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.

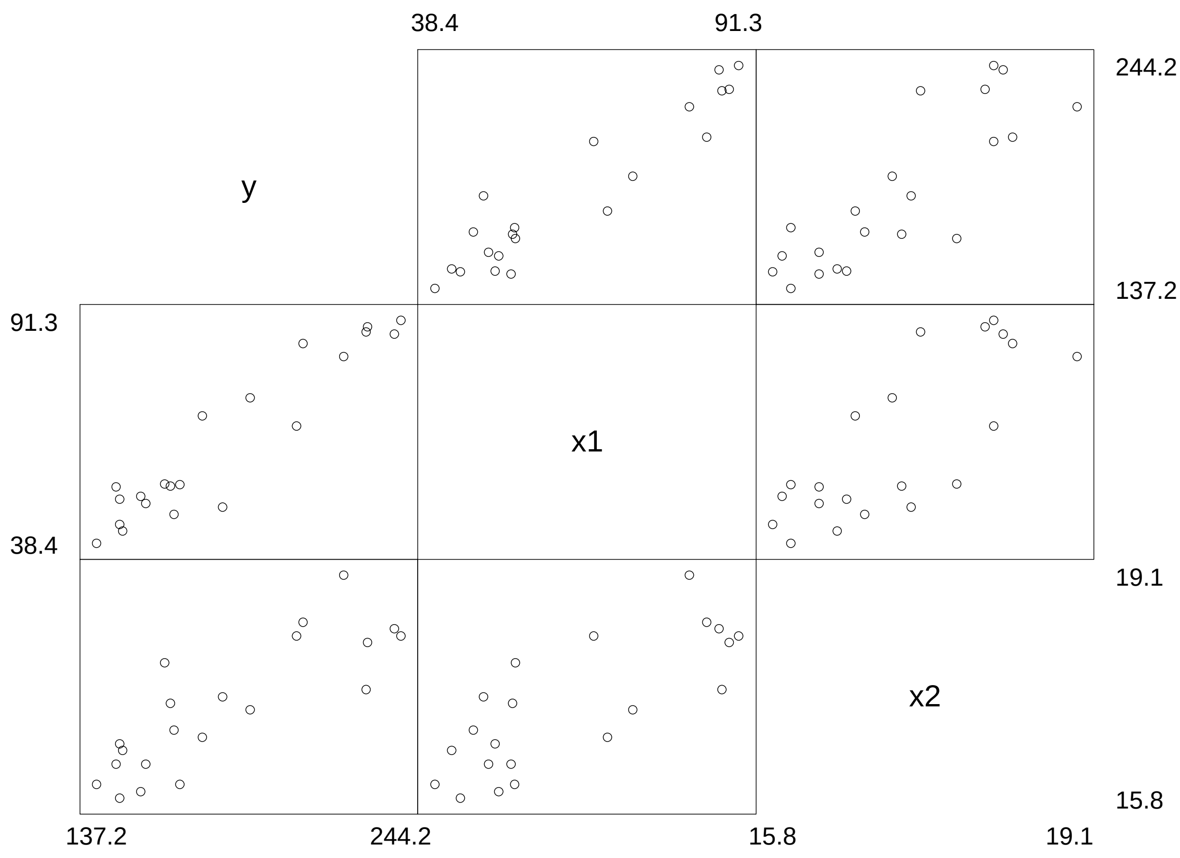
2. Strictly speaking:
   - Threats to stability and reliability include
     - influential outliers
     - Influential outliers played havoc with statistical estimation. Since 1960, many robust techniques of estimation have developed that have been resistant to the effects of such outliers.
       - 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 heterskedasticity
        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

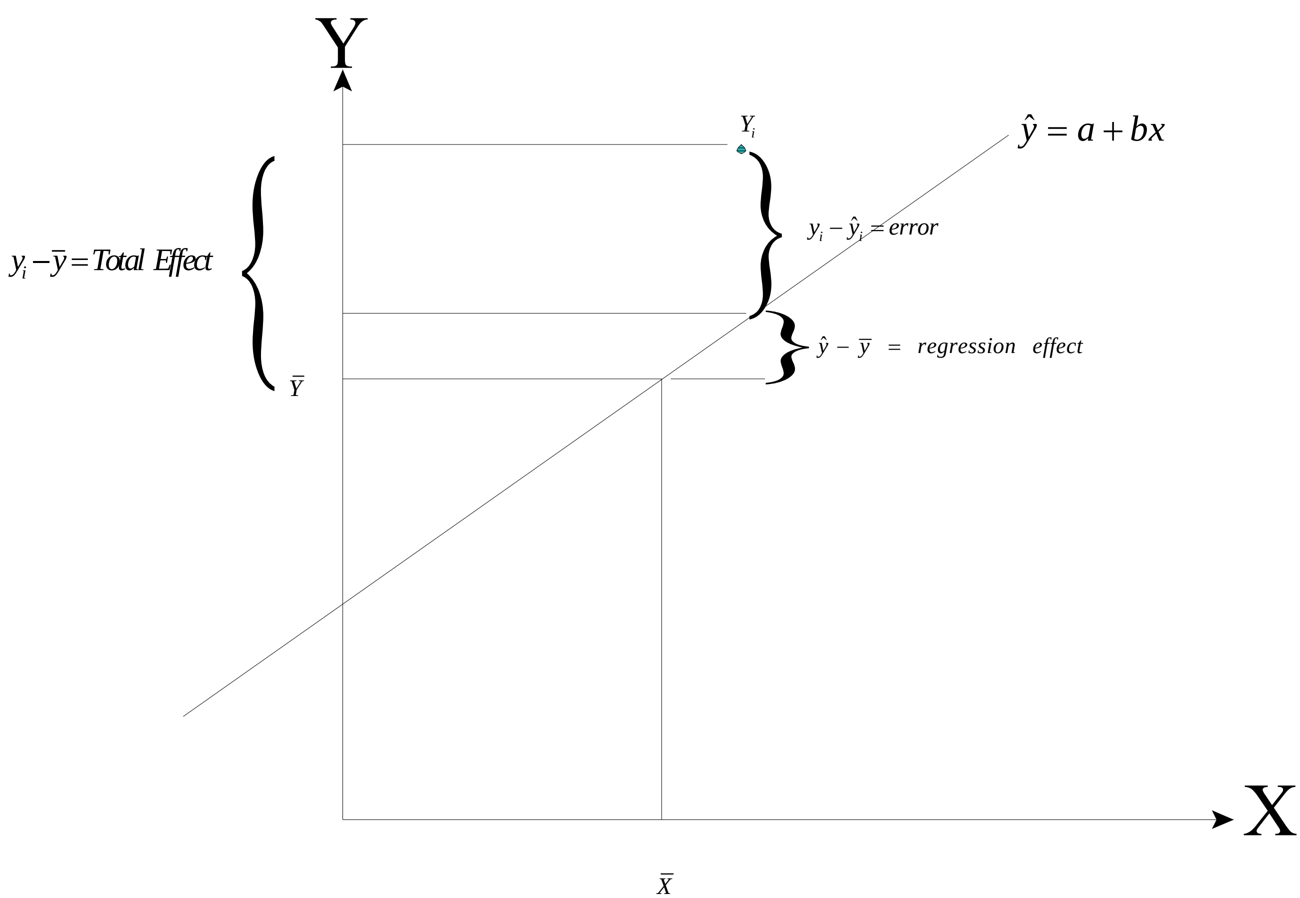
a is the intercept

b is the regression coefficient

x is the predictor variable

Graphical Decomposition of Effects

Decomposition of Effects



Derivation of the Intercept

$$y = a + bx + e$$

$$e = y - a - bx$$

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

$$\text{Because by definition} \sum_{i=1}^{n} e_i = 0$$

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

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

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

$$a = \bar{y} - b\bar{x}$$

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:

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

$$\begin{aligned} &where \\ &x = x_i - \overline{x} \\ &y = y_i - \overline{y} \end{aligned}$$

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

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 $b_{y.12}$

$$\beta_{yx_1.x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}} * \frac{sd_y}{sd_x} \qquad (6)$$

$$\beta_{yx_2.x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r^2_{x_1x_2}} * \frac{sd_y}{sd_x} \qquad (7)$$

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

$$a = \overline{Y} - b_1\overline{x}_1 - b_2\overline{x}_2 \qquad (8)$$

Linear Multiple Regression

- Suppose that we have the following data set.

```
. list y x1 x2 x1x2 x1sq x1cubed x2sq x2cubed
```

	y	x1	x2	x1x2	x1sq	x1cubed	x2sq	x2cubed
1.	137.2	38.4	16	614.4	1474.56	56623.11	256	4096
2.	146.4	41.3	16.5	681.45	1705.69	70444.99	272.25	4492.125
3.	145.3	42.9	15.8	677.82	1840.41	78953.59	249.64	3944.312
4.	166.5	52.3	16	836.8	2735.29	143055.7	256	4096
5.	163.2	52	17.2	894.4	2704	140608	295.84	5088.449
6.	164.4	45.2	16.8	759.36	2043.04	92345.41	282.24	4741.631
7.	144	51.7	16.3	842.71	2672.89	138188.4	265.69	4330.747
8.	161.1	52.5	17.8	934.4999	2756.25	144703.1	316.84	5639.751
9.	181.6	46.9	17.3	811.37	2199.61	103161.7	299.29	5177.716
10.	207.5	66.1	18.2	1203.02	4369.21	288804.8	331.24	6028.569
11.	152.8	49.5	15.9	787.05	2450.25	121287.4	252.81	4019.679
12.	154.6	47.8	16.3	779.14	2284.84	109215.3	265.69	4330.747
13.	145.4	48.9	16.6	811.7401	2391.21	116930.2	275.56	4574.296
14.	174.4	68.5	16.7	1143.95	4692.25	321419.1	278.89	4657.464
15.	191.1	72.8	17.1	1244.88	5299.84	385828.4	292.41	5000.211
16.	224.1	82.7	19.1	1579.57	6839.29	565609.3	364.81	6967.872
17.	209.7	85.7	18.4	1576.88	7344.489	629422.8	338.56	6229.503
18.	241.9	87.9	18.3	1608.57	7726.41	679151.5	334.89	6128.486
19.	232	88.4	17.4	1538.16	7814.56	690807.1	302.76	5268.023
20.	232.6	89.6	18.1	1621.76	8028.16	719323.1	327.61	5929.741
21.	244.2	91.3	18.2	1661.66	8335.69	761048.6	331.24	6028.569

Stata OLS regression model syntax

```
. regress y x1 x2

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  2,    18) =   99.10
       Model |  24015.2826     2  12007.6413           Prob > F      =  0.0000
    Residual |  2180.92749    18  121.162638           R-squared     =  0.9167
-------------+------------------------------           Adj R-squared =  0.9075
       Total |  26196.2101    20   1309.8105           Root MSE      =  11.007

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    1.45456   .2117818     6.87   0.000     1.009623    1.899497
          x2 |   9.365501   4.063958     2.30   0.033     .8274414    17.90356
       _cons |  -68.85708   60.01695    -1.15   0.266     -194.948    57.23386
------------------------------------------------------------------------------

.
```

We now see that the significance levels reveal that x1 and x2 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.   1.271483
 13.  -12.33821
 14.  -12.78413
 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.  -.1531504
 19.  -.2030219
 20.    .4539774
 21.    -2.63822

.
```

Generation of studentized residuals

- Predict rstud, rstudent

```
. predict rstud, rstudent

. list rstud

         rstud
  1.   1.209475
  2.  -.7051386
  3.   1.617904
  4.   2.051797
  5.  -.5305036
  6.  -.9632038
  7.  -.8259467
  8.  -.4736521
  9.  -1.048586
 10.    .426188
 11.    .878292
 12.   .9667067
 13.  -.4687306
 14.    -.01695
 15.   .8006164
 16.    .291185
 17.  -.5995858
 18.  -.1486803
 19.  -.1971994
 20.    .443117
 21.  -3.330493

.
```

Testing the Residuals for Normality

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

sktest resid

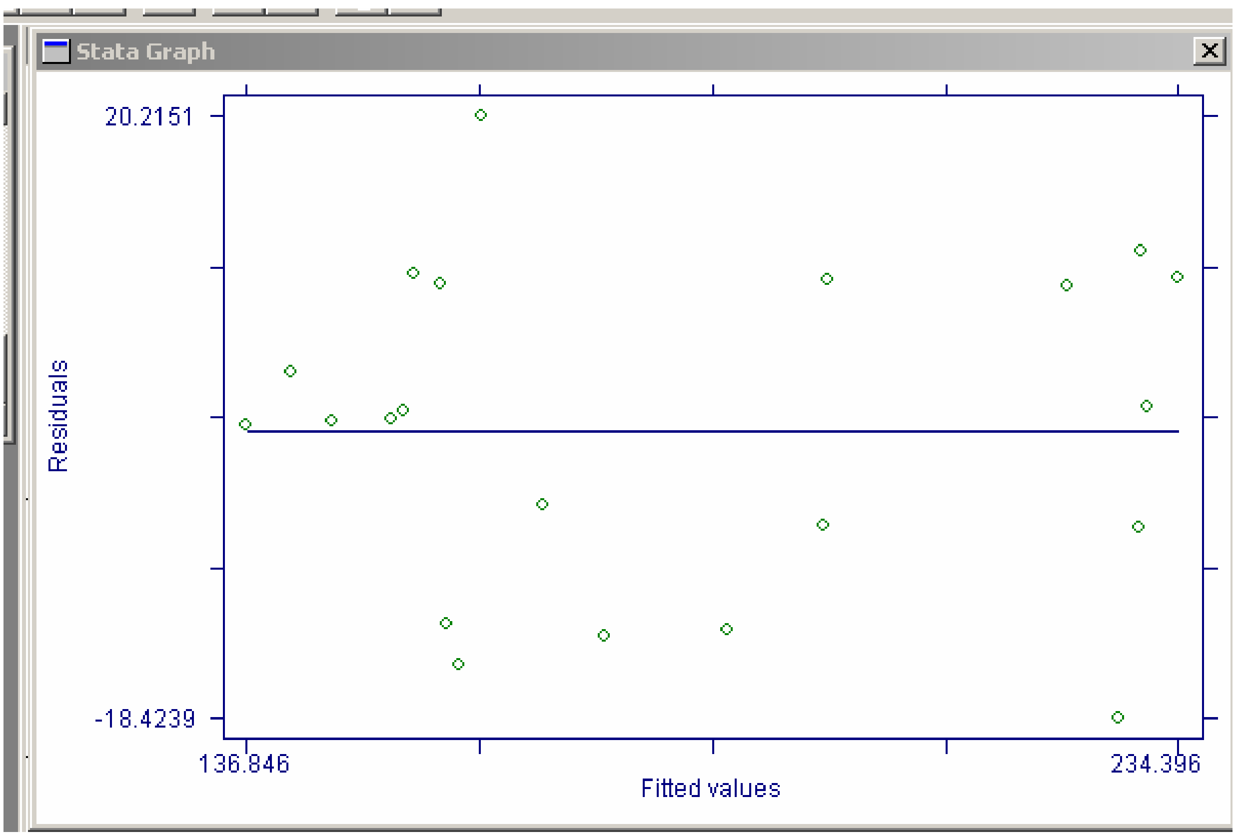
```
. sktest resid

                   Skewness/Kurtosis tests for Normality
                                                        ------- joint ------
    Variable | Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)     Prob>chi2
-------------+-------------------------------------------------------
       resid |     0.837          0.370           0.91        0.6348
```

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: 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

$$Var(e_i) = \sigma^2 \exp(zt)$$

where

$e_i = \text{error in regression model}$

$z = x\hat{\beta}$ *or variable list supplied by user*

The test is whether $t = \mathbf{0}$

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

it forms a score test $S = \dfrac{SS\ of\ model}{\mathbf{2}}$

$h\mathbf{0}: S_{df=p} \sim \chi\mathbf{2}$ *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, an 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 STATA command is :   corrgram resid

```
. corrgram resid
```

LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	0.1053	0.1103	.26761	0.6049		
2	-0.1253	-0.1445	.66703	0.7164		
3	0.0749	0.1229	.81748	0.8453		
4	-0.3449	-0.4979	4.1972	0.3800		
5	-0.2008	-0.1370	5.4149	0.3674		
6	0.0637	-0.0657	5.5455	0.4760		
7	-0.1090	-0.1877	5.9551	0.5450		
8	-0.1116	-0.1975	6.4177	0.6005		

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

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.

$$resid_{standardized} = \frac{\hat{y}_i - y_i}{s}$$

$$where\ s = std\ dev\ of\ residuals$$

Standardized residuals

predict residstd, rstandard
list residstd
tabulate residstd

```
. tabulate residstd
```

Standardized residuals	Freq.	Percent	Cum.
-1.805076	1	4.76	4.76
-1.426554	1	4.76	9.52
-1.339868	1	4.76	14.29
-1.23931	1	4.76	19.05
-1.163855	1	4.76	23.81
-.6649696	1	4.76	28.57
-.5814227	1	4.76	33.33
-.4448118	1	4.76	38.10
.0345273	1	4.76	42.86
.0630611	1	4.76	47.62
.0731241	1	4.76	52.38
.1208419	1	4.76	57.14
.1599861	1	4.76	61.90
.3712687	1	4.76	66.67
.9229192	1	4.76	71.43
.9662861	1	4.76	76.19
.969979	1	4.76	80.95
.9763196	1	4.76	85.71
.9798224	1	4.76	90.48
1.13156	1	4.76	95.24
2.006145	1	4.76	100.00
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.

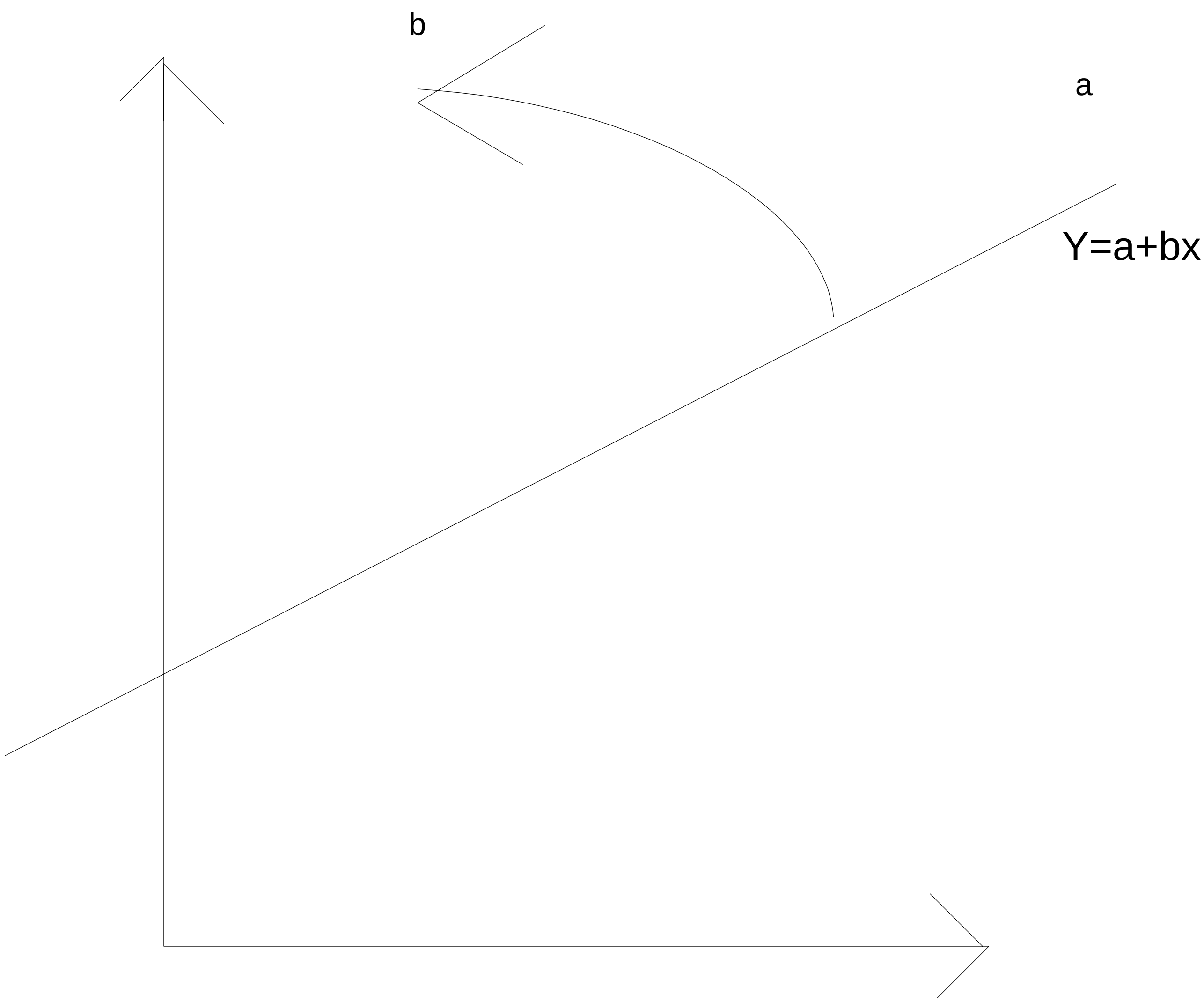
If the absolute values are less than 3.5, as these are, then there are no outliers

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.



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.

```
. tabulate residstd
```

Standardized residuals	Freq.	Percent	Cum.
-6.658807	1	4.76	4.76
-1.816594	1	4.76	9.52
-1.068852	1	4.76	14.29
-.9231888	1	4.76	19.05
-.4742897	1	4.76	23.81
-.3774867	1	4.76	28.57
-.2336429	1	4.76	33.33
-.1958244	1	4.76	38.10
-.1440506	1	4.76	42.86
-.1213372	1	4.76	47.62
-.1061324	1	4.76	52.38
-.0915209	1	4.76	57.14
-.0190165	1	4.76	61.90
-.0183856	1	4.76	66.67
-.0131741	1	4.76	71.43
.0905697	1	4.76	76.19
.3354265	1	4.76	80.95
.3532699	1	4.76	85.71
.3931204	1	4.76	90.48
.539899	1	4.76	95.24
4.038815	1	4.76	100.00
Total	21	100.00	

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

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

$where$

$$e_i^{\,s} = studentized\ residual$$

$$s_{(i)} = standard\ deviation\ where\ ith\ obs\ is\ deleted$$

$$h_i = leverage\ statistic$$

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:

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.

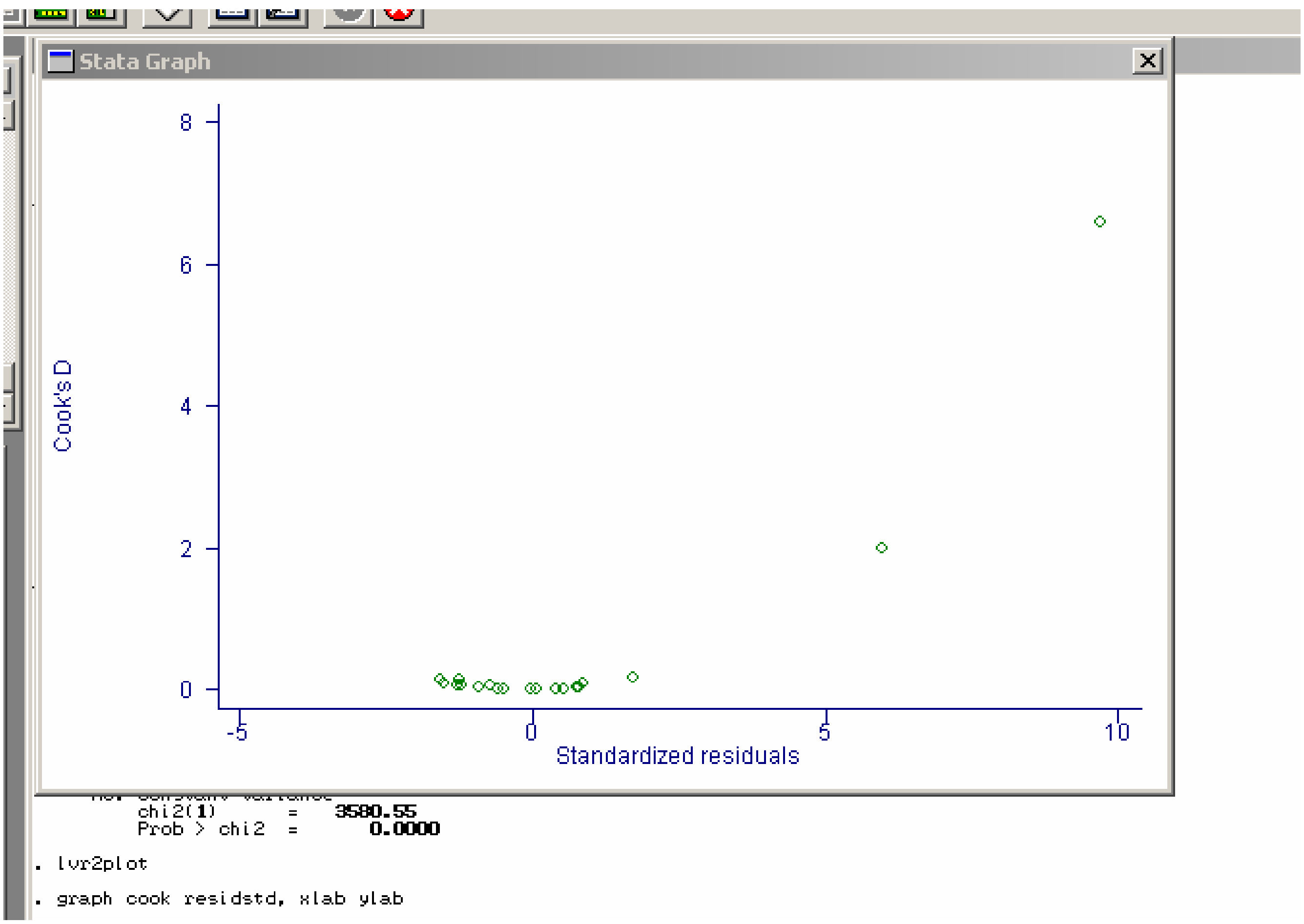
Using Cook’s D in STATA

- 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

            cook
 10.    1.022088
 16.    5.713506

. list y cook if cook > 4/21

               y        cook
 10.         388    1.022088
 16.          20    5.713506

.
```

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.

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

$$where \ \ u_j = residuals \, of$$

$$regression \, of \ \ x \ on \ \ remaining \ \ xs.$$

Obtaining DFbetas in STATA

```
. regress y x1 x2

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  2,    18) =   89.64
       Model |  1880.44276     2  940.221381           Prob > F      =  0.0000
    Residual |  188.795334    18  10.4886297           R-squared     =  0.9088
-------------+------------------------------           Adj R-squared =  0.8986
       Total |   2069.2381    20  103.461905           Root MSE      =  3.2386

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .6711544    .126691     5.30   0.000     .4049864    .9373225
          x2 |   1.295351   .3674854     3.52   0.002     .5232931     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

            id       dfbx1       dfbx2
  1.         1    .3925095    .1159482
  2.         2   -.132735   -.0392102
  3.         3    .4057813   -.0047129
  4.         4   -.4259597      .61115
  5.         5     .010505   -.0305179
  6.         6    .1094532   -.1711058
  7.         7    .2337774   -.3354144
  8.         8    .1608564   -.2307903
  9.         9    .3192272   -.3511808
 10.        10    .1325348   -.2043691
 11.        11    .1325348   -.2043691
 12.        12    .2485714    -.361683
 13.        13   -.0547803    .0844714
 14.        14   -.0350302      .06301
 15.        15   -.0674575    -.016317
 16.        16   -.0218951   -.0052961
 17.        17   -.0143292    .0052403
 18.        18   -.0143292    .0052403
 19.        19    .0116234   -.0072811
 20.        20   -.0457675    .0067213
 21.        21   -1.923631    1.701042
```

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.

$$nl \;\; exp2 \; y \;\; x$$

$$estimates \;\; Y = b_1 b_2^{\;x}$$

$$nl \;\; exp3 \;\; y \; x$$

$$estimates \;\; y = b_0 + b_1 b_2^{\;x}$$

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=b1*b2^x
-
- y Coef. Std. Err. t P>t [95% Conf. Interval]
-
- b1 58.60656 1.472156 39.81 0.000 55.42616 61.78696
- b2 .9611869 .0016449 584.36 0.000 .9576334 .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), <u>Estimation and Inference in Econometrics</u>, Oxford U Press, p. 552).
2. This variance-covariance matrix under ordinary least squares is shown on the next page.

OLS Covariance Matrix Estimator

$$(X'X)^{-1}(X'\Sigma X)(X'X)^{-1}$$

$$where \quad \Sigma = s_t^{2}/(X'X)$$

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)

Bread | Meat(tofu) | Bread

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

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

However, there are different versions:

$$HC_0 : \Omega = e_t^2$$

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

$$HC2 : \Omega = \frac{e_t^2}{1-h_t}$$

$$HC3 : \Omega = \frac{e_t^2}{(1-h_t)^2}$$

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                 Number of obs =      21
                                                       F(  2,    18) =    4.79
                                                       Prob > F      =  0.0215
                                                       R-squared     =  0.4841
                                                       Root MSE      =  144.80

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          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. $e_t^2 = \rho_1^2 e_{t-1}^2 + v_t^2$

$$
\begin{aligned}
E(e_t e_t) &= E(\rho e_{t-1} + v_t)(\rho e_{t-1} + v_{t-1}) \\
\sigma_e^2 &= \rho^2 \sigma_e^2 + \sigma_v^2 \\
\sigma_v^2 &= (1 - \rho^2)\sigma_e^2 ,
\end{aligned}
\qquad (10.15)
$$

where

σ_e^2 = *apparent (uncorrected autocorrelated) error variance*
σ_v^2 = *actual identically, independently distributed error variance.*

58

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

$$e_t^{\,2} = \frac{v_t^{\,2}}{(1-\rho^2)} \;,\quad therefore \;\; e_t = \frac{v_t}{\sqrt{(1-\rho^2)}}$$

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

```
(5 missing values generated)

. corrgram resid10

                                          -1       0       1 -1       0       1
 LAG       AC       PAC       Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.6572    0.6700   38.452  0.0000
2        0.3805   -0.0775   51.496  0.0000
3        0.3162    0.1778   60.614  0.0000
4        0.2927    0.0762    68.52  0.0000
5        0.2416   -0.0096   73.973  0.0000
6        0.1653   -0.0707   76.557  0.0000
7        0.0712   -0.1136   77.042  0.0000
8        0.0768    0.1568   77.614  0.0000
9        0.0654   -0.0771   78.035  0.0000
10       0.0250   -0.0402   78.098  0.0000
11      -0.0262   -0.0562   78.167  0.0000
12      -0.0408   -0.0091   78.337  0.0000
13      -0.0221    0.0492   78.387  0.0000
14      -0.0152   -0.0290   78.412  0.0000
```

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.sumprc, 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-Winsten 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     =  0.0703
                                                       Root MSE      =   5.228

------------------------------------------------------------------------------
             |             Semi-robust
mna12        |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sumprc       |
          L5 |   1.400647   .5589943     2.51   0.014     .2890256    2.512269
_cons        |   5.768363   1.178935     4.89   0.000     3.423921    8.112804
-------------+----------------------------------------------------------------
         rho |   .4959529
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.041513
Durbin-Watson statistic (transformed) 1.897833

.
```

Newey-West Robust Standard errors

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

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

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

$$However, there\ are\ different\ versions:$$

$$HC_0 : \Omega = e_t^{2}$$

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

$$HC2 : \Omega = \frac{e_t^{2}}{1-h_t}$$

$$HC3 : \Omega = \frac{e_t^{2}}{(1-h_t)^{2}}$$

Central Part of Newey-West Sandwich estimator

$$X'\hat{\Omega}X_{newey-west}$$

$$= X'\hat{\Omega}X_{white}$$

$$+ \frac{n}{n-k}\sum_{l=1}^{m}\left(1-\frac{l}{m+1}\right)e_i e_{i-1}\left(x_i' x_{i-1} + x_{i-1}' x_i\right)$$

$$where\ k = number\ of\ predictors$$

$$l = time\, lag$$

$$m = maximum\ time\, lag$$

Newey-West Robust Standard errors

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

```
.  newey mna10 l1mna10 l5sumprc, lag(1) t(time)

Regression with Newey-West standard errors          Number of obs  =        86
maximum lag : 1                                     F(  2,    83)  =     21.72
                                                    Prob > F       =    0.0000
```

mna10	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
l1mna10	.6482148	.0987128	6.57	0.000	.4518791	.8445505
l5sumprc	1.071262	.6551474	1.64	0.106	-.2317995	2.374324
_cons	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              Number of obs =      21
-------------+------------------------------           F(  3,    17) =   59.90
       Model |  1890.40813     3  630.136045           Prob > F      =  0.0000
    Residual |  178.829962    17  10.5194095           R-squared     =  0.9136
-------------+------------------------------           Adj R-squared =  0.8983
       Total |   2069.2381    20  103.461905           Root MSE      =  3.2434

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .7156402   .1348582     5.31   0.000     .4311143    1.000166
          x2 |   1.295286   .3680243     3.52   0.003     .5188228    2.071749
          x3 |  -.1521225    .156294    -0.97   0.344    -.4818741    .1776291
       _cons |  -39.91967     11.896    -3.36   0.004    -65.01803   -14.82132
------------------------------------------------------------------------------
```

Next we examine the residuals

Residual Assessment

```
. lvr2plot

. predict rstud, rstudent

. predict lev, hat

. predict cook, cooksd

. tabulate rstud
```

Studentized residuals	Freq.	Percent	Cum.
-3.330493	1	4.76	4.76
-1.048586	1	4.76	9.52
-.9632038	1	4.76	14.29
-.8259467	1	4.76	19.05
-.7051386	1	4.76	23.81
-.5995858	1	4.76	28.57
-.5305036	1	4.76	33.33
-.4736521	1	4.76	38.10
-.4687306	1	4.76	42.86
-.1971994	1	4.76	47.62
-.1486803	1	4.76	52.38
-.01695	1	4.76	57.14
.291185	1	4.76	61.90
.426188	1	4.76	66.67
.443117	1	4.76	71.43
.8006164	1	4.76	76.19
.878292	1	4.76	80.95
.9667067	1	4.76	85.71
1.209475	1	4.76	90.48
1.617904	1	4.76	95.24
2.051797	1	4.76	100.00
Total	21	100.00	

```
. gen id=1

. replace id=_n
(20 real changes made)

. list id cook rstud if cook > 12/21

              id        cook       rstud
 21.          21    .6919999   -3.330493

.
```

The data set is to small to drop case 21, so I use robust regression

Robust regression algorithm: 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

$$s = \frac{M}{\boldsymbol{0.6745}}$$

where

$$M = med(|\, r_i - med(r_i)\,|)$$

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):

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

$$\text{where } r_i = y - x_i b$$

$$\sigma \text{ is estimated by } s$$

Essential algorithm-cont'd

1. If this were OLS, the ρ 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

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

$$where \quad j = 1, \ldots, p$$

$$\psi = \rho'$$

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

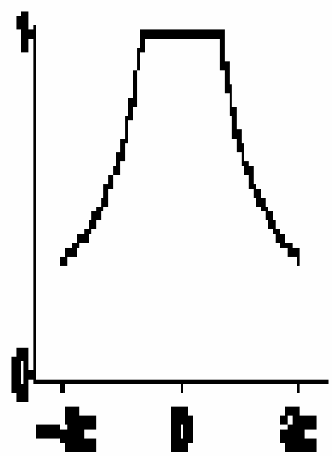
- The case weight w(x) is defined as:

$$w(x) = \frac{\psi(x)}{x}$$

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

huber $$W(x,c) = \begin{cases} 1 & \text{if } |x| < c \\ \frac{c}{|x|} & \text{otherwise} \end{cases}$$

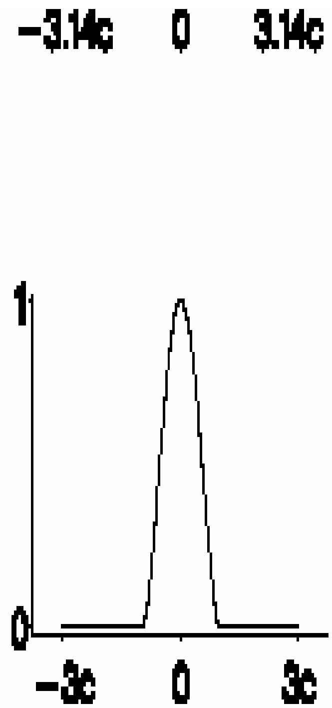


c is the tuning constant used in determining the case weights.
For the Huber weights c = 1.345 by default.

Weight Functions

Tukey biweight (bisquare)

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



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.

```
. rreg y x1 x2 x3

   Huber iteration 1:  maximum difference in weights = .48402478
   Huber iteration 2:  maximum difference in weights = .07083248
   Huber iteration 3:  maximum difference in weights = .03630349
Biweight iteration 4:  maximum difference in weights = .2114744
Biweight iteration 5:  maximum difference in weights = .04709559
Biweight iteration 6:  maximum difference in weights = .01648123
Biweight iteration 7:  maximum difference in weights = .01050023
Biweight iteration 8:  maximum difference in weights = .0027233

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

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .8526511    .1223835     6.97   0.000     .5944446    1.110858
          x2 |   .8733594    .3339811     2.61   0.018      .168721    1.577998
          x3 |  -.1224349    .1418364    -0.86   0.400    -.4216836    .1768139
       _cons |   -41.6703    10.79559    -3.86   0.001      -64.447   -18.89361
------------------------------------------------------------------------------

. reg y x1 x2 x3

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  3,    17) =   59.90
       Model |  1890.40813     3  630.136045           Prob > F      =  0.0000
    Residual |  178.829962    17  10.5194095           R-squared     =  0.9136
-------------+------------------------------           Adj R-squared =  0.8983
       Total |   2069.2381    20  103.461905           Root MSE      =  3.2434

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .7156402    .1348582     5.31   0.000     .4311143    1.000166
          x2 |   1.295286    .3680243     3.52   0.003     .5188228    2.071749
          x3 |  -.1521225     .156294    -0.97   0.344    -.4818741    .1776291
       _cons |  -39.91967      11.896    -3.36   0.004    -65.01803   -14.82132
------------------------------------------------------------------------------
```

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.

$$Y_{median} = \text{constant} + bx$$

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

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	1.82036	.4807429	3.79	0.001	.8103565	2.830363
x2	1.06053	9.564848	0.11	0.913	-19.03447	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)

```
. qreg y x1 x2 x3
Iteration  1:  WLS sum of weighted deviations =  49.699024

Iteration  1: sum of abs. weighted deviations =       56.5
Iteration  2: sum of abs. weighted deviations =  45.713592
Iteration  3: sum of abs. weighted deviations =  44.571208
Iteration  4: sum of abs. weighted deviations =  43.744836
Iteration  5: sum of abs. weighted deviations =   42.11057
Iteration  6: sum of abs. weighted deviations =  42.081159

Median regression                                    Number of obs =        21
  Raw sum of deviations       145 (about 15)
  Min sum of deviations 42.08116                     Pseudo R2     =    0.7098

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    .831884   .1305396     6.37   0.000     .5564695    1.107299
          x2 |    .573913   .3835329     1.50   0.153    -.2352706    1.383097
          x3 |  -.0608696   .1540719    -0.40   0.698    -.3859329    .2641938
       _cons |  -39.68985   11.76248    -3.37   0.004    -64.50651    -14.8732
------------------------------------------------------------------------------

. bsqreg y x1 x2 x3, reps(1000)
(estimating base model)
(bootstrapping ..........................................................
> ...........................................................................
> ...........................................................................
> ...........................................................................
> ...........................................................................
> ...........................................................................
> ...........................................................................
> ...........................................................................

Median regression, bootstrap(1000) SEs               Number of obs =        21
  Raw sum of deviations       145 (about 15)
  Min sum of deviations 42.08116                     Pseudo R2     =    0.7098

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    .831884   .2211178     3.76   0.002     .3653663    1.298402
          x2 |    .573913   .5977836     0.96   0.350    -.6873002    1.835126
          x3 |  -.0608696   .1724506    -0.35   0.728    -.4247085    .3029694
       _cons |  -39.68985   12.18281    -3.26   0.005    -65.39333   -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) saveing(mybstrap1)

```
. bs "regress y x1 x2 x3"  "_b[x1] _b[x2] _b[x3]", reps(1000)

command:      regress y x1 x2 x3
statistics:   _b[x1] _b[x2] _b[x3]
(obs=21)

Bootstrap statistics

Variable |   Reps  Observed       Bias   Std. Err.   [95% Conf. Interval]
---------+----------------------------------------------------------------
     bs1 |   1000  .7156402   .0168524   .1762298    .3698172  1.061463  (N)
         |                                           .3832314  1.064108  (P)
         |                                           .3592274  1.045956 (BC)
---------+----------------------------------------------------------------
     bs2 |   1000  1.295286  -.0626064   .4774686    .3583298  2.232243  (N)
         |                                           .3119411  2.190439  (P)
         |                                           .4578098  2.343391 (BC)
---------+----------------------------------------------------------------
     bs3 |   1000 -.1521225  -.0026936   .1229401    -.393373   .089128  (N)
         |                                          -.4155004  .0709161  (P)
         |                                          -.4453675  .0612494 (BC)
--------------------------------------------------------------------------
                             N = normal, P = percentile, BC = bias-corrected

.
```

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 bootrapped variance estimate is now known.
4. After enough resampling, an empirical distribution function is formed.

Bootstrapped Formulae

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

$$Var(x)^{b} = \sum_{b=1}^{B} (\overline{x}^{b} - avg(\overline{x}^{b})^{2} / (B - 1))$$

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.