MA397 – Outlier, Influential Point, and Residual Correlation Detection in Stata

Goals

We will see how to use Stata's built-in commands to detect outliers and influential points. We will also see an introduction to time series data in Stata.

Data

For the first part of this exercise we will be using the *birthwt.dta* dataset found on the course webpage at http://www.colby.edu/personal/l/lobrien/ma397.html.

Detecting Outliers

Of course, when looking for outliers, you should always generate residual-vs-fitted value, and residual-vs-predictor plots. Beyond that, Stata will calculate both standardized residuals and Studentized (jackknifed) residuals.

Let's first generate an ID number for each infant in our birthweight data so that we can identify them on plots more easily. Issue the command:

gen id = _n

This created a variable "id" that contains a unique number for each subject. Run the regression using head circumference and length. Now we can generate the standardized residuals by typing:

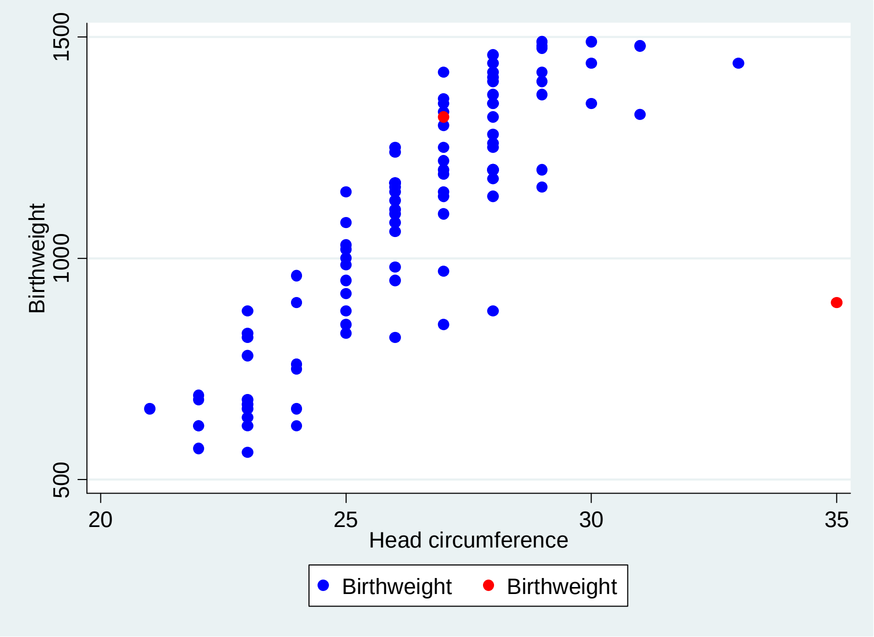
```
Quiety regress birthwt headcirc length
predict stanresid, rstandard
list id stanresid if abs(stanresid) > 3
```

We see that there are two observations that are outliers by this criterion:

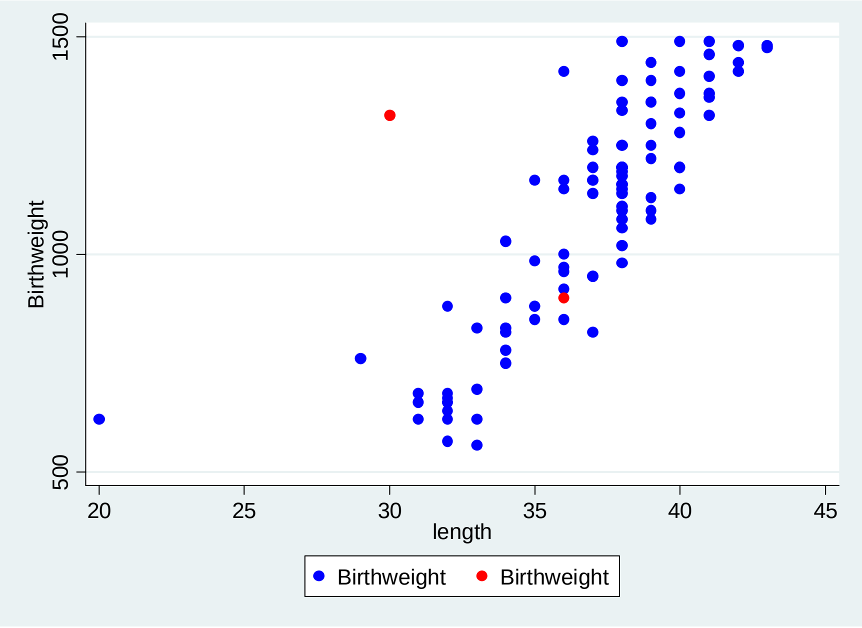
```
. list id  stanresid if abs(stanresid) > 3

     +----------------+
     | id   stanresid |
     |----------------|
  9. |  9    3.580978 |
 31. | 31   -5.007583 |
     +----------------+
```

Let's flag the outliers so that we can generate plots and label them, or generate separate subplots using different colors, to help them stand out.

```
. gen outlier = 1 if abs(stanresid) > 3
(98 missing values generated)

. replace outlier =0 if abs(stanresid) <= 3
(98 real changes made)
```



We can see the obvious outlier due to head circumference. The other outlier is due to an outlying value with respect to length.



Detecting Influential Points

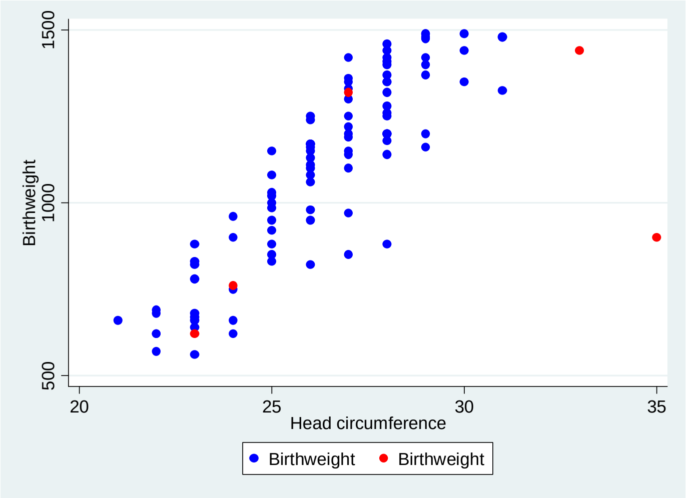
Stata has a built-in command for calculating the leverage of each observation. One you issue it, you can then get a summary of the values to determine the average leverage. To obtain these values and their summary statistics, type:

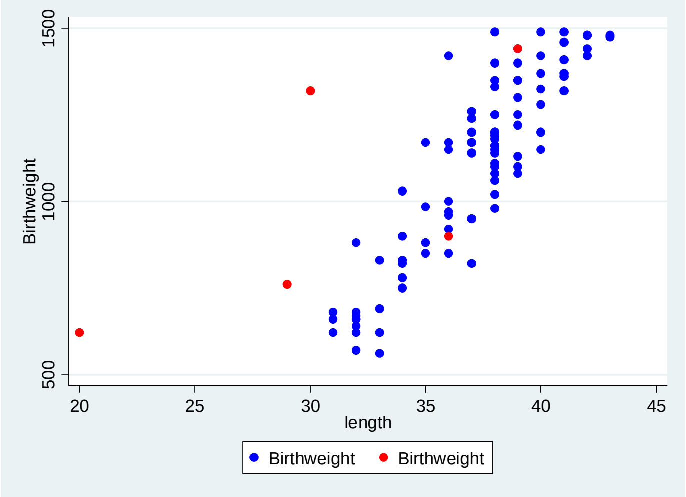
```
. predict h, lev

. summ h
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           h |       100         .03    .0410121   .0105366   .3156862
```

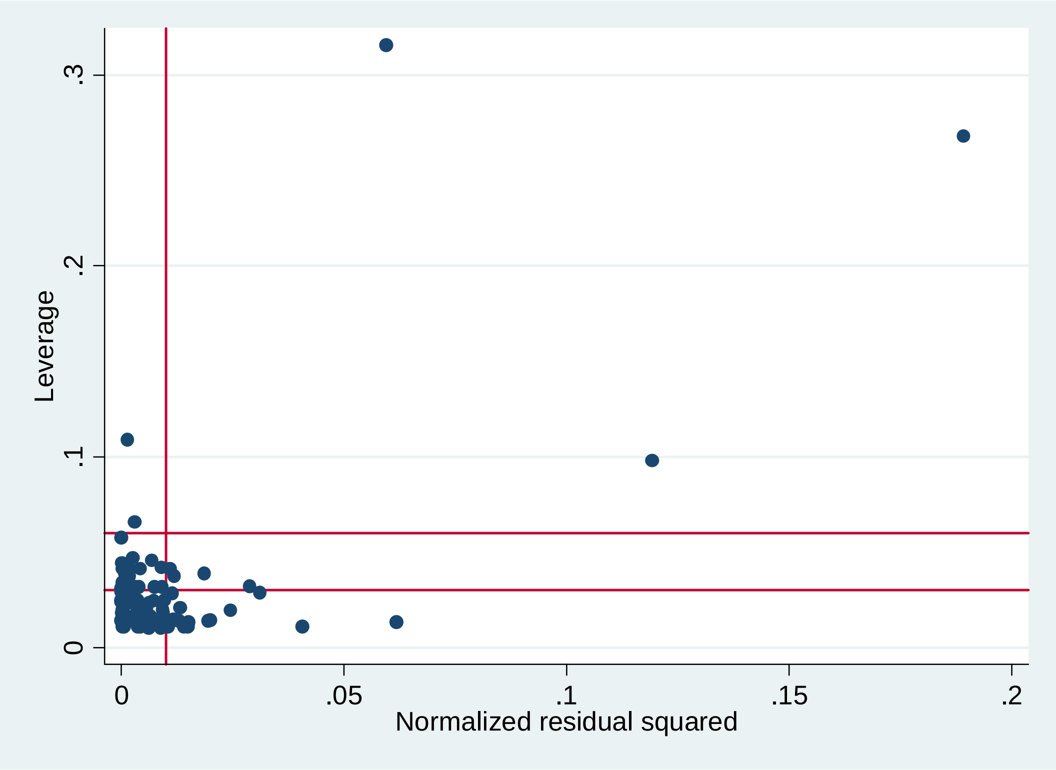
Any observation with leverage values about 0.06 may be influential. Let's generate a flag variable that indicates this and let's also plot the data using separate subplots for high leverage observations and normal observations.

```
. gen highlev = 1 if h > 0.06
(95 missing values generated)

. replace highlev = 0 if h <= 0.06
(95 real changes made)
```





Stata also allows you to generate a simultaneous plot of residuals and leverage. Go to **Graphics > Regression diagnostic plots > Leverage versus squared residual plot**. This command takes no arguments to just hit enter. You obtain a plot that shows the leverage on the y-axis and the squared residual on the x-axis. It places red lines at the average value of each. Once you obtain the plot, click on "Start graph editor" and then select **Graph > Add horizontal reference line**. Enter 0.06 since that is twice our average leverage and hit enter and enter again. You should obtain:



You can easily see the five points of high leverage above the top horizontal red line.

Cook's Distance

Stata has an option for generating Cook's distance values after a regression as well. After running the regression, type:

predict cooksd, cooksd

This will generate the variable "cooksd" which will contain the Cook's distance values. We can then look up the median value of the F-distribution with k+1 numerator and n-(k+1) denominator degrees of freedom and generate a flag of whether an observation has a high Cook's distance than this value.

```
. display invFtail(3,97, .5)
.79423686

. gen highcookd =1 if cooksd>0.79423686
(98 missing values generated)

. replace highcookd =0 if cooksd<=0.79423686
(98 real changes made)
```

Note that we can also generate jackknifed Studentized residuals by using the command:

predict studresid, rstudent

If you list these side-by-side with the standardized residuals you will see that they are virtually identical. While this may not always be the case, it oftentimes is. These follow a Student's t-distribution with (n-1) – (k+1) df.

DFBetas and DFITS

If you are interested to see how each point affects to estimates of the betas, you can ask Stata to give you the DFBETAS after the regression. Type:

```
. dfbeta
                      DFheadcirc:  DFbeta(headcirc)
                        DFlength:  DFbeta(length)
```

This created two new variables -- one for each beta. You can now list points that are labeled as high by leverage or by Cook's D.

```
. list headcirc length birthwt highlev highcookd DFheadcirc DFlength if highlev==1 |
highcookd==1

     +----------------------------------------------------------------------------+
     | headcirc   length   birthwt   highlev   highco~d   DFheadc~c    DFlength |
     |----------------------------------------------------------------------------|
  8. |       23       20       620         1          1    1.044203   -1.957564 |
 19. |       24       29       760         1          0    .0492917   -.1246761 |
 52. |       27       30      1320         1          0    .9105416   -1.191028 |
 99. |       33       39      1440         1          0   -.1296314    .0742994 |
100. |       35       36       900         1          1   -3.429696    2.553895 |
     +----------------------------------------------------------------------------+
```

Notice that those labeled as influential by Cook's distance have significantly larger DFBETA's than those labeled as high only by leverage.

If you would like to see how the predicted values from the entire data set differ from the jackknifed predicted values, you can also obtain those by typing:

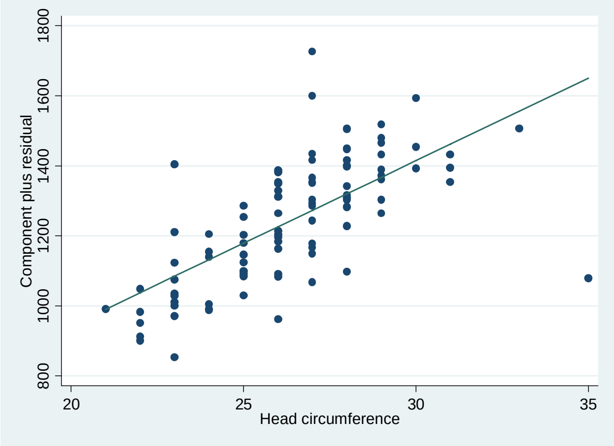
predict dfits, dfits

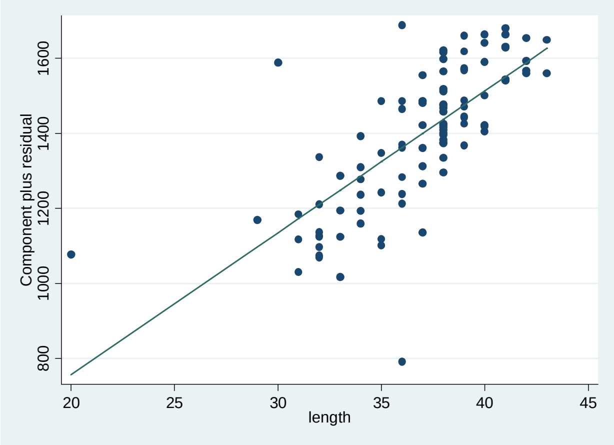
The results will generally parallel those seen by looking at the DFBETAs:

```
. list birthwt highlev highcookd DFheadcirc DFlength dfits if highlev==1 |
highcookd==1

     +-------------------------------------------------------------------+
     | birthwt   highlev   highco~d   DFheadc~c    DFlength       dfits |
     |-------------------------------------------------------------------|
  8. |     620         1          1    1.044203   -1.957564    2.053299 |
 19. |     760         1          0    .0492917   -.1246761    .1486323 |
 52. |    1320         1          0    .9105416   -1.191028    1.260319 |
 99. |    1440         1          0   -.1296314    .0742994   -.1386995 |
100. |     900         1          1   -3.429696    2.553895   -3.499185 |
     +-------------------------------------------------------------------+
```

Partial Residual Plots

Recall that partial residuals consider the residual values based on predicted values that do not consider the independent variable under consideration. Plotting the partial residuals will often lead to a better understanding of the true relationship between the response and the independent variable. To generate these in Stata, go to **Graphics > Regression diagnostic plots > Component plus residual plot** and choose the independent variable you want to consider. The plots for head circumference and length are below:





Notice that both plots look linear indicating our assumption of linearity for the deterministic part of the model is accurate.

General Test for Heteroscedasrticity

There are many tests available for detecting heteroscedasticity. One such test is called the Cook-Weisberg or Breusch-Pagan test. Its null hypothesis is that the variances of the residuals are constant for all values of the independent variables, while its alternative is that the variances are not equal. To generate this test, you can simply type "hettest" after running the regression. This considers the residuals versus the fitted values. The command "hettest, rhs" considers the residuals against the independent variables. For the birthweight data we have,

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of birthwt

         chi2(1)      =     0.22
         Prob > chi2  =   0.6413

. hettest, rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: headcirc length

         chi2(2)      =   123.59
         Prob > chi2  =   0.0000
```

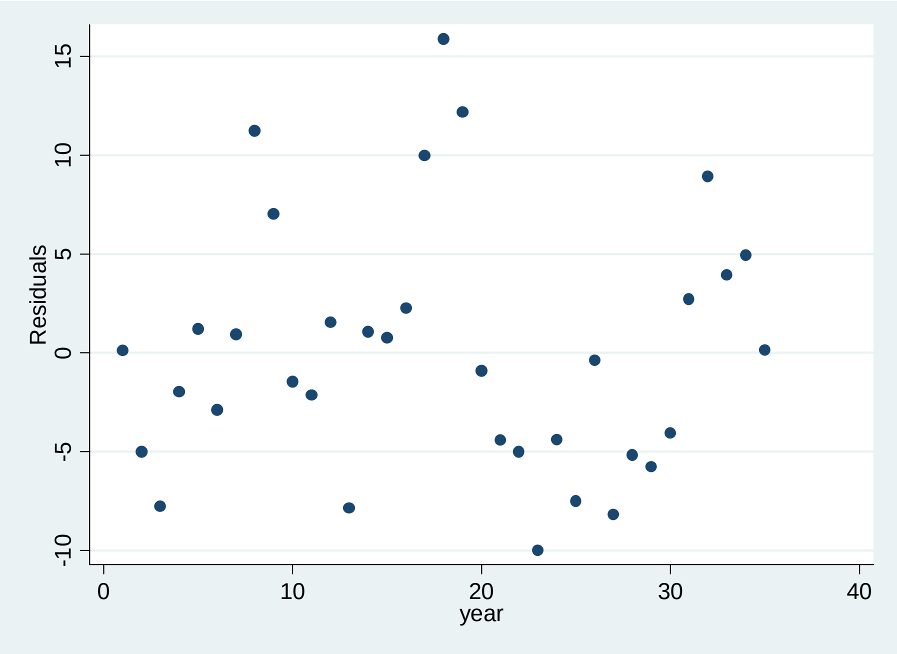
You can see that the homogeneity of variances assumption is violated when considering the independent variables, but this is not detected when considering only the fitted values. Be careful to look at both tests. They are extremely sensitive to the normality assumption of the residuals.

Introduction to Time Series Data

Time series data occur when observations are made on the same subjects repeatedly through time. Often there are seasonal variations or cyclical components to such data. For this example, consider the *sales35.dta* file on the course webpage. It contains sales data over a 350year period. If we regress sales on year we obtain:

```
. regress sales year

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  1,    33) = 1615.72
       Model |  65875.2068     1  65875.2068           Prob > F      =  0.0000
    Residual |  1345.45362    33  40.7713217           R-squared     =  0.9800
-------------+------------------------------           Adj R-squared =  0.9794
       Total |  67220.6604    34  1977.07825           Root MSE      =  6.3852

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |    4.29563   .1068669    40.20   0.000     4.078208    4.513053
       _cons |   .4015129   2.205708     0.18   0.857    -4.086034     4.88906
------------------------------------------------------------------------------
```

The residual plot looking at year is on the top of the next page. You can clearly see the cyclical nature in the data. To properly analyze this data set, we need to declare it as a time series data set in Stata.



To declare a time series data set in Stata, go to **Statistics** > **Time series** > **Setup and utilities** > **Declare dataset to be time series** and enter the time variable (year in this case) in the box.

```
. tsset year
        time variable:  year, 1 to 35
                delta:  1 unit
```

Now we can use the special built-in time series tests and commands such as that for the Durbin-Watson test. To generate the d-statistic, type:

```
. tsset year
        time variable:  year, 1 to 35
                delta:  1 unit

. regress sales year

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  1,    33) = 1615.72
       Model |  65875.2068     1  65875.2068           Prob > F      =  0.0000
    Residual |  1345.45362    33  40.7713217           R-squared     =  0.9800
-------------+------------------------------           Adj R-squared =  0.9794
       Total |  67220.6604    34  1977.07825           Root MSE      =  6.3852

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |    4.29563   .1068669    40.20   0.000     4.078208    4.513053
       _cons |   .4015129   2.205708     0.18   0.857    -4.086034     4.88906
------------------------------------------------------------------------------

. estat dwatson

Durbin-Watson d-statistic(  2,    35) =  .8207266
```

By using Table 8 in Appendix C, we can determine that this is *d* is less than the lower cutoff with an alpha of 0.05. So there is evidence of positive correlation among the residuals. We would need to control fopr this correlation to obtain accurate confidence intervals and hypothesis tests. This test is sensitive to the normality assumption of the residuals.