PPOL502-01, Spring 2016
Course Notes # 13: Functional Form: Putting it all Together, F-tests, Ramsey's RESET test

---

I. FUNCTIONAL FORM OVERVIEW

- "Functional form" refers to, well, the functional form of the variables in the model: are they in level form? in natural log? Squared? cubed? square root? Indicators? Interactions? something else?

- And the double whammy: when a particular form of a variable should be in a model but is not, the problem is referred to as "functional form misspecification." This is like an omitted variable problem.

How to make functional form decisions? Again, this is both art and science (sorry)

- Here too, base your decisions on theory, prior work in the area, and your own noggin.

- These decisions will be driven by the estimation problems that you're faced with, and the variables in the model that you care about the most. As Wooldridge points out (bottom p. 305), "***It can be difficult to pinpoint the precise reason that a functional form is misspecified. Fortunately, in many cases, using logarithms of certain variables and adding quadratics is sufficient for detecting many important nonlinear relationships in economics.***"

- The ultimate criterion often is to use the *simplest* model possible for your purposes (i.e., you don't get extra points for being a good person if you log or square or whatever everything in the model).

- O.k., so how about some specific strategies….

(1) The importance of looking at how people before you have specified their dependent and independent variables cannot be overemphasized. We hate to break it to you, but we are unlikely to be breaking new ground for functional form in our work. We can rely on previous work (checked by our own good sense) to get some guidelines about how to proceed.

(2) Adjusted R-squared (which we have already talked about) provides some information: does adding higher order terms, or transforming one or more variables into logs, increase the adjusted R-squared? i.e., is the explanatory power of the model higher with different forms of the model?

-- Does a relatively low R-squared or adjusted R-squared indicate that the model is "wrong"? -- no! As long as the assumptions of the model are met (and remember, they are not automatically met if OLS runs), then the model is o.k. See earlier notes on R-squared for further discussion of high/low R-squared.

-- Can R-squared or adjusted R-squared be helpful at all for selecting and specifying regressors for a model? We know that regular R-squared increases (or at least doesn't decrease) when additional variables are added to a model (using the same sample). Thus, looking at regular R-squared is probably not all that useful for making model specification decisions.

-- Looking at adjusted R-squared can be useful, but it *isn't* the be-all-and-end-all (an example below).

Advantages: low-tech, easy to implement and check
Disadvantages: not grounded in theory; doesn't always provide good guidance on what's going on.

Example:

		Adj R-squared
reg wage	IQ exper educ KWW momdad14 sinmom14 step14 libcrd14	0.1863
reg wage	IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14	0.1979
reg lwage	IQ exper educ KWW momdad14 sinmom14 step14 libcrd14	0.1839
reg lwage	IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14	0.1930

Q: Based on these specifications and these adjusted R-squared figures, which model would you prefer? Why?

(3) When you are thinking about including higher-order terms for one or more variables, you can estimate the model with squared (and cubed, etc) forms of the X variable(s), and then use F-tests to test joint exclusion restrictions on these higher-order terms.

Advantages: the inclusion of higher order terms can be guided by theory (i.e., you do not have to square every single variable in the model).

Drawbacks: What if you really don't know what variables are driving the nonlinearities? you can use up many degrees of freedom by throwing in lots of higher order terms.

Example

```
. use crime1.dta
.
. describe narr86 pcnv pcnvsq avgsen tottime ptime86 pt86sq qemp86 inc86 inc86sq black
hispan

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------------
narr86          byte   %9.0g                  # times arrested, 1986
pcnv            float  %9.0g                  proportion of prior convictions
pcnvsq          float  %9.0g                  pcnv^2
avgsen          float  %9.0g                  avg sentence length, mos.
tottime         float  %9.0g                  time in prison since 18 (mos.)
ptime86         byte   %9.0g                  mos. in prison during 1986
pt86sq          int    %9.0g                  ptime86^2
qemp86          float  %9.0g                  # quarters employed, 1986
inc86           float  %9.0g                  legal income, 1986, $100s
inc86sq         float  %9.0g                  inc86^2
black           byte   %9.0g                  =1 if black
hispan          byte   %9.0g                  =1 if Hispanic


. reg narr86 pcnv pcnvsq avgsen tottime ptime86 pt86sq qemp86 inc86 inc86sq black
hispan

      Source |       SS       df       MS              Number of obs =    2725
-------------+------------------------------           F( 11,  2713) =   28.46
       Model |  207.979008    11  18.9071826           Prob > F      =  0.0000
    Residual |  1802.36815  2713   .66434506           R-squared     =  0.1035
-------------+------------------------------           Adj R-squared =  0.0998
       Total |  2010.34716  2724  .738012906           Root MSE      =  .81507

------------------------------------------------------------------------------
      narr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |   .5525236   .1542372     3.58   0.000     .2500892     .854958
      pcnvsq |  -.7302119   .1561177    -4.68   0.000    -1.036334   -.4240903
      avgsen |  -.0170216   .0120539    -1.41   0.158    -.0406574    .0066142
     tottime |    .011954   .0092825     1.29   0.198    -.0062474    .0301554
     ptime86 |   .2874334   .0442582     6.49   0.000     .2006501    .3742166
      pt86sq |  -.0296076   .0038634    -7.66   0.000     -.037183   -.0220321
      qemp86 |  -.0140941   .0173612    -0.81   0.417    -.0481366    .0199485
       inc86 |  -.0034152   .0008037    -4.25   0.000    -.0049912   -.0018392
     inc86sq |   7.19e-06   2.56e-06     2.81   0.005     2.17e-06    .0000122
       black |    .292296     .04483     6.52   0.000     .2043916    .3802004
      hispan |   .1636175   .0394507     4.15   0.000     .0862609     .240974
       _cons |   .5046065   .0368353    13.70   0.000     .4323784    .5768347
------------------------------------------------------------------------------

. test (pcnvsq=0) (pt86sq=0) (inc86sq=0)

 ( 1)  pcnvsq = 0
 ( 2)  pt86sq = 0
 ( 3)  inc86sq = 0

       F(  3,  2713) =   31.40
            Prob > F =    0.0000
```

(4) Ramsey's RESET (regression specification error test). This is a test that uses squares, cubes, etc. of *predicted Y* in the model:

Original model: $Y= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k + u$

Estimate this model, obtain the *predicted Y* value for each observation. Then, take the square and cube (in some cases also the 4th power) of that term, and re-estimate the model including these terms:

RESET model: $Y= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + error$

H0: $\delta_1 = 0$ *and* $\delta_2 = 0$

Run an F-test. If we fail to reject the null hypothesis, this indicates that there are likely no functional form misspecifications in the original model.

Advantages: Easy to implement; does not eat up a lot of degrees of freedom.
Drawbacks: Doesn't really tell you where the problem is in the original model. This detects only very general forms of functional form misspecification. You will still need to be guided by theory about how to proceed.

You can use the Stata postestimation command **estat ovtest** to implement this test:

Example

```
reg wage  IQ exper educ KWW momdad14 sinmom14 step14 libcrd14
estat ovtest
reg wage  IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14
estat ovtest
reg lwage IQ exper educ KWW momdad14 sinmom14 step14 libcrd14
estat ovtest
reg lwage IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14
estat ovtest
```

- To use **Estat ovtest,** first run the regression you would like to test, then run the test command. This command, like all postestimation commands, will test the most recent regression. Stata's implementation of this test includes predicted Y values up to the fourth power (see output for 4 models below).

- Among the following 4 models, Model 4 looks the best based on this series of tests. This model is consistent with prior work – that the wage variable is logged, and that the square of experience is included along with EXPER.

MODEL 1

```
. reg wage IQ exper educ KWW momdad14 sinmom14 step14 libcrd14

      Source |       SS       df       MS              Number of obs =    2034
-------------+------------------------------           F(  8,  2025) =   59.19
       Model |  26851193.8     8  3356399.22            Prob > F      =  0.0000
    Residual |   114828203  2025  56705.2856            R-squared     =  0.1895
-------------+------------------------------           Adj R-squared =  0.1863
       Total |   141679397  2033  69689.8166            Root MSE      =  238.13

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |      1.1586   .4313011     2.69   0.007     .3127599     2.00444
       exper |    23.85762   2.008143    11.88   0.000     19.91937    27.79586
        educ |    37.98313   3.497601    10.86   0.000     31.12385     44.8424
         KWW |    5.006964   .9102044     5.50   0.000     3.221929    6.791999
    momdad14 |    41.91382   24.68697     1.70   0.090    -6.500699    90.32834
    sinmom14 |    10.26134    30.2559     0.34   0.735    -49.07461    69.59728
      step14 |   -3.052065   36.36072    -0.08   0.933    -74.36039    68.25626
    libcrd14 |    27.62837    12.9856     2.13   0.033     2.161848    53.09489
       _cons |   -469.8669   60.62699    -7.75   0.000    -588.7647   -350.9691
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of wage
       Ho:  model has no omitted variables
                F(3, 2022) =     14.77
                  Prob > F =      0.0000
```

MODEL 2

```
. reg wage  IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14

      Source |       SS       df       MS              Number of obs =    2034
-------------+------------------------------           F(  9,  2024) =   56.72
       Model |  28536335.5     9  3170703.94            Prob > F      =  0.0000
    Residual |   113143062  2024  55900.7222            R-squared     =  0.2014
-------------+------------------------------           Adj R-squared =  0.1979
       Total |   141679397  2033  69689.8166            Root MSE      =  236.43

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |    1.334431   .4294262     3.11   0.002     .4922673    2.176594
       exper |    56.95279   6.348952     8.97   0.000     44.50163    69.40395
     expersq |   -1.751863   .3190734    -5.49   0.000    -2.377609   -1.126116
        educ |     40.6186   3.505717    11.59   0.000     33.74341    47.49379
         KWW |    4.460293   .9091925     4.91   0.000     2.677242    6.243344
    momdad14 |    44.49471   24.51572     1.81   0.070    -3.583971    92.57339
    sinmom14 |    10.51109   30.04053     0.35   0.726    -48.40249    69.42467
      step14 |    2.470041   36.11586     0.07   0.945    -68.35809    73.29817
    libcrd14 |    27.42929    12.8932     2.13   0.034     2.143971    52.71461
       _cons |   -637.5092    67.4964    -9.45   0.000    -769.8789   -505.1395
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of wage
       Ho:  model has no omitted variables
                F(3, 2021) =      8.52
                  Prob > F =      0.0000
```

MODEL 3

```
. reg lwage IQ exper educ KWW momdad14 sinmom14 step14 libcrd14

      Source |       SS       df       MS              Number of obs =    2034
-------------+------------------------------           F(  8,  2025) =   58.27
       Model |  66.4068172     8  8.30085214            Prob > F      =  0.0000
    Residual |  288.454656  2025  .142446744            R-squared     =  0.1871
-------------+------------------------------           Adj R-squared =  0.1839
       Total |  354.861474  2033  .174550651            Root MSE      =  .37742

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |   .0028359   .0006836     4.15   0.000     .0014953    .0041765
       exper |   .0366581   .0031828    11.52   0.000     .0304162       .0429
        educ |   .0524187   .0055435     9.46   0.000     .0415471    .0632903
         KWW |   .0083142   .0014426     5.76   0.000     .0054851    .0111434
    momdad14 |   .0681809   .0391275     1.74   0.082    -.0085535    .1449153
    sinmom14 |   .0198705    .047954     0.41   0.679    -.0741738    .1139148
      step14 |  -.0020144   .0576298    -0.03   0.972    -.1150343    .1110054
    libcrd14 |     .03896   .0205815     1.89   0.059     -.001403    .0793231
       _cons |   4.620621   .0960905    48.09   0.000     4.432175    4.809068
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of lwage
       Ho:  model has no omitted variables
                F(3, 2022) =      4.06
                  Prob > F =      0.0069
```

MODEL 4

```
. reg lwage IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14

      Source |       SS       df       MS              Number of obs =    2034
-------------+------------------------------           F(  9,  2024) =   55.01
       Model |  69.7397877     9   7.7488653            Prob > F      =  0.0000
    Residual |  285.121686  2024  .140870398            R-squared     =  0.1965
-------------+------------------------------           Adj R-squared =  0.1930
       Total |  354.861474  2033  .174550651            Root MSE      =  .37533

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |   .0030832   .0006817     4.52   0.000     .0017463    .0044201
       exper |    .083202   .0100787     8.26   0.000     .0634363    .1029676
     expersq |  -.0024638   .0005065    -4.86   0.000    -.0034571   -.0014704
        educ |   .0561251   .0055652    10.09   0.000     .0452111    .0670392
         KWW |   .0075454   .0014433     5.23   0.000     .0047149    .0103759
    momdad14 |   .0718106   .0389176     1.85   0.065    -.0045121    .1481333
    sinmom14 |   .0202218    .047688     0.42   0.672    -.0733008    .1137444
      step14 |   .0057517   .0573323     0.10   0.920    -.1066847    .1181881
    libcrd14 |   .0386801   .0204674     1.89   0.059    -.0014592    .0788193
       _cons |   4.384855   .1071474    40.92   0.000     4.174725    4.594986
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of lwage
       Ho:  model has no omitted variables
                F(3, 2021) =      1.54
                  Prob > F =      0.2034
```