Mathematics 231

Lecture 33

Liam O'Brien

Announcements

- Today
  - Model Building

Model Building

- We have seen several examples of ways to select explanatory variables for a multiple regression model.
- We will examine all of the predictors in the birthweight dataset to determine which are important predictors.
- We had mother's age, toxemia, gestational age, infant length, and head circumference as predictors.

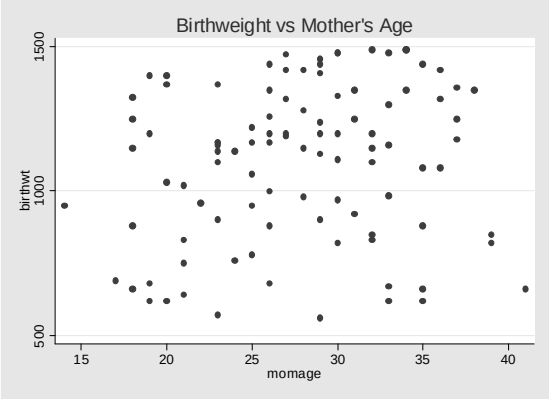
Continuous Predictors

- We should plot the response against each continuous explanatory variable to assess linearity.
- It's often useful to look at all pairwise correlations among the continuous predictors.
- We should then perform all possible simple linear regressions by regressing the outcome on each predictor, separately.
- The variable with the smallest p-value (or largest t-statistic) explains more of the variability of the response and should be included.
- This forward selection process is repeated until no variable adds to the model.
- We can also consider categorical variables in the selection.

Example

- We have 4 continuous predictors (momage, gestage, length, headcirc).
- We have one categorical (binary) predictor (toxemia).
- We have examined the plots of birthwt vs gestage, birthwt vs length, and birthwt vs headcirc.
- We now need to examine birthwt vs momage.

5



6

Regress Birthwt on Gestage

```
. regress birthwt gestage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   75.61
       Model |  3143019.07     1  3143019.07           Prob > F      =  0.0000
    Residual |  4073723.68    98   41568.609           R-squared     =  0.4355
-------------+------------------------------           Adj R-squared =  0.4298
       Total |  7216742.75    99  72896.3914           Root MSE      =  203.88

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     gestage |   70.30993   8.085854     8.70   0.000     54.26382    86.35604
       _cons |  -932.4039   234.4884    -3.98   0.000    -1397.738   -467.0693
------------------------------------------------------------------------------
```

7

Regress Birthwt on Headcirc

```
. regress birthwt  headcirc

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  172.82
       Model |  4605298.87     1  4605298.87           Prob > F      =  0.0000
    Residual |  2611443.88    98  26647.3866           R-squared     =  0.6381
-------------+------------------------------           Adj R-squared =  0.6344
       Total |  7216742.75    99  72896.3914           Root MSE      =  163.24

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    headcirc |   85.17802   6.479268    13.15   0.000     72.32013    98.03592
       _cons |  -1154.109   172.1523    -6.70   0.000    -1495.739    -812.478
------------------------------------------------------------------------------
```

8

Regress Birthwt on Length

```
. regress birthwt length

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  194.65
       Model |  4800034.87     1  4800034.87           Prob > F      =  0.0000
    Residual |  2416707.88    98  24660.2845           R-squared     =  0.6651
-------------+------------------------------           Adj R-squared =  0.6617
       Total |  7216742.75    99  72896.3914           Root MSE      =  157.04

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   61.65408   4.419149    13.95   0.000     52.88442    70.42373
       _cons |  -1171.253   163.4691    -7.16   0.000    -1495.652    -846.854
------------------------------------------------------------------------------
```


Regress Birthwt on Momage

```
. regress birthwt momage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    2.40
       Model |  172425.647     1  172425.647           Prob > F      =  0.1247
    Residual |  7044317.1     98  71880.7868           R-squared     =  0.0239
-------------+------------------------------           Adj R-squared =  0.0139
       Total |  7216742.75    99  72896.3914           Root MSE      =  268.11

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      momage |   6.975444   4.503782     1.55   0.125    -1.962164    15.91305
       _cons |   905.4209   127.7352     7.09   0.000     651.9346    1158.907
------------------------------------------------------------------------------
```


Example

- Infant length resulted in the largest t-statistic (13.95) just beating out head circumference.
- We include length in the model and add one additional continuous variable at a time to the model.
- We will also take notice to see if the predictive ability of length decreases with the addition of each variable (collinearity).

Add Headcirc

```
. regress birthwt length  headcirc

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =  154.71
       Model |  5494338.63     2  2747169.31           Prob > F      =  0.0000
    Residual |  1722404.12    97  17756.7435           R-squared     =  0.7613
-------------+------------------------------           Adj R-squared =  0.7564
       Total |  7216742.75    99  72896.3914           Root MSE      =  133.25

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   37.82793   5.346058     7.08   0.000     27.21749    48.43838
    headcirc |   47.15045   7.540371     6.25   0.000      32.1849      62.116
       _cons |  -1541.104   150.7971   -10.22   0.000    -1840.394   -1241.813
------------------------------------------------------------------------------
```

Add Gestage

```
. regress birthwt length gestage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =  106.48
       Model |  4958359.24     2  2479179.62           Prob > F      =  0.0000
    Residual |  2258383.51    97  23282.3043           R-squared     =  0.6871
-------------+------------------------------           Adj R-squared =  0.6806
       Total |  7216742.75    99  72896.3914           Root MSE      =  152.59

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   51.40366   5.821408     8.83   0.000     39.84978    62.95754
     gestage |   21.39402   8.204112     2.61   0.011     5.111133    37.67692
       _cons |  -1411.906   183.6993    -7.69   0.000    -1776.499   -1047.314
------------------------------------------------------------------------------
```

13

Add Momage

```
. regress birthwt length momage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   96.58
       Model |  4804117.61     2   2402058.8           Prob > F      =  0.0000
    Residual |  2412625.14    97  24872.4242           R-squared     =  0.6657
-------------+------------------------------           Adj R-squared =  0.6588
       Total |  7216742.75    99  72896.3914           Root MSE      =  157.71

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   62.05571   4.547482    13.65   0.000     53.03022     71.0812
      momage |  -1.099813   2.714578    -0.41   0.686    -6.487499    4.287872
       _cons |  -1155.543   168.6876    -6.85   0.000    -1490.342   -820.7452
------------------------------------------------------------------------------
```

14

Example

- Adding head circumference has slightly decreased the usefulness of length (the t-statistic decreased for length).
- Adding gestational age didn't decrease the usefulness of length as much, but it isn't as a good a predictor as head circumference.
- Notice the model with headcirc has a larger R^2 value than the model with gestage (0.76 to 0.69) and also a larger adjusted R^2.
- We will keep length and headcirc, and examine the remaining variables.

15

Add Gestage

```
. regress birthwt length headcirc  gestage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =  102.92
       Model |  5505033.78     3  1835011.26           Prob > F      =  0.0000
    Residual |  1711708.97    96  17830.3017           R-squared     =  0.7628
-------------+------------------------------           Adj R-squared =  0.7554
       Total |  7216742.75    99  72896.3914           Root MSE      =  133.53

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |    38.9962   5.565441     7.01   0.000     27.94889    50.04352
    headcirc |   51.30345   9.265343     5.54   0.000     32.91189    69.69501
     gestage |  -6.818412   8.803777    -0.77   0.441    -24.29377    10.65695
       _cons |  -1496.982   161.4911    -9.27   0.000     -1817.54   -1176.425
------------------------------------------------------------------------------
```

16

Add Momage

```
. regress birthwt length headcirc momage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =  102.18
       Model |  5495606.98     3  1831868.99            Prob > F      =  0.0000
    Residual |  1721135.77    96  17928.4976           R-squared     =  0.7615
-------------+------------------------------           Adj R-squared =  0.7541
       Total |  7216742.75    99  72896.3914           Root MSE      =   133.9

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   38.08653   5.459126     6.98   0.000     27.25025    48.92281
    headcirc |   47.08196   7.581126     6.21   0.000     32.03354    62.13037
      momage |  -.6133579   2.306035    -0.27   0.791    -5.190801    3.964085
       _cons |  -1531.805   155.5052    -9.85   0.000    -1840.481    -1223.13
------------------------------------------------------------------------------
```


Add Toxemia

```
. regress birthwt length headcirc 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    -3.50529
       _cons |  -1567.605   148.7759   -10.54   0.000    -1862.923   -1272.287
------------------------------------------------------------------------------
```


Example

- We see that adding toxemia has helped in our prediction of birthweight.
- We will include toxemia in the model, but now want to consider its interactions with head circumference and infant length.
- We have seen collinearity with gestational age and both head circumference and length and will not consider it further.
- Mother's age has given us no indication that it is a useful predictor at all.
- We consider toxemia as a modifying variable.

Add Length*Toxemia

```
. regress birthwt length headcirc toxemia  lentox

      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]
-------------+----------------------------------------------------------------
      length |   35.89579   5.271313     6.81   0.000     25.43091    46.36067
    headcirc |   46.05232   7.392782     6.23   0.000      31.3758    60.72885
     toxemia |  -854.8818   379.2735    -2.25   0.026    -1607.835   -101.9288
      lentox |   21.05083   10.10944     2.08   0.040      .981055    41.12061
       _cons |  -1427.483   160.9943    -8.87   0.000    -1747.097   -1107.869
------------------------------------------------------------------------------
```

Add Headcirc*Toxemia

```
. regress birthwt length headcirc toxemia headtox

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  4,    95) =   81.81
       Model |  5593029.44     4  1398257.36            Prob > F      =  0.0000
    Residual |  1623713.31    95  17091.7191            R-squared     =  0.7750
-------------+------------------------------           Adj R-squared =  0.7655
       Total |  7216742.75    99  72896.3914            Root MSE      =  130.74

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   37.99741   5.246506     7.24   0.000     27.58178    48.41304
    headcirc |   45.28098   7.871737     5.75   0.000     29.65361    60.90836
     toxemia |   -475.261    348.715    -1.36   0.176    -1167.548    217.0258
     headtox |   15.12886   12.89553     1.17   0.244    -10.47199    40.72971
       _cons |  -1484.175   164.6359    -9.01   0.000    -1811.019   -1157.331
------------------------------------------------------------------------------
```

21

Example

- It looks like headcirc, length, toxemia, and length*toxemia give us the best model.
- We should now at least look at the residual vs predicted value plot (looking at residual vs predictor plots when there is an interaction is difficult).

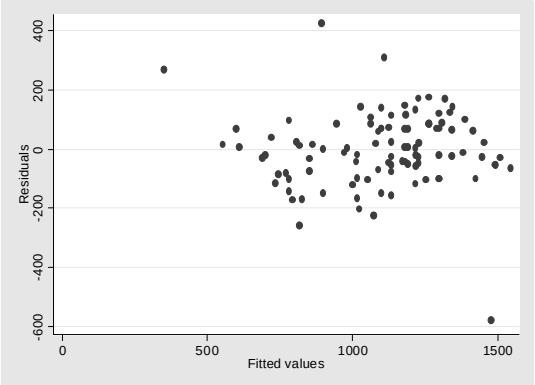
22

Regression Diagnostics in Stata

- Click on **Graphics > Regression diagnostic plots**.
- For a residual vs fitted value plot choose **residual-vs-fitted** plot.
- For an added variable plot choose the **added variable** plot and the "all variables" option. Otherwise you can examine predictors individually.
- For a leverage plot choose the **leverage versus fitted value** plot.
- For any of these plots, you may choose an identifying variance to put in the "Marker Label" box under the Plot tab.

23

Residuals versus Fitted Values



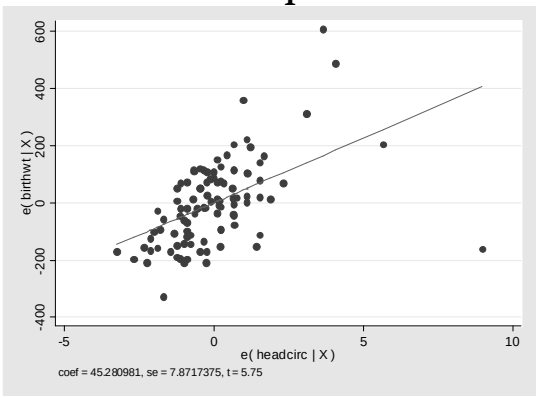
24

Added Component Plot

- These plots examine the relationship between the predicted response with each explanatory variable, averaging over the explanatory variables not under immediate consideration.
- These plots should not depart markedly from linearity.
- These aren't generally appropriate to run for a model with an interaction term.
- We will consider this plot for the predictor not involved in the interaction – head circumference.

25

Added Component Plot



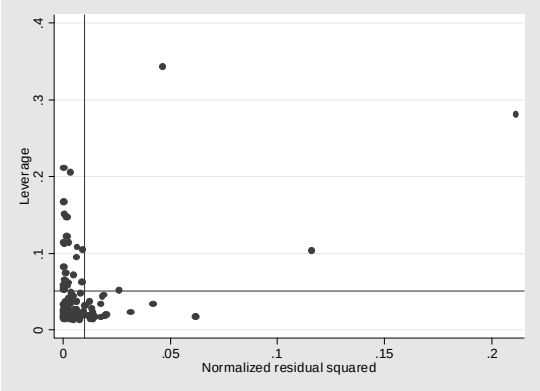
26

More Diagnostics

- We still need to consider whether we have points that have had undue influence on our model.
- This can be checked by looking at leverage plots.
- We consider the leverage plot for the final model we've obtained here.

27

Leverage Plot



28

Leverage Plot

- The horizontal line represents the average "leverage"
- Points far above it have a possibly undue influence on the regression.
- The vertical line represents the average normalized squared residual.
- Points far to the right of this line represent points with large residuals.

29