LPO 9952

Model Design

We'll be working today with the wage2 dataset, which includes monthly wages of male earners along with a variety of characteristics. We'll be attempting to esimtate some fairly standard wage models, but we'll also try to answer the most vexing question for many students: what variables should I put in my model?

The most important answer to that question is to use theory. Theory and previous results are our only guide—the data simply can't tell you by themselves what belongs in the model and what doesn't. However, we can use a combination of theory and applied data analysis to come up with a model that fits the data well and says something interesting about theory.

```
. version 15 /* Can set version here, use the most recent as default */

. capture log close /* Closes any logs, should they be open */

. log using "model_design_do.log",replace /*Open up new log */
--------------------------------------------------------------------------------
      name:  <unnamed>
       log:  /Users/doylewr/lpo_prac/lessons/s2-06-model_design/model_design_do.log
  log type:  text
 opened on:  11 Mar 2021, 08:52:19

. clear

. clear matrix

. graph drop _all

. estimates clear /* Clears any estimates hanging around */

. set more off /*Get rid of annoying "more" feature */

. ssc install nnest
checking nnest consistency and verifying not already installed...
all files already exist and are up to date.

. bcuse wage2, clear

Contains data from http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta
  obs:           935
 vars:            17                          26 Jan 2000 12:16
```

```
-----------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-----------------------------------------------------------------------------------
wage            float   %9.0g
hours           float   %9.0g
IQ              float   %9.0g
KWW             float   %9.0g
educ            float   %9.0g
exper           float   %9.0g
tenure          float   %9.0g
age             float   %9.0g
married         float   %9.0g
black           float   %9.0g
south           float   %9.0g
urban           float   %9.0g
sibs            float   %9.0g
brthord         float   %9.0g
meduc           float   %9.0g
feduc           float   %9.0g
lwage           float   %9.0g
-----------------------------------------------------------------------------------
Sorted by:

. label variable wage "Wages from work in last month"

. label variable hours "Weekly hours"

. label variable IQ "IQ test"

. label variable KWW "Knowledge of world of work"

. label variable educ "Years of education"

. label variable tenure "Months in current job"

. label variable age "Age"

. label variable married "Married"

. label variable black "African-American"

. label variable south "South"

. label variable urban "Urban"
```

```
. label variable sibs "No. Siblings"

. label variable brthord "Birth order"

. label variable meduc "Mother's years of school"

. label variable feduc "Father's years of education"

. label variable lwage "ln Wage"

. renvars *, lower

. save wage2, replace
file wage2.dta saved

. local sig=.05

. local sigtail=`sig'/2

. graph twoway scatter wage educ


. graph export "wage_educ.pdf", replace
(file /Users/doylewr/lpo_prac/lessons/s2-06-model_design/wage_educ.pdf written in PDF

. graph twoway qfit wage age||scatter wage age


. graph export "wage_age.pdf", replace
(file /Users/doylewr/lpo_prac/lessons/s2-06-model_design/wage_age.pdf written in PDF f
```

Missing Data

Let's talk again about how Stata handles missing data. Let's assume that we want to estimate several nested models, first with hours, education and age, then the same model with mother's education, then the same model with father's education, then a final model with all variables. Our results look like this.

```
. di _N
935

. reg lwage hours educ age

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(3, 931)       =     46.91
```

```
       Model |  21.7514568         3  7.25048559   Prob > F        =    0.0000
    Residual |  143.904838       931   .15457018   R-squared       =    0.1313
-------------+----------------------------------   Adj R-squared   =    0.1285
       Total |  165.656294       934  .177362199   Root MSE        =    .39315

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0047011   .0017887    -2.63   0.009    -.0082115   -.0011906
        educ |   .0616404   .0058814    10.48   0.000     .0500981    .0731827
         age |   .0227339   .0041411     5.49   0.000     .0146069    .0308608
       _cons |   5.403279   .1732026    31.20   0.000     5.063366    5.743191
------------------------------------------------------------------------------


. reg lwage hours educ age meduc

      Source |       SS           df       MS      Number of obs   =       857
-------------+----------------------------------   F(4, 852)       =     38.47
       Model |  22.8514162         4  5.71285406   Prob > F        =    0.0000
    Residual |  126.509635       852  .148485487   R-squared       =    0.1530
-------------+----------------------------------   Adj R-squared   =    0.1490
       Total |  149.361051       856  .174487209   Root MSE        =    .38534

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0058052   .0018374    -3.16   0.002    -.0094115   -.0021989
        educ |   .0525597   .0064521     8.15   0.000     .0398957    .0652236
         age |   .0243798   .0042747     5.70   0.000     .0159896      .03277
       meduc |   .0184424   .0049725     3.71   0.000     .0086826    .0282022
       _cons |    5.33402   .1776844    30.02   0.000      4.98527    5.682771
------------------------------------------------------------------------------


. reg lwage hours educ age feduc

      Source |       SS           df       MS      Number of obs   =       741
-------------+----------------------------------   F(4, 736)       =     34.17
       Model |  20.2139719         4  5.05349299   Prob > F        =    0.0000
    Residual |  108.836202       736  .147875274   R-squared       =    0.1566
-------------+----------------------------------   Adj R-squared   =    0.1521
       Total |  129.050173       740  .174392126   Root MSE        =    .38455

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
       hours |   -.007041   .0019639    -3.59   0.000    -.0108964   -.0031855
        educ |   .0475258   .0070026     6.79   0.000     .0337785    .0612732
         age |   .0262759   .0045806     5.74   0.000     .0172834    .0352685
       feduc |   .0172076   .0047569     3.62   0.000     .0078689    .0265462
       _cons |   5.421121   .1897058    28.58   0.000     5.048692     5.79355
------------------------------------------------------------------------------
```

```
. reg lwage hours educ age feduc meduc

      Source |       SS           df       MS      Number of obs   =       722
-------------+----------------------------------   F(5, 716)       =     27.64
       Model |   20.519095         5  4.10381899   Prob > F        =    0.0000
    Residual |  106.292836       716  .148453682   R-squared       =    0.1618
-------------+----------------------------------   Adj R-squared   =    0.1560
       Total |  126.811931       721    .1758834   Root MSE        =     .3853

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0071408   .0019821    -3.60   0.000    -.0110321   -.0032495
        educ |    .046006   .0072322     6.36   0.000     .0318071    .0602049
         age |   .0249456   .0046705     5.34   0.000     .0157762     .034115
       feduc |   .0114239   .0055571     2.06   0.040     .0005138     .022334
       meduc |   .0132414   .0063094     2.10   0.036     .0008543    .0256285
       _cons |   5.404781   .1929204    28.02   0.000     5.026024    5.783539
------------------------------------------------------------------------------
```

The results are extermely problematic because each set of results is on a different sample! The first set has 857 observations, the second 741, and down to 722 for the final one. Stata performs casewise deletion when running regressions, and doesn't adjust unless you tell it to. In this case none of the standard tests of model fit are relevant, because it's not the same sample.

The solution is to use the `e(sample)` command to limit the sample to the relevant analysis sample. First, run the model that restricts the data the most (has the most missing data), then limit subsequent models using the statement `if e(sample)==1`.

```
. gen analytic_sample_flag=e(sample)

. reg lwage hours educ age if analytic_sample_flag==1

      Source |       SS           df       MS      Number of obs   =       722
-------------+----------------------------------   F(3, 718)       =     39.52
       Model |   17.972245         3  5.99074834   Prob > F        =    0.0000
```

```
    Residual |  108.839686       718  .151587307   R-squared       =    0.1417
-------------+----------------------------------   Adj R-squared   =    0.1381
       Total |  126.811931       721    .1758834   Root MSE        =    .38934

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0067617   .0019997    -3.38   0.001    -.0106877   -.0028357
        educ |   .0593979   .0065158     9.12   0.000     .0466056    .0721902
         age |   .0236215   .0046986     5.03   0.000      .014397    .0328461
       _cons |    5.50891   .1932456    28.51   0.000     5.129516    5.888304
------------------------------------------------------------------------------


. reg lwage hours educ age meduc if analytic_sample_flag==1

      Source |       SS           df       MS      Number of obs   =       722
-------------+----------------------------------   F(4, 717)       =     33.35
       Model |  19.8917203         4  4.97293007   Prob > F        =    0.0000
    Residual |  106.920211       717  .149121633   R-squared       =    0.1569
-------------+----------------------------------   Adj R-squared   =    0.1522
       Total |  126.811931       721    .1758834   Root MSE        =    .38616

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0071587   .0019865    -3.60   0.000    -.0110588   -.0032587
        educ |    .050238   .0069486     7.23   0.000      .036596    .0638801
         age |   .0240817    .004662     5.17   0.000      .014929    .0332345
       meduc |   .0196873   .0054874     3.59   0.000      .008914    .0304606
       _cons |   5.423663   .1931347    28.08   0.000     5.044485     5.80284
------------------------------------------------------------------------------


. gen meduc_flag=meduc==.

. gen feduc_flag=feduc==.

. tab meduc_flag feduc_flag

           |      feduc_flag
meduc_flag |         0          1 |     Total
-----------+----------------------+----------
         0 |       722        135 |       857
         1 |        19         59 |        78
-----------+----------------------+----------
```

```
   Total |       741        194 |       935
```

The log transformation

The variable `lwage` is the natural log of wages. This means that it has been transformed by taking the natural log of the underlying variable:

$$log_e(y_i) = x \equiv e^x = y_i$$

Where e is Euler's constant,

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = \frac{1}{1} + \frac{1}{1 \times 2} + \frac{1}{1 \times 2 \times 3} \dots$$

The log transformation is used all the time, and particularly in econometrics. It's useful whenever you have a variable that follows some kind of exponential distribution, with widely disparate levels. Earnings, school sizes, revenues of instititions of higher education and state populations are all examples of these kinds of situations.

When the dependent variable is log transformed but the independent variable is not, this is called a log-level regression. In a log-level regression, the following applies:

$$log(y_i) = \beta_o + \beta_1 x_i + \epsilon_i$$

Which implies that

$$y_i = e^{\beta_o + \beta_1 x_i + \epsilon_i}$$

And . . .

$$\frac{dy}{dx} = \beta e^{\beta_0 + \beta_1 x_1 + \epsilon} = \beta_1 y$$

Which means that the coefficent, β_1

$$\beta_1 = \frac{dy}{dx}\frac{1}{y}$$

This changes our interpretation to mean that for a one unit increase in x, y is predicted to increase by β_1 proportion of y or more commonly by $100 * \beta_1$

percent. It changes the scale of the dependent variable to be on the $1/y$ scale as opposed to the y scale, so everything is about a proportional (or percentage) increase in y.

Quick Exercise

Interpret the coefficients from the basic earnings regression of log wages on years of education.

```
. preserve

. clear

. di log(0)
.

. di log(1)
0

. di log(10)
2.3025851

. di log(100)
4.6051702

. di log(1000)
6.9077553

. set obs 1000
number of observations (_N) was 0, now 1,000

. egen fakenumber= fill(1(10)1000)

. gen log_fakenumber=log(fakenumber)

. graph twoway line log_fakenumber fakenumber


. restore
```

Stepwise Regression: Proceed with Caution

When selecting variables for a model, students are sometimes tempted by the dark side of stepwise regression, which is a step on the path toward the greater evil that is data mining. I will illustrate why this is a bad idea. The basic idea with stepwise regression is to eliminate variables from the model one at a time—if the variable is not significant, it gets dropped. However, this method is very sensitive to the overall group of variables used, essentially just pushing decisions one step back, and then using an arbirtray non-theoertical standard for variable inclusion. There is no good theoretical reason to use this procedure.

In data science, these approaches are used all of the time with the assumption that we won't learn anything meaningful about the parameters, but instead will get an accurate prediction. In cases where all we want is an accurate predcition, this approach is ok, but stepwise regression isn't used in modern practice any more.

```
. stepwise, pr(.2): reg lwage hours educ age meduc feduc tenure south married black ur
                      begin with full model
p = 0.3500 >= 0.2000  removing feduc
p = 0.2712 >= 0.2000  removing sibs
p = 0.2237 >= 0.2000  removing brthord

      Source |       SS           df       MS      Number of obs   =       663
-------------+----------------------------------   F(11, 651)      =     23.26
       Model |  31.7400529        11  2.88545936   Prob > F        =    0.0000
    Residual |  80.7431222       651  .124029374   R-squared       =    0.2822
-------------+----------------------------------   Adj R-squared   =    0.2700
       Total |  112.483175       662  .169914162   Root MSE        =    .35218

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0067126   .0019326    -3.47   0.001    -.0105074   -.0029177
        educ |   .0332989   .0078237     4.26   0.000     .0179361    .0486617
         age |   .0157517    .005244     3.00   0.003     .0054545    .0260488
       meduc |   .0134071   .0053659     2.50   0.013     .0028706    .0239437
         kww |   .0034485   .0023844     1.45   0.149    -.0012335    .0081305
      tenure |   .0081709   .0028567     2.86   0.004     .0025613    .0137804
       south |  -.0585439   .0305602    -1.92   0.056    -.1185524    .0014646
     married |   .2077864   .0461308     4.50   0.000     .1172033    .2983696
       black |  -.0938556    .054557    -1.72   0.086    -.2009845    .0132733
       urban |   .1980571   .0312361     6.34   0.000     .1367215    .2593928
          iq |   .0034329    .001223     2.81   0.005     .0010314    .0058343
       _cons |   5.151168   .2180037    23.63   0.000     4.723092    5.579243
------------------------------------------------------------------------------

. stepwise, pr(.05): reg lwage hours educ age meduc feduc tenure south married black u
                      begin with full model
p = 0.3500 >= 0.0500  removing feduc
p = 0.2712 >= 0.0500  removing sibs
p = 0.2237 >= 0.0500  removing brthord
p = 0.1486 >= 0.0500  removing kww
```

```
p = 0.0689 >= 0.0500  removing south

      Source |       SS           df       MS      Number of obs   =       663
-------------+----------------------------------   F(9, 653)       =     27.69
       Model |  31.0680934         9  3.45201038   Prob > F        =    0.0000
    Residual |  81.4150818       653  .124678533   R-squared       =    0.2762
-------------+----------------------------------   Adj R-squared   =    0.2662
       Total |  112.483175       662  .169914162   Root MSE        =     .3531

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0064609   .0019327    -3.34   0.001    -.0102559   -.0026659
        educ |   .0347812   .0076565     4.54   0.000     .0197469    .0498156
         age |   .0195908   .0047192     4.15   0.000     .0103242    .0288574
       meduc |   .0150286   .0053333     2.82   0.005      .004556    .0255011
          iq |   .0040984   .0011802     3.47   0.001     .0017809    .0064159
      tenure |   .0088182   .0028499     3.09   0.002     .0032221    .0144143
       urban |   .2102756   .0308692     6.81   0.000     .1496607    .2708905
     married |   .2095021   .0461979     4.53   0.000     .1187877    .3002164
       black |  -.1163507   .0538301    -2.16   0.031    -.2220516   -.0106498
       _cons |   5.000151    .206967    24.16   0.000      4.59375    5.406552
------------------------------------------------------------------------------

. stepwise, pr(.2) : reg lwage south brthord iq kww sibs feduc tenure  married black u
                      begin with full model
p = 0.3500 >= 0.2000  removing feduc
p = 0.2712 >= 0.2000  removing sibs
p = 0.2237 >= 0.2000  removing brthord

      Source |       SS           df       MS      Number of obs   =       663
-------------+----------------------------------   F(11, 651)      =     23.26
       Model |  31.7400529        11  2.88545936   Prob > F        =    0.0000
    Residual |  80.7431222       651  .124029374   R-squared       =    0.2822
-------------+----------------------------------   Adj R-squared   =    0.2700
       Total |  112.483175       662  .169914162   Root MSE        =    .35218

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       south |  -.0585439   .0305602    -1.92   0.056    -.1185524    .0014646
        educ |   .0332989   .0078237     4.26   0.000     .0179361    .0486617
          iq |   .0034329    .001223     2.81   0.005     .0010314    .0058343
         kww |   .0034485   .0023844     1.45   0.149    -.0012335    .0081305
         age |   .0157517    .005244     3.00   0.003     .0054545    .0260488
       meduc |   .0134071   .0053659     2.50   0.013     .0028706    .0239437
```

```
      tenure |   .0081709   .0028567     2.86   0.004     .0025613    .0137804
     married |   .2077864   .0461308     4.50   0.000     .1172033    .2983696
       black |  -.0938556    .054557    -1.72   0.086    -.2009845    .0132733
       urban |   .1980571   .0312361     6.34   0.000     .1367215    .2593928
       hours |  -.0067126   .0019326    -3.47   0.001    -.0105074   -.0029177
       _cons |   5.151168   .2180037    23.63   0.000     4.723092    5.579243
------------------------------------------------------------------------------

. stepwise, pr(.05): reg lwage south brthord iq kww sibs feduc tenure  married black u
                      begin with full model
p = 0.3500 >= 0.0500  removing feduc
p = 0.2712 >= 0.0500  removing sibs
p = 0.2237 >= 0.0500  removing brthord
p = 0.1486 >= 0.0500  removing kww
p = 0.0689 >= 0.0500  removing south

      Source |       SS           df       MS      Number of obs   =       663
-------------+----------------------------------   F(9, 653)       =     27.69
       Model |  31.0680934         9  3.45201038   Prob > F        =    0.0000
    Residual |  81.4150818       653  .124678533   R-squared       =    0.2762
-------------+----------------------------------   Adj R-squared   =    0.2662
       Total |  112.483175       662  .169914162   Root MSE        =     .3531

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |   .2102756   .0308692     6.81   0.000     .1496607    .2708905
        educ |   .0347812   .0076565     4.54   0.000     .0197469    .0498156
          iq |   .0040984   .0011802     3.47   0.001     .0017809    .0064159
       hours |  -.0064609   .0019327    -3.34   0.001    -.0102559   -.0026659
         age |   .0195908   .0047192     4.15   0.000     .0103242    .0288574
       meduc |   .0150286   .0053333     2.82   0.005      .004556    .0255011
      tenure |   .0088182   .0028499     3.09   0.002     .0032221    .0144143
     married |   .2095021   .0461979     4.53   0.000     .1187877    .3002164
       black |  -.1163507   .0538301    -2.16   0.031    -.2220516   -.0106498
       _cons |   5.000151    .206967    24.16   0.000      4.59375    5.406552
------------------------------------------------------------------------------

. stepwise, pr(.2) : reg lwage south brthord  kww sibs feduc tenure  married black  ho
                      begin with full model
p = 0.5103 >= 0.2000  removing sibs
p = 0.2146 >= 0.2000  removing brthord

      Source |       SS           df       MS      Number of obs   =       663
-------------+----------------------------------   F(10, 652)      =     19.75
       Model |  26.1546559        10  2.61546559   Prob > F        =    0.0000
```

```
    Residual |  86.3285193       652  .132405704   R-squared       =    0.2325
-------------+----------------------------------   Adj R-squared   =    0.2207
       Total |  112.483175       662  .169914162   Root MSE        =    .36388

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       south |  -.0886972   .0312376    -2.84   0.005    -.1500357   -.0273588
         age |   .0120413   .0053403     2.25   0.024     .0015551    .0225275
         kww |   .0066253   .0023708     2.79   0.005       .00197    .0112805
       meduc |   .0106574   .0062851     1.70   0.090    -.0016842    .0229989
       feduc |   .0084136   .0055693     1.51   0.131    -.0025224    .0193497
      tenure |   .0078097   .0029507     2.65   0.008     .0020157    .0136036
     married |   .1987501   .0476411     4.17   0.000     .1052016    .2922985
       black |  -.0799257   .0545557    -1.47   0.143    -.1870519    .0272004
       hours |   -.006741   .0019951    -3.38   0.001    -.0106585   -.0028235
        educ |   .0412033   .0075936     5.43   0.000     .0262926    .0561141
       _cons |   5.508661   .2030461    27.13   0.000     5.109958    5.907365
------------------------------------------------------------------------------

. stepwise, pr(.05): reg lwage south brthord  kww sibs feduc tenure  married black  ho
                      begin with full model
p = 0.5103 >= 0.0500  removing sibs
p = 0.2146 >= 0.0500  removing brthord
p = 0.1434 >= 0.0500  removing black
p = 0.1099 >= 0.0500  removing feduc

      Source |       SS           df       MS      Number of obs   =       663
-------------+----------------------------------   F(8, 654)       =     24.00
       Model |  25.5304974         8  3.19131217   Prob > F        =    0.0000
    Residual |  86.9526778       654  .132955165   R-squared       =    0.2270
-------------+----------------------------------   Adj R-squared   =    0.2175
       Total |  112.483175       662  .169914162   Root MSE        =    .36463

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       south |  -.1005921   .0307895    -3.27   0.001    -.1610502    -.040134
         age |   .0108531   .0053138     2.04   0.042      .000419    .0212872
         kww |   .0074534   .0023359     3.19   0.001     .0028667    .0120401
       meduc |   .0161784   .0055007     2.94   0.003     .0053773    .0269795
       hours |  -.0066665   .0019986    -3.34   0.001    -.0105909   -.0027422
      tenure |   .0077155    .002955     2.61   0.009     .0019132    .0135178
     married |   .2023529   .0476962     4.24   0.000     .1086967    .2960091
        educ |   .0438651   .0073657     5.96   0.000     .0294018    .0583284
       _cons |     5.4996   .2016607    27.27   0.000      5.10362    5.895581
```

```
------------------------------------------------------------------------------
```

F Test

In choosing among model specifications that are nested, the F test is our basic guide. The F test looks at whether a linear restriction in the fully specified model results in a statistically significant decrease in model fit.

```
. reg lwage south brthord kww sibs feduc tenure married black hours educ age meduc

      Source |       SS       df       MS              Number of obs =     663
-------------+------------------------------           F(12, 650)    =   16.63
       Model |   26.416293    12  2.20135775           Prob > F      =  0.0000
    Residual |  86.0668822   650  .132410588           R-squared     =  0.2348
-------------+------------------------------           Adj R-squared =  0.2207
       Total |  112.483175   662  .169914162           Root MSE      =  .36388

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       south |     -.0838   .0314578    -2.66   0.008    -.1455711   -.0220289
     brthord |  -.0167837   .0119862    -1.40   0.162      -.04032    .0067526
         kww |   .0067305   .0023903     2.82   0.005     .0020369    .0114241
        sibs |   .0053631   .0081421     0.66   0.510    -.0106248     .021351
       feduc |   .0079674   .0055787     1.43   0.154     -.002987    .0189218
      tenure |   .0078633    .002952     2.66   0.008     .0020668    .0136598
     married |   .1967992   .0476677     4.13   0.000     .1031979    .2904005
       black |  -.0846613   .0558859    -1.51   0.130       -.1944    .0250775
       hours |  -.0069189   .0019991    -3.46   0.001    -.0108444   -.0029934
        educ |   .0409417   .0076049     5.38   0.000     .0260086    .0558747
         age |   .0118407   .0053439     2.22   0.027     .0013473    .0223341
       meduc |   .0096308   .0063914     1.51   0.132    -.0029196    .0221812
       _cons |   5.560053   .2098306    26.50   0.000     5.148025     5.97208
------------------------------------------------------------------------------


. test meduc feduc

 ( 1)  meduc = 0
 ( 2)  feduc = 0

       F(  2,   650) =    4.01
            Prob > F =    0.0187

. test meduc=feduc
```

```
 ( 1)  - feduc + meduc = 0

       F(  1,   650) =    0.03
            Prob > F =    0.8711

. test educ tenure hours

 ( 1)  educ = 0
 ( 2)  tenure = 0
 ( 3)  hours = 0

       F(  3,   650) =   15.51
            Prob > F =    0.0000
```

RESET Test

One question that comes up frequently is whether one or more variables ought to be expressed as quadratic or higher-order polynomials in the equation. The RESET test can help with this problem. Specifying the RESET test without any options means that Stata will fit the model with the second, third and fourth powers of $\hat{y}$. Specifying the option `rhs` will use powers of the individual regressors.

In Stata, we would run:

```
. reg lwage hours age educ

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(3, 931)       =     46.91
       Model |  21.7514568         3  7.25048559   Prob > F        =    0.0000
    Residual |  143.904838       931   .15457018   R-squared       =    0.1313
-------------+----------------------------------   Adj R-squared   =    0.1285
       Total |  165.656294       934  .177362199   Root MSE        =    .39315

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0047011   .0017887    -2.63   0.009    -.0082115   -.0011906
         age |   .0227339   .0041411     5.49   0.000     .0146069    .0308608
        educ |   .0616404   .0058814    10.48   0.000     .0500981    .0731827
       _cons |   5.403279   .1732026    31.20   0.000     5.063366    5.743191
------------------------------------------------------------------------------


. estat ovtest
```

```
Ramsey RESET test using powers of the fitted values of lwage
       Ho:  model has no omitted variables
                 F(3, 928) =      0.32
                  Prob > F =      0.8089

. estat ovtest, rhs

Ramsey RESET test using powers of the independent variables
       Ho:  model has no omitted variables
                 F(9, 922) =      2.12
                  Prob > F =      0.0255
```

The result of the first test is not significant, but the result of the second test is. This indicates that we might want to include some additional powers of the right hand variables. Let's begin by introducing a quadratic function of age:

```
. gen agesq=age^2

. label var agesq "Age squared"

. reg lwage hours educ age agesq

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(4, 930)       =     35.15
       Model |  21.7551592         4  5.43878981   Prob > F        =    0.0000
    Residual |  143.901135       930  .154732403   R-squared       =    0.1313
-------------+----------------------------------   Adj R-squared   =    0.1276
       Total |  165.656294       934  .177362199   Root MSE        =    .39336

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |     -.0047   .0017897    -2.63   0.009    -.0082123   -.0011876
        educ |   .0615585   .0059083    10.42   0.000     .0499634    .0731536
         age |   .0388675   .1043805     0.37   0.710    -.1659811    .2437162
       agesq |  -.0002425   .0015678    -0.15   0.877    -.0033193    .0028343
       _cons |   5.138356   1.721371     2.99   0.003     1.760134    8.516578
------------------------------------------------------------------------------


. test age agesq

 ( 1)  age = 0
 ( 2)  agesq = 0

       F(  2,   930) =   15.07
```

```
                 Prob > F =    0.0000
```

The two terms for age are jointly significant, but it looks like we could safely exclude age squared from the model without any loss of model fit.

Now let's try education squared:

```
. gen educsq=educ^2

. la var educsq "Education squared"

. reg lwage hours age educ educsq

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(4, 930)       =     36.27
       Model |  22.3576551         4  5.58941378   Prob > F        =    0.0000
    Residual |  143.298639       930  .154084558   R-squared       =    0.1350
-------------+----------------------------------   Adj R-squared   =    0.1312
       Total |  165.656294       934  .177362199   Root MSE        =    .39254

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |   -.004608   .0017866    -2.58   0.010    -.0081141   -.0011018
         age |   .0243563   .0042147     5.78   0.000     .0160848    .0326278
        educ |   .2161619   .0781252     2.77   0.006     .0628397     .369484
      educsq |  -.0054912   .0027685    -1.98   0.048    -.0109245    -.000058
       _cons |   4.286926   .5887928     7.28   0.000      3.13141    5.442443
------------------------------------------------------------------------------


. test educ educsq

 ( 1)  educ = 0
 ( 2)  educsq = 0

       F(  2,   930) =   57.06
            Prob > F =    0.0000
```

This does result in a statistically significant increase in model fit. The way I would prefer approaching this problem is to fully specify the model, then restrict it appropriately, like so:

```
. reg lwage hours age agesq educ educsq

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(5, 929)       =     29.00
       Model |  22.3674226         5  4.47348452   Prob > F        =    0.0000
```

```
   Residual |  143.288872       929  .154239905   R-squared       =    0.1350
-------------+----------------------------------   Adj R-squared   =    0.1304
      Total |  165.656294       934  .177362199   Root MSE        =    .39273

------------------------------------------------------------------------------
      lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hours |  -.0046056   .0017875    -2.58   0.010    -.0081135   -.0010976
        age |    .050602   .1043806     0.48   0.628     -.154247     .255451
      agesq |  -.0003944   .0015672    -0.25   0.801    -.0034699    .0026812
       educ |   .2169837   .0782328     2.77   0.006     .0634502    .3705171
     educsq |  -.0055252   .0027732    -1.99   0.047    -.0109676   -.0000828
      _cons |   3.849225   1.836393     2.10   0.036     .2452658    7.453184
------------------------------------------------------------------------------


. test age agesq

 ( 1)  age = 0
 ( 2)  agesq = 0

       F(  2,   929) =   16.71
            Prob > F =    0.0000

. test educ educsq

 ( 1)  educ = 0
 ( 2)  educsq = 0

       F(  2,   929) =   56.44
            Prob > F =    0.0000
```

Davidson-Mackinnon Test

In many situations, models are based on competing hypotheses, and so they don't nest within one another. Let's say we have one model that posits education as the key to wages, another that posits iq as the key to wages. To test whether one is better than the other, we use the Davidson-Mackinnon test:

```
. reg lwage hours iq

     Source |       SS         df       MS      Number of obs   =       935
-------------+----------------------------------   F(2, 932)       =     54.14
      Model |  17.2420918         2  8.62104588   Prob > F        =    0.0000
   Residual |  148.414203       932  .159242707   R-squared       =    0.1041
```

```
-------------+----------------------------------   Adj R-squared   =    0.1022
       Total |  165.656294       934  .177362199   Root MSE        =    .39905

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0041302   .0018124    -2.28   0.023     -.007687   -.0005734
          iq |   .0089535   .0008698    10.29   0.000     .0072465    .0106606
       _cons |   6.053607   .1150416    52.62   0.000     5.827837    6.279378
------------------------------------------------------------------------------


. nnest educ age

                   Competing Models
  --------------------------------------------------
  M1 : Y = [lwage]
       X = [hours iq]
  M2 : Y = [lwage]
       Z = [educ age]
  --------------------------------------------------
             J test for non-nested models
  --------------------------------------------------
                     Dist          Stat      P>|Stat|
  H0:M1 / H1:M2    t(931)          8.18       0.000
  H0:M2 / H1:M1    t(931)          6.77       0.000
  --------------------------------------------------
        Cox-Pesaran test for non-nested models
  --------------------------------------------------
                     Dist          Stat      P>|Stat|
  H0:M1 / H1:M2    N(0,1)        -12.47       0.000
  H0:M2 / H1:M1    N(0,1)        -10.17       0.000
  --------------------------------------------------
```

The results of this test indicate that it would be better to include both of these models, in a sort of "super" model.

Binary- Binary Interaction

Let's say we're interested in whether marriage is associated with wages differently for black and white men. The specification of an interaction between the two binary variables of white and married would look like this:

```
. gen black_marry=black*married

. eststo black_marry: reg lwage hours age educ i.black##i.married iq meduc south urbar
```

```
      Source |       SS           df       MS      Number of obs   =       857
-------------+----------------------------------   F(10, 846)      =     29.83
       Model |  38.9381574        10  3.89381574   Prob > F        =    0.0000
    Residual |  110.422894       846  .130523515   R-squared       =    0.2607
-------------+----------------------------------   Adj R-squared   =    0.2520
       Total |  149.361051       856  .174487209   Root MSE        =    .36128

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0063301   .0017273    -3.66   0.000    -.0097204   -.0029398
         age |   .0220385   .0040456     5.45   0.000     .0140979    .0299792
        educ |   .0363765   .0068217     5.33   0.000     .0229871    .0497659
     1.black |  -.2750052   .1036788    -2.65   0.008     -.478503   -.0715073
   1.married |   .1817023   .0438419     4.14   0.000     .0956506     .267754
             |
      black#|
     married |
        1 1  |    .153022   .1108707     1.38   0.168    -.0645919    .3706359
             |
          iq |   .0039649   .0010438     3.80   0.000     .0019162    .0060137
       meduc |   .0104632   .0047967     2.18   0.029     .0010483    .0198781
       south |  -.0729291   .0276262    -2.64   0.008    -.1271531   -.0187051
       urban |    .179466   .0280946     6.39   0.000     .1243228    .2346092
       _cons |   5.084363   .1866763    27.24   0.000     4.717961    5.450766
------------------------------------------------------------------------------
```

Quick Exercise

Run a regression with an interaction between urban and south. Interpret the results.

Continuous-Binary Interaction

Let's say we're interested in whether education affects wages differently for black and white men. If possible, we should start by plotting the data to see if these patterns are evident.

```
. gen educ_adj=educ+.2

. graph twoway (scatter wage educ if black==0, msize(small) mcolor(red)) ///
     (scatter wage educ_adj if black==1, msize(small) mcolor(blue))   ///
         (lfit wage educ if black==0, lcolor(red))  ///
             (lfit wage educ if black==1, lcolor(blue)), ///
                        legend(order(1 "White" 2 "Black"))
```

```
. graph twoway (scatter wage educ if married==0, msize(small) mcolor(red)) ///
     (scatter wage educ_adj if married==1, msize(small) mcolor(blue))   ///
         (lfit wage educ if married==0, lcolor(red))  ///
             (lfit wage educ if married==1, lcolor(blue)), ///
                 legend(order(1 "Unmarried" 2 "Married"))
```

```
. graph export interact1.pdf, replace
(file /Users/doylewr/lpo_prac/lessons/s2-06-model_design/interact1.pdf written in PDF
```

The specification of an interaction between a binary variable and a continous variable would look like this:

****/

```
. eststo black_educ: reg lwage hours age i.black##c.educ married  iq meduc south urban

      Source |       SS           df       MS      Number of obs   =       857
-------------+----------------------------------   F(10, 846)      =     29.69
       Model |  38.7974052        10  3.87974052   Prob > F        =    0.0000
    Residual |  110.563646       846  .130689889   R-squared       =    0.2598
-------------+----------------------------------   Adj R-squared   =    0.2510
       Total |  149.361051       856  .174487209   Root MSE        =    .36151

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0063045     .00173    -3.64   0.000    -.0097001    -.002909
         age |   .0215072   .0040455     5.32   0.000     .0135669    .0294476
     1.black |   .1034702   .2766783     0.37   0.709    -.4395862    .6465265
        educ |   .0378404   .0069883     5.41   0.000     .0241239    .0515569
             |
black#c.educ |
          1  |  -.0196117   .0215854    -0.91   0.364    -.0619789    .0227554
             |
     married |   .2051596   .0402851     5.09   0.000     .1260892      .28423
          iq |   .0040191   .0010442     3.85   0.000     .0019695    .0060687
       meduc |   .0102747   .0047964     2.14   0.032     .0008604    .0196889
       south |   -.069586   .0276172    -2.52   0.012    -.1237922   -.0153798
       urban |    .179133   .0281112     6.37   0.000     .1239572    .2343088
       _cons |    5.05532    .187864    26.91   0.000     4.686586    5.424055
------------------------------------------------------------------------------
```

```
. test 1.black educ 1.black#c.educ

 ( 1)  1.black = 0
 ( 2)  educ = 0
 ( 3)  1.black#c.educ = 0

       F(  3,   846) =   12.94
            Prob > F =    0.0000
```

Interactions with two continuous variables

Finally let's say we think that eductation will affect your wages differently depending on your age. The specification of an interaction between two continous variables would look like this:

```
. eststo age_educ : reg lwage hours age educ c.age#c.educ black married iq meduc south

      Source |       SS           df       MS      Number of obs   =       857
-------------+----------------------------------   F(10, 846)      =     30.08
       Model |  39.1769054        10  3.91769054   Prob > F        =    0.0000
    Residual |  110.184146       846  .130241307   R-squared       =    0.2623
-------------+----------------------------------   Adj R-squared   =    0.2536
       Total |  149.361051       856  .174487209   Root MSE        =    .36089

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0066151   .0017294    -3.83   0.000    -.0100095   -.0032208
         age |  -.0280105   .0260076    -1.08   0.282    -.0790575    .0230364
        educ |  -.0876756   .0645406    -1.36   0.175    -.2143541    .0390029
             |
c.age#c.educ |   .0037019   .0019136     1.93   0.053    -.0000542    .0074579
             |
       black |  -.1466181   .0430343    -3.41   0.001    -.2310847   -.0621515
     married |   .2063299   .0402134     5.13   0.000     .1274003    .2852596
          iq |   .0040003   .0010423     3.84   0.000     .0019545    .0060461
       meduc |   .0100804   .0047844     2.11   0.035     .0006897     .019471
       south |   -.06698    .0276104    -2.43   0.015    -.121173    -.012787
       urban |   .182844     .02813      6.50   0.000     .1276312    .2380569
       _cons |   6.747921   .8848456     7.63   0.000     5.011171    8.484672
------------------------------------------------------------------------------
```

```
. estimates replay black_marry

----------------------------------------------------------------------------------------
Model black_marry
----------------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =       857
-------------+----------------------------------   F(10, 846)      =     29.83
       Model |  38.9381574        10  3.89381574   Prob > F        =    0.0000
    Residual |  110.422894       846  .130523515   R-squared       =    0.2607
-------------+----------------------------------   Adj R-squared   =    0.2520
       Total |  149.361051       856  .174487209   Root MSE        =    .36128

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0063301   .0017273    -3.66   0.000    -.0097204   -.0029398
         age |   .0220385   .0040456     5.45   0.000     .0140979    .0299792
        educ |   .0363765   .0068217     5.33   0.000     .0229871    .0497659
     1.black |  -.2750052   .1036788    -2.65   0.008     -.478503   -.0715073
   1.married |   .1817023   .0438419     4.14   0.000     .0956506     .267754
             |
      black#|
    married |
        1 1  |    .153022   .1108707     1.38   0.168    -.0645919    .3706359
             |
          iq |   .0039649   .0010438     3.80   0.000     .0019162    .0060137
       meduc |   .0104632   .0047967     2.18   0.029     .0010483    .0198781
       south |  -.0729291   .0276262    -2.64   0.008    -.1271531   -.0187051
       urban |    .179466   .0280946     6.39   0.000     .1243228    .2346092
       _cons |   5.084363   .1866763    27.24   0.000     4.717961    5.450766
------------------------------------------------------------------------------

. estimates restore black_marry
(results black_marry are active now)

. local mydf=e(df_r)

. quietly margins , predict(xb) at(black=(1 0) married=(0 1) south=1 urban=1 (mean) ho

. mat mypred=e(b)'

. mata: st_matrix("mypred", exp(st_matrix("mypred")))
```

```
. svmat mypred

. mat mypred1=e(b)'

. svmat mypred1

. local no_predict=rowsof(mypred)

. di "no of preds is `no_predict'"
no of preds is 4

. egen mycount=fill(1(1)`no_predict')

. graph twoway bar mypred1 mycount in 1/4, barw(.6) base(0) ytick(100(100)1000) ylabel


. estimates restore black_marry
(results black_marry are active now)

. quietly margins , predict(stdp) at((mean) _all black=(1 0) married=(0 1) south=1 urb

. mat mystdp=e(b)'

. svmat mystdp

. gen ub_log=mypred11+ (invttail(`mydf',`sigtail')*mystdp)
(931 missing values generated)

. gen lb_log=mypred11- (invttail(`mydf',`sigtail')*mystdp)
(931 missing values generated)

. gen ub=exp(ub_log)
(931 missing values generated)

. gen lb=exp(lb_log)
(931 missing values generated)

. graph twoway (bar mypred1 mycount if mycount==1, barw(.6) base(0) ytick(100(100)1000
               (bar mypred1 mycount if mycount==2, barw(.6) base(0) ytick(100(100)1000
                (bar mypred1 mycount if mycount==3, barw(.6) base(0) ytick(100(100)100
                 (bar mypred1 mycount if mycount==4, barw(.6) base(0) ytick(100(100)10
                     (rcap ub lb mycount in 1/`no_predict'), ///
                         xlabel(1 "Unmarried, Black" 2 "Married, Black" 3 "Unmarried,
```

```
. drop mypred*

. drop mystdp*

. drop ub*

. drop lb*

. drop mycount

. eststo urban_south: reg lwage hours age educ black married iq meduc i.south##i.urban

      Source |       SS           df       MS      Number of obs   =       857
-------------+----------------------------------   F(10, 846)      =     30.26
       Model |  39.3509478        10  3.93509478   Prob > F        =    0.0000
    Residual |  110.010103       846  .130035583   R-squared       =    0.2635
-------------+----------------------------------   Adj R-squared   =    0.2548
       Total |  149.361051       856  .174487209   Root MSE        =     .3606

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0064086   .0017237    -3.72   0.000    -.0097919   -.0030252
         age |   .0223906   .0040422     5.54   0.000     .0144567    .0303244
        educ |   .0366152   .0068088     5.38   0.000     .0232511    .0499793
       black |  -.1413414   .0430185    -3.29   0.001     -.225777   -.0569058
     married |   .2032162   .0401949     5.06   0.000     .1243228    .2821096
          iq |   .0041836   .0010445     4.01   0.000     .0021334    .0062338
       meduc |   .0111274   .0048022     2.32   0.021     .0017018     .020553
     1.south |   .0219466   .0494206     0.44   0.657    -.0750549     .118948
     1.urban |   .2284794   .0355802     6.42   0.000     .1586435    .2983153
             |
 south#urban |
        1 1  |  -.1307251   .0579628    -2.26   0.024    -.2444928   -.0169573
             |
       _cons |   4.985266   .1903604    26.19   0.000     4.611632      5.3589
------------------------------------------------------------------------------


. local mydf=e(df_r)

. quietly margins , predict(xb) at(urban=(1 0) south=(0 1) black=1 married=1 (mean) ho

. mat mypred=e(b)'

. mata: st_matrix("mypred", exp(st_matrix("mypred")))
```

```
. svmat mypred

. mat mypred1=e(b)'

. svmat mypred1

. local no_predict=rowsof(mypred)

. di "no of preds is `no_predict'"
no of preds is 4

. egen mycount=fill(1(1)`no_predict')

. graph twoway bar mypred1 mycount in 1/4, barw(.6) base(0) ytick(100(100)1000) ylabel


. estimates restore black_marry
(results black_marry are active now)

. quietly margins , predict(stdp) at((mean) _all urban=(1 0) south=(0 1) black=1 marri

. mat mystdp=e(b)'

. svmat mystdp

. gen ub_log=mypred11+ (invttail(`mydf',`sigtail')*mystdp)
(931 missing values generated)

. gen lb_log=mypred11- (invttail(`mydf',`sigtail')*mystdp)
(931 missing values generated)

. gen ub=exp(ub_log)
(931 missing values generated)

. gen lb=exp(lb_log)
(931 missing values generated)

. graph twoway (bar mypred1 mycount if mycount==1, barw(.6) base(0) ytick(100(100)1000
               (bar mypred1 mycount if mycount==2, barw(.6) base(0) ytick(100(100)1000
                (bar mypred1 mycount if mycount==3, barw(.6) base(0) ytick(100(100)100
                 (bar mypred1 mycount if mycount==4, barw(.6) base(0) ytick(100(100)10
                     (rcap ub lb mycount in 1/`no_predict'), ///
                         xlabel(1 "Urban, Non-South" 2 "Urban, South" 3 "Non-Urban, No
```

```
. drop mypred*

. drop mystdp*

. drop ub*

. drop lb*

. sum age, detail

                             Age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           28             28
 5%           29             28
10%           29             28       Obs                 935
25%           30             28       Sum of Wgt.         935

50%           33                      Mean           33.08021
                        Largest       Std. Dev.      3.107803
75%           36             38
90%           38             38       Variance       9.658441
95%           38             38       Skewness       .1185453
99%           38             38       Kurtosis       1.743208

. local mymin=r(min)

. local mymax=r(max)

. foreach myeduc of numlist 10(2)16{
. estimates restore age_educ
.  quietly margins, predict(xb) at((mean) _all age=(`mymin'(1)`mymax') educ=`myeduc')
. mat pred_ed`myeduc'=e(b)'
. svmat pred_ed`myeduc'
. estimates restore age_educ
. quietly margins, predict(stdp) at((mean) _all age=(`mymin'(1)`mymax') educ=`myeduc')
. mat pred_se_ed`myeduc'=e(b)'
. svmat pred_se_ed`myeduc'
. }
(results age_educ are active now)
(results age_educ are active now)
(results age_educ are active now)
(results age_educ are active now)
(results age_educ are active now)
```

```
(results age_educ are active now)
(results age_educ are active now)
(results age_educ are active now)

. foreach myeduc of numlist 10(2)16{
.     gen exp_pred`myeduc'=exp(pred_ed`myeduc'1)
.     gen ub`myeduc'=exp(pred_ed`myeduc'+(invttail(`mydf',`sigtail')*pred_se_ed`myeduc
.     gen lb`myeduc'=exp(pred_ed`myeduc'-(invttail(`mydf',`sigtail')*pred_se_ed`myeduc
. }
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)
(924 missing values generated)

. egen age_levels=fill(`mymin'(1)`mymax')

. twoway line exp_pred10 exp_pred12 exp_pred14 exp_pred16 age_levels in 1/11, ///
        legend(order(1 "10 Years" 2 "12 Years" 3 "14 Years" 4 "16 Years"))  ytitle("Wa



. twoway (rarea ub10 lb10 age_levels in 1/11, color(gs14)) ///
     (rarea ub16 lb16 age_levels in 1/11, color(gs14)) ///
         (line exp_pred10 age_levels in 1/11, lcolor(blue) ) ///
             (line lb10 age_levels in 1/11, lcolor(blue) lwidth(thin) lpattern(dash))
                 (line ub10 age_levels in 1/11, lcolor(blue) lwidth(thin) lpattern(das
                     (line exp_pred16 age_levels in 1/11, lcolor(red) ) ///
                         (line ub16 age_levels in 1/11, lcolor(red) lwidth(thin) lpatt
                             (line lb16 age_levels in 1/11, lcolor(red) lwidth(thin) l
                                 legend(order( 3 "Less than HS" 6 "College Grad"))  xt




. drop lb* ub* exp_pred* pred* pred_ed* pred_se_ed*

. eststo ten_educ: reg lwage hours age black married iq meduc c.tenure##c.educ

      Source |       SS           df       MS      Number of obs   =       857
```

```
-------------+----------------------------------   F(9, 847)       =     27.24
       Model |  33.5243055         9  3.72492283   Prob > F        =    0.0000
    Residual |  115.836746       847  .136761211   R-squared       =    0.2245
-------------+----------------------------------   Adj R-squared   =    0.2162
       Total |  149.361051       856  .174487209   Root MSE        =    .36981

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0059456   .0017729    -3.35   0.001    -.0094254   -.0024658
         age |   .0185637   .0042976     4.32   0.000     .0101285    .0269989
       black |  -.1269117   .0433577    -2.93   0.004    -.2120129   -.0418105
     married |   .1838169   .0412587     4.46   0.000     .1028355    .2647982
          iq |   .0043803   .0010608     4.13   0.000     .0022983    .0064623
       meduc |   .0118471   .0048859     2.42   0.016     .0022572    .0214371
      tenure |   .0056798   .0165314     0.34   0.731    -.0267674    .0381271
        educ |   .0375757   .0115939     3.24   0.001     .0148194    .0603319
             |
    c.tenure#|
      c.educ |   .0002596   .0012336     0.21   0.833    -.0021617    .0026809
             |
       _cons |    5.14412   .2228437    23.08   0.000     4.706729     5.58151
------------------------------------------------------------------------------


. sum educ, detail

                     Years of education
-------------------------------------------------------------
      Percentiles      Smallest
 1%            9              9
 5%           11              9
10%           12              9       Obs                 935
25%           12              9       Sum of Wgt.         935

50%           12                      Mean           13.46845
                        Largest       Std. Dev.      2.196654
75%           16             18
90%           17             18       Variance       4.825288
95%           18             18       Skewness       .5477959
99%           18             18       Kurtosis       2.262651

. local mymin=r(min)

. local mymax=r(max)
```

```
. foreach mytenure of numlist 0(5)20{
.     estimates restore ten_educ
.  quietly   margins, predict(xb) at((mean) _all educ=(`mymin'(1)`mymax') tenure=`myte
.     mat pred_ed`mytenure'=e(b)'
.    svmat pred_ed`mytenure'
. estimates restore ten_educ
. quietly margins, predict(stdp) at((mean) _all educ=(`mymin'(1)`mymax') tenure=`myten
.     mat pred_se_ed`mytenure'=e(b)'
.     svmat pred_se_ed`mytenure'
. }
(results ten_educ are active now)
(results ten_educ are active now)
(results ten_educ are active now)
(results ten_educ are active now)
(results ten_educ are active now)
(results ten_educ are active now)
(results ten_educ are active now)
(results ten_educ are active now)
(results ten_educ are active now)
(results ten_educ are active now)

. foreach mytenure of numlist 0(5)20{
.     gen exp_pred`mytenure'=exp(pred_ed`mytenure'1)
.     gen ub`mytenure'=exp(pred_ed`mytenure'+(invttail(`mydf',`sigtail')*pred_se_ed`my
.     gen lb`mytenure'=exp(pred_ed`mytenure'-(invttail(`mydf',`sigtail')*pred_se_ed`my
. }
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)

. egen educ_levels=fill(`mymin'(1)`mymax')

. twoway line exp_pred0 exp_pred5 exp_pred10 exp_pred15 exp_pred15 educ_levels in 1/10
       legend(order(1 "0 Years" 2 "5 Years" 3 "10 Years" 4 "15 Years" 5 "20 Years"))
```

```
. eststo age_iq : reg lwage hours age iq c.iq#c.age black married meduc south urban

      Source |       SS           df       MS      Number of obs   =       857
-------------+----------------------------------   F(9, 847)       =     28.99
       Model |  35.1762546         9  3.90847273   Prob > F        =    0.0000
    Residual |  114.184796       847  .134810858   R-squared       =    0.2355
-------------+----------------------------------   Adj R-squared   =    0.2274
       Total |  149.361051       856  .174487209   Root MSE        =    .36717

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0060656   .0017613    -3.44   0.001    -.0095225   -.0026086
         age |   -.014751   .0297362    -0.50   0.620    -.0731163    .0436144
          iq |  -.0056246   .0095494    -0.59   0.556    -.0243679    .0131188
             |
  c.iq#c.age |   .0003674    .000288     1.28   0.202    -.0001978    .0009327
             |
       black |  -.1364958   .0438374    -3.11   0.002    -.2225385   -.0504531
     married |   .1935177   .0408545     4.74   0.000     .1133298    .2737057
       meduc |   .0162645   .0047216     3.44   0.001      .006997    .0255319
       south |  -.0631237   .0280021    -2.25   0.024    -.1180853   -.0081621
       urban |   .1892403   .0285395     6.63   0.000     .1332239    .2452567
       _cons |    6.43898   .9981104     6.45   0.000      4.47992     8.39804
------------------------------------------------------------------------------


. sum iq, detail

                           IQ test
-------------------------------------------------------------
      Percentiles      Smallest
 1%           64             50
 5%           74             54
10%           82             55       Obs                 935
25%           92             59       Sum of Wgt.         935

50%          102                      Mean           101.2824
                        Largest       Std. Dev.      15.05264
75%          112            134
90%          120            134       Variance       226.5819
95%          125            137       Skewness      -.3404246
99%          132            145       Kurtosis       2.977035
```

```
. sum iq, detail

                           IQ test
-------------------------------------------------------------
      Percentiles      Smallest
 1%           64             50
 5%           74             54
10%           82             55       Obs                 935
25%           92             59       Sum of Wgt.         935

50%          102                      Mean           101.2824
                        Largest       Std. Dev.      15.05264
75%          112            134
90%          120            134       Variance       226.5819
95%          125            137       Skewness      -.3404246
99%          132            145       Kurtosis       2.977035

. scalar iqlo=round(r(p10))

. scalar iqhi=round(r(p90))

. local iqhi=round(r(p90))

. scalar diff=iqhi-iqlo

. scalar step=round(diff/10)

. local iqlo=iqlo

. local iqhi=iqhi

. local step=step

. foreach myiq of numlist `iqlo'(`step')`iqhi'{
. estimates restore age_iq
. quietly margins, predict(xb) at((mean) _all age=(`mymin'(1)`mymax') iq=`myiq') post
. mat pred_ed`myiq'=e(b)'
. svmat pred_ed`myiq'
. estimates restore age_iq
. quietly margins, predict(stdp) at((mean) _all age=(`mymin'(1)`mymax') iq=`myiq') nos
. mat pred_se_ed`myiq'=e(b)'
. svmat pred_se_ed`myiq'
. }
(results age_iq are active now)
(results age_iq are active now)
```

```
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)
(results age_iq are active now)

. foreach myiq of numlist `iqlo'(`step')`iqhi'{
.     gen exp_pred`myiq'=exp(pred_ed`myiq'1)
.     gen ub`myiq'=exp(pred_ed`myiq'+(invttail(`mydf',`sigtail')*pred_se_ed`myiq'1))
.     gen lb`myiq'=exp(pred_ed`myiq'-(invttail(`mydf',`sigtail')*pred_se_ed`myiq'1))
. }
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
```

```
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)
(925 missing values generated)

. local iqhi=118

. twoway line exp_pred`iqlo' exp_pred90 exp_pred102 exp_pred`iqhi' age_levels in 1/11,
       legend(order(1 "10 Years" 2 "12 Years" 3 "14 Years" 4 "16 Years"))  ytitle("Wa



.     twoway (rarea ub`iqlo' lb`iqlo' age_levels in 1/11, color(gs14)) ///
                 (line exp_pred`iqlo' age_levels in 1/11, lcolor(blue) ) ///
                     (line ub`iqlo' age_levels in 1/11, lcolor(blue) lwidth(thin) lpat
                         (line lb`iqlo' age_levels in 1/11, lcolor(blue) lwidth(thin)
                             (rarea ub`iqhi' lb`iqhi' age_levels in 1/11, color(gs14))
                                  (line exp_pred`iqhi' age_levels in 1/11, lcolor(red)
                                      (line ub`iqhi' age_levels in 1/11, lcolor(red) l
                                          (line lb`iqhi' age_levels in 1/11, lcolor(re
                                              legend(order( 2 "10th Percentile" 6 "90t
```

Not-So-Quick Exercise

In pairs, I would like you to estimate the best possible model using the wage2 dataset. Think about model specification and functional form, with an eye toward possible non-linearities and other issues. Generate a do file that walks through your process of identifying the best model. Generate a fancy graph that shows the predictions made by your model.

```
. exit
```