MA397 – Stepwise Regression and Residual Analyses in Stata

Goals

We will see how to run automated stepwise regressions in Stata and how to perform some basic residual diagnostics.

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.

Running an Automated Stepwise Regression in Stata

Stata has a built-in command for automatically running a stepwise regression analysis. It can perform forward selection and stepwise modeling, as well as backward selection and stepwise modeling. Which method is used depends mainly on which entrance/exclusion criteria you enter. The basic format of the stepwise command is:

sw regress y x1 x2 x3…xk, pe(alpha1) pr(alpha2)

The "pe()" option specifies the alpha level at which an independent variable has to be significant for entrance into the model. The "pr()" option specifies the alpha level at which an independent variable must fail to achieve significance in order to be removed from the model.

- If you include only the "pe()" option, Stata will perform a forward selection.
- If you include only the "pr()" option, Stata will perform a backward selection.
- If you include both the "pe()" and "pr()" options, Stata will perform a backward stepwise.
- If you include the "pe()" and "pr()" options *with* the "forward" option, Stata will perform a forward stepwise.

Note that alpha2 must be larger than alpha1 in Stata.

If we ran a forward stepwise regression on the birthweight data, we would obtain the following:

```
. sw regress birthwt headcirc length gestage momage toxemia, pe(.05) pr(.1) forward
                     begin with empty model
p = 0.0000 <  0.0500  adding   length
p = 0.0000 <  0.0500  adding   headcirc
p = 0.0390 <  0.0500  adding   toxemia

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =  108.20
       Model |  5569504.96     3  1856501.65           Prob > F      =  0.0000
    Residual |  1647237.79    96   17158.727           R-squared     =  0.7717
-------------+------------------------------           Adj R-squared =  0.7646
       Total |  7216742.75    99  72896.3914           Root MSE      =  130.99

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |    38.0639   5.256473     7.24   0.000     27.62989    48.49792
    headcirc |   48.36315   7.434922     6.50   0.000     33.60495    63.12135
     toxemia |  -67.92159   32.45179    -2.09   0.039    -132.3379   -3.505288
       _cons |  -1567.605   148.7759   -10.54   0.000    -1862.923   -1272.287
------------------------------------------------------------------------------
```

Note that if we forgot the "forward" option a backward stepwise regression would have been run:

```
. sw regress birthwt headcirc length gestage momage toxemia, pe(.05) pr(.1)
                     begin with full model
p = 0.8883 >= 0.1000  removing momage
p = 0.7442 >= 0.1000  removing gestage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =  108.20
       Model |  5569504.96     3  1856501.65           Prob > F      =  0.0000
    Residual |  1647237.79    96   17158.727           R-squared     =  0.7717
-------------+------------------------------           Adj R-squared =  0.7646
       Total |  7216742.75    99  72896.3914           Root MSE      =  130.99

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    headcirc |   48.36315   7.434922     6.50   0.000     33.60495    63.12135
      length |    38.0639   5.256473     7.24   0.000     27.62989    48.49792
     toxemia |  -67.92159   32.45179    -2.09   0.039    -132.3379   -3.505288
       _cons |  -1567.605   148.7759   -10.54   0.000    -1862.923   -1272.287
------------------------------------------------------------------------------
```

We obtained the same model, but this will not always be the case. It is entirely possible to obtain very different models using forward and backward methods.

If you would like to include higher-order terms or interactions in the stepwise process, you can do so but will need to require them to be considered along with their lower-order terms or main effects. This will not allow you to consider theses variables individually but may be useful if you know complex relationships exist in the data. To do so, simply place the variables to be considered together in parentheses. Let's consider an example where we created the variable *lentox* as the interaction between length and toxemia:

```
. sw regress birthwt headcirc gestage momage (length toxemia lentox), pe(.05) pr(.1)
                      begin with full model
p = 0.8352 >= 0.1000  removing momage
p = 0.8483 >= 0.1000  removing gestage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  4,    95) =   85.05
       Model |  5641405.82     4  1410351.46           Prob > F      =  0.0000
    Residual |  1575336.93    95   16582.494           R-squared     =  0.7817
-------------+------------------------------           Adj R-squared =  0.7725
       Total |  7216742.75    99  72896.3914           Root MSE      =  128.77

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    headcirc |   46.05232   7.392782     6.23   0.000      31.3758    60.72885
      length |   35.89579   5.271313     6.81   0.000     25.43091    46.36067
     toxemia |  -854.8818   379.2735    -2.25   0.026    -1607.835   -101.9288
      lentox |   21.05083   10.10944     2.08   0.040     .9810546    41.12061
       _cons |  -1427.483   160.9943    -8.87   0.000    -1747.097   -1107.869
------------------------------------------------------------------------------
```

Generating and Plotting Residuals in Stata

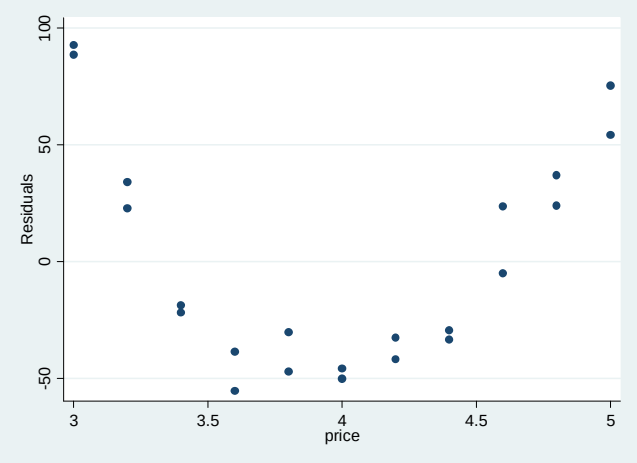
After running a regression in Stata, there are many diagnostic checks that can be made. Foremost among these is checking the residuals. Consider the data set *coffee2.dta* on the course webpage. The response variable is "price" (measured in dollars/pound), and the independent variables are "demand" (measured in pounds per week) and "advertisement" which indicated whether advertisement was used. The regression results are:

```
. regress demand price advertisement

      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  2,    19) =  373.71
       Model |  1859298.92     2  929649.461           Prob > F      =  0.0000
    Residual |  47264.8952    19  2487.62606           R-squared     =  0.9752
-------------+------------------------------           Adj R-squared =  0.9726
       Total |  1906563.82    21  90788.7532           Root MSE      =  49.876

------------------------------------------------------------------------------
      demand |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -456.2954   16.81323   -27.14   0.000    -491.4859   -421.1049
advertisem~t |   70.18182   21.26724     3.30   0.004     25.66897    114.6947
       _cons |   2400.182   68.91374    34.83   0.000     2255.944     2544.42
------------------------------------------------------------------------------
```

We can generate the residuals by typing, "predict resid, r". The "r" option tells Stata that we want the estimated residuals. We can easily generate both residual-versus-predictor plots, and residual-versus-fitted value plots. To do the former, go to **Graphics > Residual diagnostic plots > Residual versus predictor** and enter the name of the independent variable you want to consider in the box ("price" in this case).

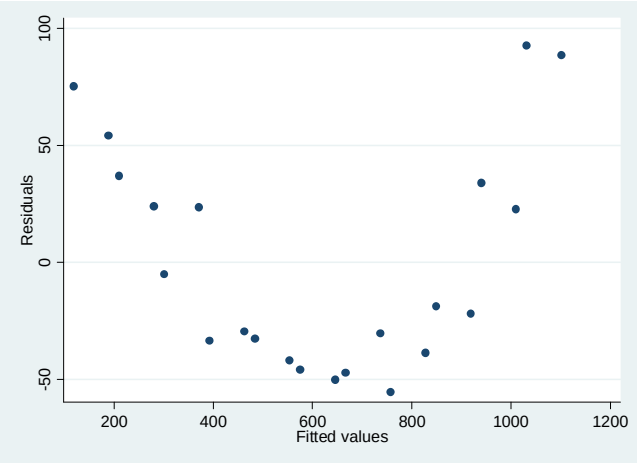


You should look for three things in this plot:

1. There should be no obvious pattern.
2. The amount of spread in the y-direction should be constant across the x-axis.
3. About 95% of the residuals should fall within 2s of the horizontal line at 0.

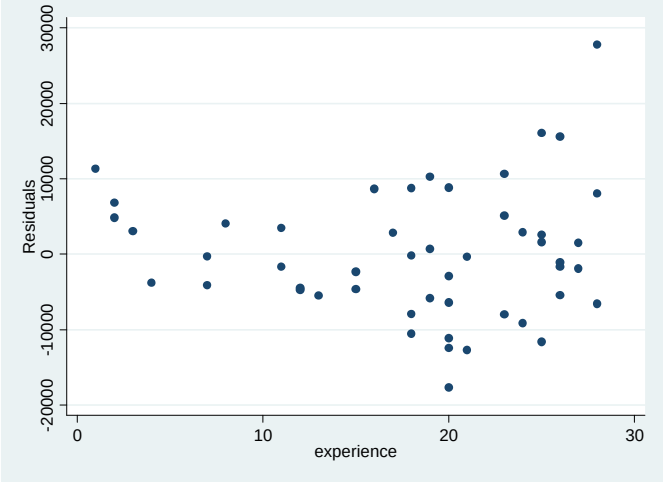
There is a clear violation of the first condition in this plot. This occurred since the relationship between price and demand is not linear.

To generate the residual-versus-fitted value plot go to **Graphics > Residual diagnostic plots > Residual versus fitted**. You should obtain the following:



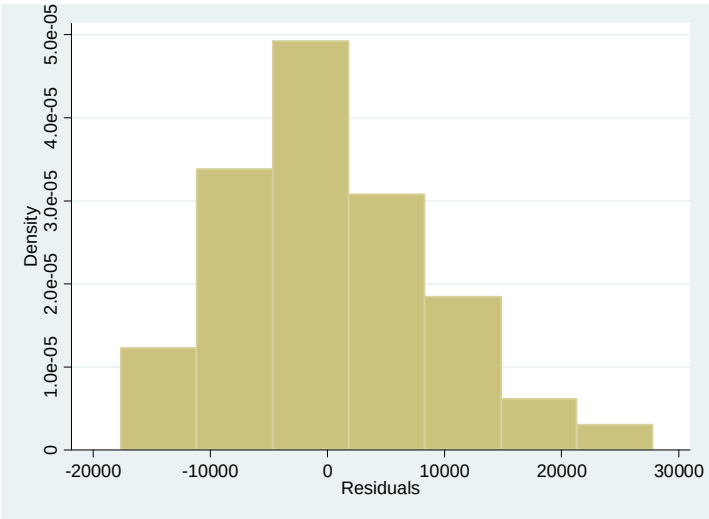
The same three conditions should hold for this plot. Again, it is clear the first condition does not. You should transform the response (or predictor) and try again.

Let's now consider the *socwork.dta* data set on the course webpage. The response variable is "salary" and the independent variable is "experience". Run the regression and generate a residual-versus-predictor plot using "experience". The plot is below,

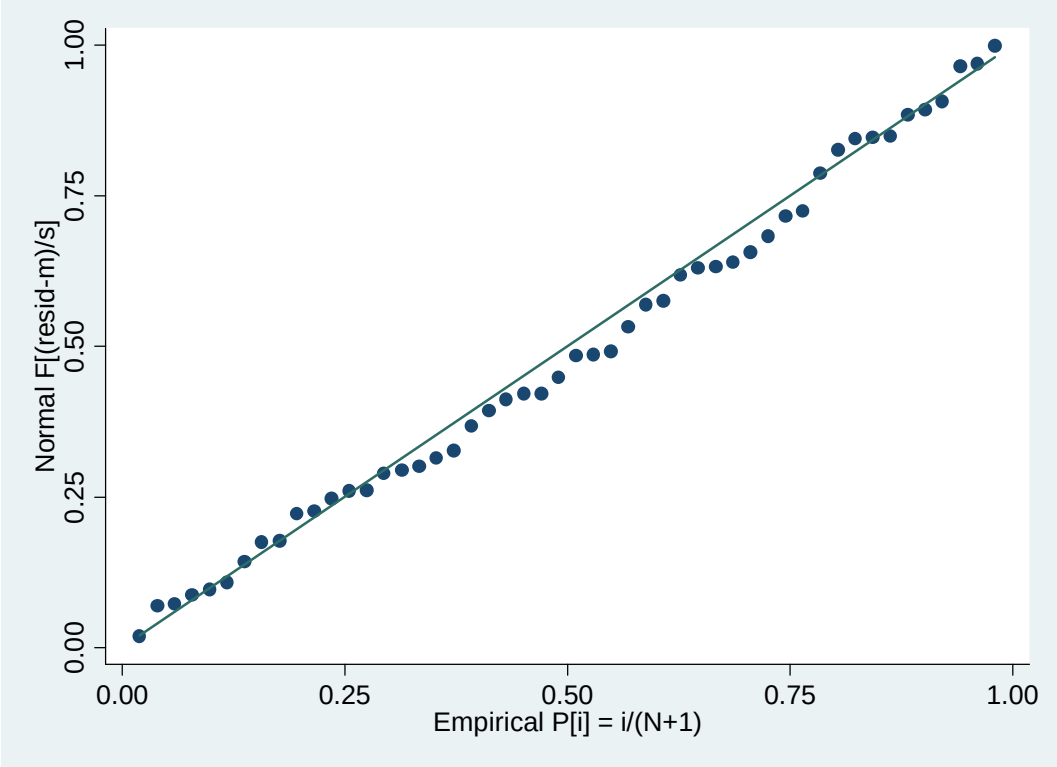


Although we see no non-linear pattern in this plot, we do see a violation of homoscedasticity. Taking the natural log of salary would rectify this. For brevity we leave this example at that, but also note that you should also generate a residual-versus-fitted value plot.

In order to check the normality assumption, we need to generate a histogram or normal probability plot of the residuals. To generate a histogram go to **Graphics > Histogram** and enter the name of the residuals in variable box. The plot is below:



The plot should look roughly bell-shaped which this one does. Now we generate the normal probability plot by going to **Graphics > Distribution plots > Normal probability plot** and entering the name of the variable containing the residuals, You should get,



This plot should follow a straight line which this one does. The assumption id normally distributed errors is upheld (although we violated the homoscedasticity assumption.