^2} \)
Regression with robust standard errors for heteroskedasticity

Regress y x1 x2, robust

```
. hettest resid

Cook-Weisberg test for heteroskedasticity using variables specified
    Ho: Constant variance
    chi2(1)  =  3580.55
    Prob > chi2 =  0.0000

. regress y x1 x2, robust

Regression with robust standard errors

|     | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-----|--------|-----------|-------|------|---------------------|
| y   |        |           |       |      |                     |
|     |        |           |       |      |                     |
| x1  | 5.904387 | 2.871935  | 2.06  | 0.055 | -.129324             | 11.9381   |
| x2  | 28.97995 | 36.26385  | 0.80  | 0.435 | -47.20757            | 105.1675  |
| _cons | -609.3122 | 558.5422  | -1.09 | 0.290 | -1782.766           | 564.1413  |

Options other than robust, are hc2 and hc3 referring to the versions mentioned by Davidson and McKinnon above.
Robust options for the VCV matrix in Stata

• Regress y x1 x2, hc2
• Regress y x1 x2, hc3

• These correspond to the Davidson and McKinnon’s versions of the heteroskedastically consistent vcv options 2 and 3.
Problems with Autoregressive Errors

1. Problems in estimation with OLS
   1. When there is first-order autocorrelation of the residuals,
   2. \( e_t = \rho_1 e_{t-1} + v_t \)

2. Effect on the Variance
   1. \( e_t^2 = \rho_1^2 e_{t-1}^2 + v_t^2 \)

\[
\begin{align*}
E(e_t e_t) &= E(\rho e_{t-1} + v_t)(\rho e_{t-1} + v_{t-1}) \\
\sigma_e^2 &= \rho^2 \sigma_e^2 + \sigma_v^2 \\
\sigma_v^2 &= (1 - \rho^2) \sigma_e^2, \\
\end{align*}
\]

where

\( \sigma_e^2 \) = apparent (uncorrected autocorrelated) error variance
\( \sigma_v^2 \) = actual identically, independently distributed error variance.
Sources of Autocorrelation

1. Lagged endogenous variables
2. Misspecification of the model
3. Simultaneity, feedback, or reciprocal relationships
4. Seasonality or trend in the model
Prais-Winston Transformation-cont’d

\[ e_t^2 = \frac{v_t^2}{(1-\rho^2)}, \text{ therefore } e_t = \frac{v_t}{\sqrt{(1-\rho^2)}} \]

It follows that

\[ Y_t = a + bx_t + \frac{v_t}{\sqrt{(1-\rho^2)}} \]

\[ \sqrt{(1-\rho^2)} Y_t = \sqrt{(1-\rho^2)} a + \sqrt{(1-\rho^2)} bx_t + v_t \]

\[ \therefore \quad Y_t^* = a^* + bx_t^* + v_t \]
Autocorrelation of the residuals: prais & newey regression

To test whether the variable is autocorrelated

- `tsset time`
- `corrgram y`
- `prais y x1 x2, robust`
- `newey y x1 x2, lag(1) t(time)`
Testing for autocorrelation of residuals

```
regress mna10 l5sumprc
predict resid10, residual
corrgram resid10
```
Prais-Winston Regression for AR(1) errors

Using the robust option here guarantees that the White heteroskedasticity consistent sandwich variance-covariance estimator will be used in the autoregression procedure.

```
. prais mna12 l5,sumproc, robust
Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.4768
Iteration 2:  rho = 0.4953
Iteration 3:  rho = 0.4959
Iteration 4:  rho = 0.4960
Iteration 5:  rho = 0.4960

Prais-Winston AR(1) regression -- iterated estimates

Regression with robust standard errors

Number of obs =     86
F(  2,    84) =   23.50
Prob > F      =  0.0000
R-squared     =  0.0708
Root MSE      =  5.228

<table>
<thead>
<tr>
<th></th>
<th>Semi-robust Coef.</th>
<th></th>
<th></th>
<th></th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>mna12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coef.</td>
<td>Std. Err.</td>
<td>t</td>
<td>P&gt;</td>
<td>t</td>
</tr>
<tr>
<td>sumproc</td>
<td>1.400647</td>
<td>0.5589943</td>
<td>2.51</td>
<td>0.014</td>
<td>0.2890256</td>
</tr>
<tr>
<td></td>
<td>5.768363</td>
<td>1.178935</td>
<td>4.89</td>
<td>0.000</td>
<td>3.423921</td>
</tr>
<tr>
<td>_cons</td>
<td>0.4959529</td>
<td></td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rho</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Durbin-Watson statistic (original) 1.041513
Durbin-Watson statistic (transformed) 1.897883
```
Newey-West Robust Standard errors

• An autocorrelation correction is added to the meat or tofu in the White Sandwich estimator by Newey-West.

\[ n^{-1} (X' X)^{-1} (n^{-1} X' \Omega X)(n^{-1} X' X)^{-1} \]

where \( \Omega = \frac{e_t^2}{1-h_t^2} \)

However, there are different versions:

\( HC_0 : \Omega = e_t^2 \)

\( HC_1 : \Omega = \frac{n}{n-k} e_t^2 \)

\( HC_2 : \Omega = \frac{e_t^2}{1-h_t} \)

\( HC_3 : \Omega = \frac{e_t^2}{(1-h_t)^2} \)
Central Part of Newey-West Sandwich estimator

\[ X' \hat{\Omega} X_{\text{newey-west}} \]

\[ = X' \hat{\Omega} X_{\text{white}} + \frac{n}{n-k} \sum_{l=1}^{m} \left(1 - \frac{l}{m+1}\right) e_i e_{i-1} (x_i' x_{i-1} + x_{i-1}' x_i) \]

where \( k = \text{number of predictors} \)

\( l = \text{timelag} \)

\( m = \text{maximum timelag} \)
Newey-West Robust Standard errors

Newey West standard errors are robust to autocorrelation and heteroskedasticity with time series regression models.

```
newey mna10 l1mna10 l5sumpro, lag(1) t(time)

Regression with Newey-West standard errors
maximum lag : 1

Number of obs = 86
F( 2, 83) = 21.72
Prob > F = 0.0000

|      | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------|-------|-----------|-------|-----|----------------------|
| mna10| .6482148 | .0987128 | 6.57  | 0.000 | .4518791 .8445505 |
| l1mna10 | 1.071262 | .6551474 | 1.64  | 0.106 | -.2319955 2.374324 |
| l5sumpro | 3.327793 | 1.364233 | 2.44  | 0.017 | .6143872 6.041198 |
```
Assume OLS regression

- We regress y on x1 x2 x3
- We obtain the following output

```
. regress y x1 x2 x3

Source | SS df MS
----------|--------|------
Model | 1890.40813  3  630.136045
Residual | 178.829962  17  10.5194095
Total | 2069.2381  20  103.461905

Number of obs = 21
F( 3, 17) =  59.90
Prob > F =  0.0000
R-squared =  0.9136
Adj R-squared =  0.8963
Root MSE =  3.2434

y | Coef. Std. Err.  t  P>|t| [95% Conf. Interval]
----------|-----------|--------|-----|-------------------|
x1 | 0.7156402  0.1348682  5.31  0.000 0.4311143  1.000166
x2 | 1.295206  0.3660243  3.52  0.003 0.5182280  2.071749
x3 | -0.152225  0.156294 -0.97  0.344 -0.4910741  0.176291
_cons | -39.91967  11.096 -3.56  0.004 -65.01803 -14.82132
```

Next we examine the residuals
Residual Assessment

The data set is too small to drop case 21, so I use robust regression.
Robust regression algorithm: \texttt{rreg}

1. A regression is performed and absolute residuals are computed.
   \[ r_i = | y_i - x_i b | \]

2. These residuals are computed and scaled:
   \[ u_i = \frac{r_i}{s} \]
   \[ = \frac{y_i - x_i b}{s} \]
Scaling the residuals

\[ s = \frac{M}{0.6745} \]

where

\[ M = \text{med}(| r_i - \text{med}(r_i) |) \]

The residuals are scaled by the median absolute value of the median residual.
Essential Algorithm

• The estimator of the parameter $b$ minimizes the sum of a less rapidly increasing function of the residuals (SAS Institute, The Robustreg Procedure, draft copy, p.3505, forthcoming):

$$Q(b) = \sum_{i=1}^{n} \rho \left( \frac{r_i}{\sigma} \right)$$

where $r_i = y - x_i b$

$\sigma$ is estimated by $s$
1. If this were OLS, the $\rho$ would be a quadratic function.

2. If we can ascertain $s$, we can by taking the derivatives with respect to $b$, find a first order solution to

$$
\sum_{i=1}^{n} \psi \left( \frac{r_i}{s} \right) x_{ij} = 0,
$$

where $j = 1, \ldots, p$

$$
\psi = \rho'
$$
Case weights are developed from weight functions

1. Case weights are formed based on those residuals.

2. Weight functions for those case weights are first the Huber weights and then the Tukey bisquare weights:

3. A weighted regression is rerun with the case weights.
Iteratively reweighted least squares

- The case weight \( w(x) \) is defined as:

\[
w(x) = \frac{\psi(x)}{x}
\]

It is updated at each iteration until it converges on a value and the change from iteration to iteration declines below a criterion.
Weights functions for reducing outlier influence

\[ W(z,c) = \begin{cases} 
1, & \text{if } |z| < c \\
\frac{c}{|z|}, & \text{otherwise} 
\end{cases} \]

c is the tuning constant used in determining the case weights. For the Huber weights \( c = 1.345 \) by default.
Weight Functions

Tukey biweight (bisquare)

\[
W(x, c) = \begin{cases} 
(1 - \left(\frac{x}{c}\right)^2)^2 & \text{if } |x| < c \\
0 & \text{otherwise}
\end{cases}
\]

\[\text{C is also the biweight tuning constant. C is set at 4.685 for the biweight.}\]
Tuning Constants

• When the residuals are normally distributed and the tuning constants are set at the default, they give the procedure about 95% of the efficiency of OLS.

• The tuning constants may be adjusted to provide downweighting of the outliers at the expense of Gaussian efficiency.

• Higher tuning constants cause the estimator to more closely approximate OLS.
Robust Regression algorithm – cont’d

3. WLS regression is performed using those case weights

4. Iterations case when case weights drop below a tolerance level

5. Weights are based initially on Huber weights. Then Beaton and Tukey biweights are used.

6. Caveat: M estimation is not that robust with regard to leverage points.
Robust Regression for down-weighting outliers

- `rreg y x1 x2 x3`

Uses Huber and Tukey biweights to downweight the influence of outliers in the estimation of the mean of y in the upper panel whereas ols regression is given in the lower panel.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1890.40813</td>
<td>3</td>
<td>630.136045</td>
</tr>
<tr>
<td>Residual</td>
<td>178.829962</td>
<td>17</td>
<td>10.5194095</td>
</tr>
<tr>
<td>Total</td>
<td>2069.2381</td>
<td>20</td>
<td>103.461905</td>
</tr>
</tbody>
</table>

| y       | Coef. | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|---------|-------|-----------|------|-----|---------------------|
| x1      | 0.852651 | 0.1228395 | 6.97 | 0.000 | 0.594446 - 1.108858 |
| x2      | 0.873394 | 0.339811  | 2.61 | 0.018 | 0.168721 - 1.577998 |
| x3      | -1.224349 | 0.1418364 | -8.66 | 0.000 | -2.416836 - 0.179999 |
| _cons   | -41.6703 | 10.79559  | -3.86 | 0.001 | -64.447 - 18.89361  |

| y       | Coef. | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|---------|-------|-----------|------|-----|---------------------|
| x1      | 0.715602 | 0.1348582 | 5.31 | 0.000 | 0.431143 - 1.000166 |
| x2      | 1.295286 | 0.3680243 | 3.52 | 0.008 | 0.5188228 - 2.071749 |
| x3      | -1.521225 | 0.156294  | -0.97 | 0.344 | -0.4818741 - 0.1776291 |
| _cons   | -39.91967 | 11.896    | -3.36 | 0.004 | -65.01803 - 14.82132 |

Number of obs = 21
F( 3, 17) = 74.15
Prob > F = 0.0000

Robust regression estimates
A Corrective Option for Nonnormality of the Residuals

1. Quantile regression (median regression is the default) is one option.

2. Algorithm
   1. Minimizes the sum of the absolute residuals
   2. The residual in this case is the value minus the unconditional median.
   3. This produces a formula that predicts the median of the dependent variable

\[ Y_{med} = a + bx \]
Quantile Regression

qreg in STATA estimates least absolute value (LAV or MAD or L1 norm regression).

The algorithm minimizes the sum of the absolute deviations about the median.

The formula generated estimates the median rather than the mean, as rreg does.

\[ Y_{\text{median}} = \text{constant} + bx \]
Median regression

```
. qreg y x1 x2
Iteration 1:  WLS sum of weighted deviations =  646.79574
Iteration 1: sum of abs. weighted deviations =  632.51404
Iteration 2: sum of abs. weighted deviations =  630.18984
Iteration 3: sum of abs. weighted deviations =  630.04748

Median regression
Raw sum of deviations   969.2 (about 164.39999)
Min sum of deviations   630.0475
Number of obs = 21
Pseudo R2 = 0.3499

|    | Coef. | Std. Err. |    t | P>|t| | 95% Conf. Interval |
|----|-------|-----------|------|------|-------------------|
| y  |       |           |      |      |                   |
| x1 | 1.82036 | 0.4807429 | 3.79 | 0.001 | 0.8103565 - 2.830363 |
| x2 | 1.06053 | 9.564848  | 0.11 | 0.913 | -1.903447 - 21.15553 |
| _cons | 50.30016 | 141.5986 | 0.36 | 0.727 | -247.1875 - 347.7878 |
```
Bootstrapping

- Bootstrapping may be used to obtain empirical regression coefficients, standard errors, confidence intervals, etc. when the distribution is non-normal.
- Bootstrapping may be applied to qreg with bsqreg
Bootstrapping quantile or median regression standard errors

- `qreg y x1 x2 x3`
- `bsqreg y x1 x2 x3, reps(1000)`

| y     | Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|-------|--------|-----------|-------|-------|----------------------|
| x1    | .831894| .1305396  | 6.37  | 0.000 | .556495 - 1.107299   |
| x2    | .573913| .3895329  | 1.50  | 0.153 | -.2352706 - 1.383097 |
| x3    | -.060966| .1540719  | -0.40 | 0.698 | -.3859329 - .2641938 |
| _cons | -39.6895| 11.76248  | -3.37 | 0.004 | -64.50651 - 14.8732  |

- `bsqreg y x1 x2 x3, reps(1000)`

| y     | Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|-------|--------|-----------|-------|-------|----------------------|
| x1    | .831894| .2211178  | 3.76  | 0.002 | .3653663 - 1.298402  |
| x2    | .573913| .5978836  | 0.96  | 0.350 | -1.887002 - 1.885126 |
| x3    | -.060966| .1724506  | -0.35 | 0.728 | -.4247085 - .3029694 |
| _cons | -59.6895| 12.18281  | -4.96 | 0.006 | -55.39333 - 13.98637 |
Methods of Model Validation

• These methods may be necessary where the sampling distributions of the parameters of interest are nonnormal or unknown.
• Bootstrapping
• Cross-validation
• Data-splitting
Bootstrapping

• When the distribution of the residuals is nonnormal or the distribution is unknown, bootstrapping can provide proper regression coefficients, standard errors, and confidence intervals.
Stata Bootstrapping Syntax

- Bs “regress y x1 x2 x3”, “_b[x1] _b[x2] _b[x3]”, reps(1000) saving(mybstrap1)
Internal Validation
$R^2$ and adjusted $R^2$

1. Plot $\hat{Y}$ against $Y$. Compute an $R^2$ and an adjusted $R^2$. 
Cross-validation

- Jacknifing
- This is repeated sampling, where one group or observation is left out.
- The analysis is reiterated and the results are averaged to obtain a validation.
Resampling

1. Bootstrapping was performed developed by Efron. Resampling generally needs to be done at least B=100 times.

2. Resampling with replacement is performed on a sample. From each bootstrapped sample, a mean is computed. The average of all of these b bootstrapped means is the mean.

3. The bootstrapped means are used to compute a bootstrapped variance estimate. If b is the number of bootstraps, then b is the n used in the computation. A bootstrapped variance estimate is now known.

4. After enough resampling, an empirical distribution function is formed.
Bootstrapped Formulae

$$\bar{x}^b = \sum_{n} x_i^b / n$$

$$Var(x)^b = \sum_{b=1}^{B} (\bar{x}^b - \text{avg}(\bar{x}^b))^2 / (B - 1)$$
Data-splitting

1. Sample Splitting
   1. Subset the sample into a training and a validation subsample. One has to be careful about the tail wagging the dog, as David Reilly is wont to say.
   2. This results in poorer accuracy and loss of power unless there is plenty of data.
   3. Tests for parameter constancy
Comparison of STATA, SAS, and S-PLUS

Stata has rreg, qreg, bsqreg
  Rreg is M estimation with Huber and Tukey bisquare weight functions
  qreg is quantile regression
  Bsqreg is bootstrapped quantile regression
  Bootstrapping

SAS has M, Least Trimmed squares, S, and MM estimation in Proc Robustreg in version 9. It can perform Robust ANOVA as well. SAS has 10 different weight functions that may be applied. It does not have bootstrapping

SPLUS has a robust library of procedures. Among the procedures it can apply are robust regression, robust ANOVA, robust principal components analysis, robust covariance matrix estimation, robust discriminant function analysis, robust distribution estimation for asymmetric distributions. SPLUS has procedures to run OLS regression side by side with robust MM regression to show the differences. It has a wide variety of graphical diagnostics as well.