Statistics/Data Analysis

```
      name:  <unnamed>
       log:  C:\Users\Conor\Documents\Conor\Grad School\TA Work\Econ 103 - Econometric
> s\STATA Work\Week 7\wk7_section_log.smcl
  log type:  smcl
 opened on:   6 Feb 2018, 12:18:37

. 
. // Demonstration STATA code for week 7
. // Principles of Econometrics 4th Edition
. // Covered Problems: 6.14
. 
. set more off

. clear all

. use hwage.dta, clear

. 
. ////////////////////////////////////////\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
> //////////////////////////////// Question 6.14 \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
> ////////////////////////////////////////\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
> 
. ********************************************************************************
. *Setup: In the context of a wage regression, use the RESET framework to conduct
. * a model selection exercise.
. *
. * Parts (A) - (G)
. ********************************************************************************
. 
. ********************************************************************************
. *Part A: Estimate the model:
. *
. * HW = beta1 + beta2*HE + beta3*HA+e
. *
. * What effects do changes in the level of education and age have on wages?
. ********************************************************************************
. 
. reg hw he ha
```

Source	SS	df	MS
Model	31825.8982	2	15912.9491
Residual	160479.81	750	213.97308
Total	192305.708	752	255.725676

Number of obs	=	753
F(2, 750)	=	74.37
Prob > F	=	0.0000
R-squared	=	0.1655
Adj R-squared	=	0.1633
Root MSE	=	14.628

hw	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
he	2.193289	.1800506	12.18	0.000	1.839826	2.546752
ha	.1996641	.0674912	2.96	0.003	.06717	.3321583
_cons	-8.123578	4.158325	-1.95	0.051	-16.28692	.039763

```
. 
. /* Discussion:
> 
> Based on the estimated coefficients, we would say:
> 
> beta2: An increase of 1 year in a husband's education would be expected to raise
>        the husband's wage by 2.19 in 2006 dollars
> 
> beta3: An increase of 1 year in a husband's age would be expected to raise the
>        husband's wage by 0.20 in 2006 dollars
> 
> */
```

```
29.
30. ******************************************************************************
31. *Part B: Does RESET suggest that the model in part (a) is adequate?
32. ******************************************************************************
33.
34. // To run the RESET test we:
35. // (1) calcuate the square and the cube of the fitted values
36. // (2) Run the auxillary regression:
37. //     hw = beta1 + beta2*he + beta3*ha + gamma1*hw_hat^2 + e
38. //     hw = beta1 + beta2*he + beta3*ha + delta1*hw_hat^2 + delta2*hw_hat^3 + e
39. //
40. // (3) Test (a) H0: gamma1 = 0, and (b) H0: delta1 = 0 and delta2 = 0
41. predict fithw_a, xb

42. gen fithw_a2 = fithw_a^2

43. gen fithw_a3 = fithw_a^3

44.
45. reg hw he ha fithw_a2

      Source |       SS           df       MS      Number of obs   =       753
-------------+----------------------------------   F(3, 749)       =     53.32
       Model |  33841.6935         3  11280.5645   Prob > F        =    0.0000
    Residual |  158464.015       749  211.567443   R-squared       =    0.1760
-------------+----------------------------------   Adj R-squared   =    0.1727
       Total |  192305.708       752  255.725676   Root MSE        =    14.545

------------------------------------------------------------------------------
          hw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          he |  -1.472529   1.201025    -1.23   0.221    -3.830304    .8852459
          ha |  -.1575674   .1337819    -1.18   0.239    -.4201995    .1050647
    fithw_a2 |   .0302375    .009796     3.09   0.002     .0110067    .0494683
       _cons |   28.32094   12.50994     2.26   0.024     3.762224    52.87965
------------------------------------------------------------------------------

46. // To test gamma1 = 0 can just look at the p-value for the fithw_a2 coefficient
47. qui reg hw he ha fithw_a2 fithw_a3

48. test (_b[fithw_a2]=0) (_b[fithw_a3]=0)

 ( 1)  fithw_a2 = 0
 ( 2)  fithw_a3 = 0

       F(  2,   748) =    4.79
            Prob > F =    0.0086

49.
50. /* Discussion:
  >
  > For both auxillary regressions, we are able to reject the null that the powers
  > of the fitted hw values have no effect in the regression. This suggests that the
  > model is mis-specified and we should explore adding in higher powers of the
  > RHS variables. We will do this in Part (C) below.
  >
  > */
51.
52. // STATA comment:
53. // STATA also has a built-in function to do the RESET test using the 2nd, 3rd,
```

```
54. // and 4th powers of the fitted value. Compare the F-stats reported using
55. // estat ovtest following reg hw he ha, and the f-test for the coefficients on
56. // fithw_a2 fithw_a3 and fithw_a4 in the second regression below. The textbook
57. // authors argue that the 4th power test is not recommend because adding more
58. // polynomial terms reduces the power of the test (i.e. increases rate of
59. // failure to reject the null even when the null is false).
60. qui reg hw he ha

61. estat ovtest

  Ramsey RESET test using powers of the fitted values of hw
         Ho:  model has no omitted variables
                   F(3, 747) =      6.65
                    Prob > F =      0.0002

62.
63. gen fithw_a4 = fithw_a^4

64. qui reg hw he ha fithw_a2 fithw_a3 fithw_a4

65. test (_b[fithw_a2]=0) (_b[fithw_a3]=0) (_b[fithw_a4]=0)

 ( 1)  fithw_a2 = 0
 ( 2)  fithw_a3 = 0
 ( 3)  fithw_a4 = 0

       F(  3,   747) =    6.65
            Prob > F =    0.0002

66.
67. ******************************************************************************
68. *Part C: Add the variables HE^2 and HA^2 to the original equation and
69. * re-estimate it. Describe the effect that education and age have on wages in
70. * this newly estimated model.
71. ******************************************************************************
72.
73. reg hw he ha c.he#c.he c.ha#c.ha
```

Source	SS	df	MS
Model	36876.7034	4	9219.17584
Residual	155429.005	748	207.792787
Total	192305.708	752	255.725676

Number of obs	=	753
F(4, 748)	=	44.37
Prob > F	=	0.0000
R-squared	=	0.1918
Adj R-squared	=	0.1874
Root MSE	=	14.415

hw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
he	-1.457971	1.122786	-1.30	0.195	-3.662157	.7462154
ha	2.889541	.7328868	3.94	0.000	1.450781	4.328301
c.he#c.he	.1511426	.0458277	3.30	0.001	.0611762	.2411089
c.ha#c.ha	-.0301212	.0081339	-3.70	0.000	-.0460892	-.0141532
_cons	-45.56754	17.54364	-2.60	0.010	-80.00817	-11.12691

```
74.
75. disp "Education Tipping Point = beta2/(-2*beta4) = " _b[he]/(-2*_b[c.he#c.he])
  Education Tipping Point = beta2/(-2*beta4) = 4.823164
```

```
76. margins, dydx(he) at(he=(12 16))

  Average marginal effects                        Number of obs     =        753
  Model VCE     : OLS

  Expression    : Linear prediction, predict()
  dy/dx w.r.t. : he

  1._at         : he              =          12

  2._at         : he              =          16

  ------------------------------------------------------------------------------
               |            Delta-method
               |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
  he           |
           _at |
            1  |   2.169451   .1776556    12.21   0.000     1.820688    2.518214
            2  |   3.378591   .3993905     8.46   0.000     2.594531    4.162651
  ------------------------------------------------------------------------------

77.
78. /* Discussion - Education:
  >
  > The point estimate for beta2 (he) and beta4 (he^2) suggest that education has
  > an initially negative effect on wages, and then an positive effect with the
  > marginal effect getting larger. We can see this because beta2 is negative and
  > beta4 is positive. As the calculation above suggests, the model estimates that
  > the marginal effect of education is positive for years 5 and upwards. If we look
  > at the distribution of he, this indicates that marginal effect is positive for
  > almost all individuals in the sample. In addition, comparing the marginal effect
  > at he=12 to the simple linear estimate shows an effect of a similar magnitude,
  > while at higher levels of education the marginal return is higher than in the
  > linear model.
  >
  > */
79.
80. disp "Age Tipping Point = beta3/(-2*beta5) = " _b[ha]/(-2*_b[c.ha#c.ha])
  Age Tipping Point = beta3/(-2*beta5) = 47.96519

81. margins, dydx(ha) at(ha=(30 35 40 45 50 55))

  Average marginal effects                        Number of obs     =        753
  Model VCE     : OLS

  Expression    : Linear prediction, predict()
  dy/dx w.r.t. : ha

  1._at         : ha              =          30

  2._at         : ha              =          35

  3._at         : ha              =          40

  4._at         : ha              =          45

  5._at         : ha              =          50

  6._at         : ha              =          55
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ha           |
         _at |
          1  |   1.082267   .2508286     4.31   0.000     .5898554    1.574679
          2  |   .7810549   .1737602     4.50   0.000     .4399393    1.122171
          3  |   .4798426   .1034517     4.64   0.000     .2767523    .6829329
          4  |   .1786303   .0666634     2.68   0.008     .0477606       .3095
          5  |   -.122582   .1068543    -1.15   0.252    -.332352     .0871881
          6  |  -.4237943   .1778307    -2.38   0.017    -.7729009   -.0746876
------------------------------------------------------------------------------

82.
83. /* Discussion - Age:
  >
  > The parabolic shape is flipped for age, with an initially positive relationship
  > (beta3 > 0) between age and wage declining and eventually turning negative
  > (beta5 < 0). The relative magnitude of beta3 and beta5 is larger than that for
  > beta2 and beta4 (i.e. |beta3/beta5| > |beta2/beta4|) which lets us know that it
  > will take longer to reach the tipping point for age than it did for education.
  > It turns out that the tipping point for age is around 48, or a little after
  > the mean age in the dataset. The relatively small beta5 also tells us that the
  > marginal effect is changing more slowly than for education. The estimated
  > marginal effect at 45 (approximately the mean age) is similar to that seen in
  > the linear estimate from part (a).
  >
  > */
84.
85. ******************************************************************************
86. *Part D: Does RESET suggest that the model in pact (c) is adequate?
87. ******************************************************************************
88.
89. predict fithw_c, xb

90. gen fithw_c2 = fithw_c^2

91. gen fithw_c3 = fithw_c^3

92.
93. qui reg hw he ha c.he#c.he c.ha#c.ha fithw_c2

94. test (_b[fithw_c2]=0)

 ( 1)  fithw_c2 = 0

       F(  1,   747) =    0.33
            Prob > F =    0.5680

95.
96. qui reg hw he ha c.he#c.he c.ha#c.ha fithw_c2 fithw_c3

97. test (_b[fithw_c2]=0) (_b[fithw_c3]=0)

 ( 1)  fithw_c2 = 0
 ( 2)  fithw_c3 = 0

       F(  2,   746) =    0.88
            Prob > F =    0.4143
```

```
/* Discussion:
>
> In both regression, we fail to reject the null that the polynomials of the
> fitted values have no effect. The RESET test suggests the model is adequate
> with regards to including higher powers of the current set of RHS variables.
>
> */

*******************************************************************************
*Part E: Reestimate the model in part (c) with the variable CIT included. What
* can you say about the level of wages in large cities relative to outside those
* cities?
*******************************************************************************

reg hw he ha c.he#c.he c.ha#c.ha cit
```

Source	SS	df	MS
Model	**46984.2168**	**5**	**9396.84337**
Residual	**145321.492**	**747**	**194.540149**
Total	**192305.708**	**752**	**255.725676**

Number of obs	=	**753**
F(5, 747)	=	**48.30**
Prob > F	=	**0.0000**
R-squared	=	**0.2443**
Adj R-squared	=	**0.2393**
Root MSE	=	**13.948**

hw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
he	**-2.207574**	**1.091357**	**-2.02**	**0.043**	**-4.350066**	**-.0650814**
ha	**2.621256**	**.7101069**	**3.69**	**0.000**	**1.227214**	**4.015299**
c.he#c.he	**.1687597**	**.0444096**	**3.80**	**0.000**	**.0815773**	**.2559421**
c.ha#c.ha	**-.0277679**	**.007877**	**-3.53**	**0.000**	**-.0432316**	**-.0123042**
cit	**7.937853**	**1.101249**	**7.21**	**0.000**	**5.775942**	**10.09976**
_cons	**-37.05403**	**17.01601**	**-2.18**	**0.030**	**-70.45893**	**-3.649139**

```
/* Discussion:
>
> The interpretation of the beta for cit is that wages are about $7.9 (2006$)
> higher in large cities, on average, than are wages in other (non-large) cities.
>
> */

*******************************************************************************
*Part F: Do you think CIT should be included in the equation?
*******************************************************************************

/* Discussion:
>
> Yes, I would recommend including CIT in the regression. First, the coefficient
> for CIT is strongly significant. Second, its inclusion in the regression had
> a noticeable effect on the estimated coefficients for the education and age,
> with a particularly large effect for education. Third, including cit in the
> regression lead to smaller standard errors for all the other RHS beta estimates,
> indicating that it is adding significant new information to the regression.
>
> We discuss this effect on the other beta estimates in more detail in part (g).
>
> Below, we organize the beta estimates, standard errors, and t-statistics between
> the regressions with and without cit to make it easier to compare.
>
> */
```

```
// Run regressions and collect the beta estimates
//
// Note: the equations for standard errors and t-stats are somewhat complicated
// matrix algebra equations. You do not need to worry about understanding what
// is being done in those steps.
qui reg hw he ha c.he#c.he c.ha#c.ha

matrix beta_c = e(b)

matrix se_c = vecdiag(cholesky(diag(vecdiag(e(V)))))

matrix t_c = vecdiag(diag(beta_c)*inv(diag(se_c)))

qui reg hw he ha c.he#c.he c.ha#c.ha cit

matrix beta_f = e(b)

matrix se_f = vecdiag(cholesky(diag(vecdiag(e(V)))))

matrix t_f = vecdiag(diag(beta_f)*inv(diag(se_f)))


// Compile the beta and std error estimates into (2 x K) matricies, making sure
// that estimates for the same variables are in the same column
matrix beta_compare = [ [beta_c, J(1,1,.)] \ [beta_f[1,1..4], beta_f[1,6], beta_f[1,
> 5]] ]

matrix se_compare = [ [se_c, J(1,1,.)] \ [se_f[1,1..4], se_f[1,6], se_f[1,5]] ]

matrix t_compare = [ [t_c, J(1,1,.)] \ [t_f[1,1..4], t_f[1,6], t_f[1,5]] ]


// Add row and column names
local compareMats beta_compare se_compare t_compare

foreach x of local compareMats {
  2.         matrix rownames `x' = "without cit" "with cit"
  3.         matrix colnames `x' = "he" "ha" "heSqr" "haSqr" "cnst" "cit"
  4. }


// Compare betas:
matrix list beta_compare

beta_compare[2,6]
                     he           ha        heSqr        haSqr         cnst          cit
without cit  -1.4579706    2.889541   .15114255  -.03012123    -45.56754            .
   with cit  -2.2075739   2.6212563   .16875972  -.02776792   -37.054033    7.9378532

// Compare standard errors:
matrix list se_compare

se_compare[2,6]
                    he          ha       heSqr       haSqr        cnst         cit
without cit  1.1227857  .73288684  .04582774   .0081339  17.543637           .
   with cit  1.0913574  .71010685  .04440957  .00787702  17.016012    1.101249

// Compare t-stats:
matrix list t_compare

t_compare[2,6]
                     he           ha        heSqr        haSqr         cnst          cit
without cit  -1.2985298   3.9426837   3.2980578    -3.70317   -2.5973827            .
   with cit  -2.0227782   3.6913547   3.8000753  -3.5251817    -2.177598    7.2080459
```

```
****************************************************************************************
*Part G: For both the model estimated in part (c) and the model estimated in
* part (e), evaluate the following four derivatives:
*
* (i)  dHW/dHE for HE = 6 and HE = 15
* (ii) dHW/dHA for HA = 35 and HA = 50
*
* Does the omission of CIT lead to omitted-variable bias? Can you suggest why?
****************************************************************************************

// The models in part (c) and part (f) both have the same formula for the
// marginal effects:
//
// (i)  dHW/dHE = beta2 + 2*beta4*HE
// (ii) dHW/dHA = beta3 + 2*beta5*HA
//
// where beta2 is the term on HE, beta3 is the term on HA, beta4 is the term on
// HE^2, and beta5 is the term on HA^2
//
// The only difference between the marginal effects in (c) and (f) comes from
// the different estimates from beta2-beta5 that are calculated without or with
// the cit term.

// Calculate marginal effects for education (he)
matrix me_he = J(2,2,.)

matrix rownames me_he = "without cit" "with cit"

matrix colnames me_he = "he=6" "he=15"

matrix me_he[1,1] = beta_c[1,2-1]+2*beta_c[1,4-1]*6

matrix me_he[1,2] = beta_c[1,2-1]+2*beta_c[1,4-1]*15

matrix me_he[2,1] = beta_f[1,2-1]+2*beta_f[1,4-1]*6

matrix me_he[2,2] = beta_f[1,2-1]+2*beta_f[1,4-1]*15


// Calculate marginal effects for age (ha)
matrix me_ha = J(2,2,.)

matrix rownames me_ha = "without cit" "with cit"

matrix colnames me_ha = "ha=35" "he=50"

matrix me_ha[1,1] = beta_c[1,3-1]+2*beta_c[1,5-1]*35

matrix me_ha[1,2] = beta_c[1,3-1]+2*beta_c[1,5-1]*50

matrix me_ha[2,1] = beta_f[1,3-1]+2*beta_f[1,5-1]*35

matrix me_ha[2,2] = beta_f[1,3-1]+2*beta_f[1,5-1]*50


// View results:
//
// Marginal effects for HE
matrix list me_he

me_he[2,2]
                  he=6       he=15
without cit   .35573999   3.0763059
   with cit  -.18245732   2.8552176
```

```
193 disp "HE marginal effect - with CIT ME is lower if HE < " (-1)*(beta_compare[1,1]-be
  > ta_compare[2,1])/(2*(beta_compare[1,3]-beta_compare[2,3]))
  HE marginal effect - with CIT ME is lower if HE < 21.2748

194 // Marginal effects for HA
195 matrix list me_ha

  me_ha[2,2]
                    ha=35       he=50
  without cit   .78105492  -.12258199
     with cit   .67750202  -.15553551

196 disp "HA marginal effect - with CIT ME is lower if HA < " (-1)*(beta_compare[1,2]-be
  > ta_compare[2,2])/(2*(beta_compare[1,4]-beta_compare[2,4]))
  HA marginal effect - with CIT ME is lower if HA < 57.00152

197
198 // Look at correlation between cit an other RHS variables (bottom row)
199 gen heSqr = he^2

200 gen haSqr = ha^2

201 corr he ha heSqr haSqr cit
  (obs=753)

               |       he       ha    heSqr    haSqr      cit
  -------------+---------------------------------------------
            he |   1.0000
            ha |  -0.1953   1.0000
         heSqr |   0.9878  -0.1839   1.0000
         haSqr |  -0.1938   0.9960  -0.1823   1.0000
           cit |   0.2333   0.0676   0.2233   0.0639   1.0000


202
203 /* Discussion:
  >
  > The marginal effects fell for both education and age at the values specified.
  > We also show a calculation that says that for all levels of education in the
  > sample, the ME is lower in the CIT regression, while it is lower for almost all
  > ages in the sample (5% of observations have age > 57).
  >
  > As we saw earlier when we compared the betas, the estimates for the linear terms
  > shifted down, while the estimates for the quadratic (X^2) terms shifted up. The
  > shift was larger for the linear term for both HE and HA, while coefficient
  > changes were larger for HE than for HA.
  >
  > Given these changes, we would say that without including cit the estimates for
  > HE and HA were biased UPWARDS and the estimates for HE^2 and HA^2 were biased
  > DOWNWARDS.
  >
  > The presence of omitted variable bias tells us that (1) cit is correlated with
  > the RHS terms we're already using, and (2) cit is correlated the LHS term
  > OVER AND ABOVE its correlation with the RHS terms. In a more intuitive phrasing,
  > omitted variable bias occurs because cit incorporate additional information about
  > both a RHS variable AND and LHS variable that is not already featured in the
  > regression.
  >
  > The simplest way to see that cit has the needed features is that (1) we can
  > look at the correlation table for cit, and (2) the coefficient on cit in the
  > hw regression is strongly significant. The bias is larger for the HE terms
  > because the correlation between cit and HE and HE^2 is larger than is the
  > correlation between cit and HA and HA^2.
  >
  > */
```

```
//////////////////////////////////////\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
> //\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\//////////////////////////////////////\\

/* Advanced Discussion:
>
> If may seem odd that the heSqr and haSqr beta estimates shifted in the opposite
> direction from the estimates on he and ha given that the correlations between
> the linear and quadratic terms is of similar magnitude and has the same sign
> (i.e. cit is positively correlated with he, ha, heSqr, and haSqr).
>
> In general, if we have a collection of RHS terms X, a LHS term y, and a single
> omitted variable z, then the bias for the beta for a single X variable, xk
> will be:
>
> BIAS(xk) = beta(y,z; [X z])*beta(z,xk; X)
>
> where beta(y,z;[X Z]) is the beta estimate on z when we run
>
> reg y x1 x2 x3... z
>
> and beta(z, xk; X) is the beta on xk when we run
>
> reg z x1 x2 x3...
>
> */

// bias = change in estimate between the two regressions
matrix bias_direct = beta_compare[1,1..5] - beta_compare[2,1..5]

// grab beta on cit (i.e. beta(y,z; [X z])
scalar beta_hw_cit = beta_compare[2,6]

// calculate the beta for cit, i.e. beta(z, xk; X) for all xk
reg cit he ha c.he#c.he c.ha#c.ha
```

Source	SS	df	MS
Model	12.490557	4	3.12263925
Residual	160.412497	748	.21445521
Total	172.903054	752	.229924275

Number of obs	=	753
F(4, 748)	=	14.56
Prob > F	=	0.0000
R-squared	=	0.0722
Adj R-squared	=	0.0673
Root MSE	=	.46309

cit	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
he	.094434	.0360703	2.62	0.009	.0236229	.1652451
ha	.0337982	.0235445	1.44	0.152	-.0124231	.0800194
c.he#c.he	-.0022194	.0014722	-1.51	0.132	-.0051096	.0006708
c.ha#c.ha	-.0002965	.0002613	-1.13	0.257	-.0008095	.0002165
_cons	-1.07252	.5636022	-1.90	0.057	-2.17895	.0339104

```
// Calculate the bias using beta(y,z; [X z])*beta(z,xk; X)
matrix bias_indirect = beta_hw_cit*e(b)
```

```
218 // Compare the bias generated by
219 matrix list bias_direct

  bias_direct[1,5]
                    he          ha       heSqr       haSqr        cnst
  with cit   .74960327   .26828477  -.01761716  -.00235331  -8.5135066

220 matrix list bias_indirect

  bias_indirect[1,5]
                                      c.he#       c.ha#
              he          ha        c.he        c.ha       _cons
  y1   .74960327   .26828477  -.01761716  -.00235331  -8.5135066

221
222 //Convert log file (smcl) to pdf
```