Extra Exercises: Logarithm and Interaction

Antonio Pacificoi
Federica Romei

May 1, 2012

1. Use the data in WAGE2.dta for this exercise.

```
 obs:           935
vars:            17                          14 Nov 2010 21:35
size:        67,320 (99.4% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
wage            float  %9.0g                  monthly earnings
hours           float  %9.0g                  average weekly hours
IQ              float  %9.0g                  IQ score
KWW             float  %9.0g                  knowledge of world work score
educ            float  %9.0g                  years of education
exper           float  %9.0g                  years of work experience
tenure          float  %9.0g                  years with current employer
age             float  %9.0g                  age in years
married         float  %9.0g                  =1 if married
black           float  %9.0g                  =1 if black
south           float  %9.0g                  =1 if live in south
urban           float  %9.0g                  =1 if live in SMSA
sibs            float  %9.0g                  number of siblings
brthord         float  %9.0g                  birth order
meduc           float  %9.0g                  mother's education
feduc           float  %9.0g                  father's education
lwage           float  %9.0g                  natural log of wage
-------------------------------------------------------------------------------
```

(a) Estimate the model

$$log(wage) = \beta_0 + \beta_1 * educ + \beta_2 * exper + \beta_3 * tenure + \beta_4 * married + \beta_5 * black + \beta_6 * south + \beta_7 * urban * + u$$

and report the results in the usual form. Holding other factors fixed, what is the approximate difference in monthly salary between blacks and nonblacks? Is this difference statistically significant?

Solution:

```
. reg lwage educ exper tenure married black south urban, r

Linear regression                                      Number of obs =     935
                                                       F(  7,   927) =   50.83
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2526
                                                       Root MSE      =  .36547


------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0654307   .0064093    10.21   0.000     .0528524    .0780091
       exper |    .014043   .0032386     4.34   0.000     .0076872    .0203988
      tenure |   .0117473   .0025387     4.63   0.000      .006765    .0167295
     married |   .1994171   .0396937     5.02   0.000     .1215171    .2773171
       black |  -.1883499   .0367035    -5.13   0.000    -.2603816   -.1163182
       south |  -.0909036    .027363    -3.32   0.001    -.1446043    -.037203
       urban |   .1839121   .0271125     6.78   0.000     .1307031     .237121
       _cons |   5.395497   .1131274    47.69   0.000     5.173481    5.617512
------------------------------------------------------------------------------
```

The coefficient on black implies that, at given levels of the other explanatory variables, black men earn about 18.8% less than nonblack men. The *t-statistic* is about -4.95, then β_{black} is significant at 5% level.

(b) Add the variables $exper^2$ and $tenure^2$ to the equation and test their jointly significance at 20%.

Solution:

```
. g exper2=exper*exper

. g tenure2=tenure*tenure

. reg lwage educ exper tenure married black south urban exper2 tenure2, r

Linear regression                                      Number of obs =     935
                                                       F(  9,   925) =   39.85
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2550
                                                       Root MSE      =  .36528

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0642761   .0065215     9.86   0.000     .0514773    .0770748
       exper |   .0172146   .0133512     1.29   0.198    -.0089875    .0434167
      tenure |   .0249291   .0077777     3.21   0.001     .0096651     .040193
     married |    .198547   .0395432     5.02   0.000     .1209422    .2761518
       black |  -.1906636    .036513    -5.22   0.000    -.2623216   -.1190057
       south |  -.0912153   .0273367    -3.34   0.001    -.1448645   -.0375661
       urban |   .1854241    .027081     6.85   0.000     .1322768    .2385713
      exper2 |  -.0001138   .0005721    -0.20   0.842    -.0012365    .0010089
     tenure2 |  -.0007964   .0004134    -1.93   0.054    -.0016077    .0000148
       _cons |   5.358676   .1245028    43.04   0.000     5.114335    5.603016
------------------------------------------------------------------------------
```

```
. test (exper2=0) (tenure2=0)

 ( 1)  exper2 = 0
 ( 2)  tenure2 = 0

       F(  2,   925) =    1.90
            Prob > F =    0.1501
```

The $p-value$ is less than 0.2, hence you will reject the Null Hypothesis that they are jointly equal to zero.

(c) Extend the original model to allow the return to education to depend on race and test whether the return to education does depend on race.

Solution:

```
. g blackeduc=black*educ

. reg lwage educ exper tenure married black south urban blackeduc, r

Linear regression                                      Number of obs =     935
                                                       F(  8,   926) =   44.61
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2536
                                                       Root MSE      =  .36542

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0671153   .0066867    10.04   0.000     .0539925    .0802382
       exper |   .0138259    .003242     4.26   0.000     .0074633    .0201885
      tenure |    .011787   .0025382     4.64   0.000     .0068057    .0167684
     married |   .1989077   .0396634     5.01   0.000     .1210672    .2767483
       black |   .0948094   .2108264     0.45   0.653    -.3189435    .5085624
       south |  -.0894495   .0273466    -3.27   0.001     -.143118    -.035781
       urban |   .1838523   .0271105     6.78   0.000     .1306472    .2370574
   blackeduc |  -.0226237   .0164937    -1.37   0.171     -.054993    .0097457
       _cons |   5.374817   .1155165    46.53   0.000     5.148112    5.601521
------------------------------------------------------------------------------
```

We add the interaction $black * educ$ to the equation in part a. The coefficient on the interaction is about -.0226 (SE .0164). Therefore, the point estimate is that the return to another year of education is about 2.3 percentage points lower for black men than nonblack men. (The estimated return for nonblack men is about 6.7%.)
$t - statistic < 1.96$, then we cannot reject the Null Hypothesis that $\beta_{blackeduc} = 0$ at 5% significant level.

(d) Again, start with the original model, but now allow wages to differ across four groups of people: married and black, married and nonblack, single and black, and single and nonblack. What is the estimated wage differential between married blacks and married nonblacks?

Solution:

We choose the base group to be single, nonblack. Then we add dummy variables marrnonblck, singblck, and marrblck for the other three groups. The result is

```
 g marrblck= 0

g singblck = 0

g singlnblck=0

g marrnblck =0

replace marrblck=1 if black==1 & married==1

replace singlnblck=1 if married==0 & black==0

 replace singlblck =1 if married==0 & black==1

 replace marrnblck=1 if married==1 & black==0


.  reg lwage educ exper tenure south urban marrblck singlblck  singlnblck marrnblck, r

Linear regression                                      Number of obs =     935
                                                       F(  8,   926) =   44.66
                                                       Prob > F      =  0.0000
```

```
                                                       R-squared     =  0.2528
                                                       Root MSE      =   .3656

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0654751   .0064153    10.21   0.000     .0528848    .0780654
       exper |   .0141462   .0032479     4.36   0.000     .0077721    .0205203
      tenure |   .0116628   .0025493     4.57   0.000     .0066597     .016666
       south |  -.0919894   .0274988    -3.35   0.001    -.1459566   -.0380222
       urban |   .1843501   .0271884     6.78   0.000      .130992    .2377081
    marrblck |   .2502685   .0803059     3.12   0.002     .0926659    .4078712
   singlblck |  (dropped)
  singlnblck |   .2408201   .0829401     2.90   0.004     .0780478    .4035924
    marrnblck |   .4297348   .0731179     5.88   0.000     .2862388    .5732309
       _cons |   5.162973    .126034    40.96   0.000     4.915628    5.410319
------------------------------------------------------------------------------
```

As you can see if we put all four dummy variables and the constant Stata will automatically drop one of them. To see what is the difference in wage between a married black and married non black we should take the difference between :

$$\beta_{marrblck} - \beta_{marrnblack} = 0.25 - 0.42 = -0.17$$

This means that a married black person will earn 17% less than a non black. We can also test if this difference is different to zero using :

```
test marrblck=marrnblck

 ( 1)  marrblck - marrnblck = 0

       F(  1,   926) =   19.81
            Prob > F =    0.0000
```

As you can see the p-value is less than 0.5 then we can reject H_0.

2. The dataset for this exercise is VOTE1.dta that contains the following data:

```
  obs:           173
 vars:            10                          14 Nov 2010 20:48
 size:         7,612 (99.9% of memory free)
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
state           float  %9.0g                  state postal code
district        float  %9.0g                  congressional district
democA          float  %9.0g                  =1 if A is democrat
voteA           float  %9.0g                  percent vote for A
expendA         float  %9.0g                  campaign expends. by A, $1000s
expendB         float  %9.0g                  campaign expends. by B, $1000s
prtystrA        float  %9.0g                  % vote for president
lexpendA        float  %9.0g                  log(expendA)
lexpendB        float  %9.0g                  log(expendB)
shareA          float  %9.0g                  100*(expendA/(expendA+expendB))
--------------------------------------------------------------------------------
```

Consider a model with an interaction between expenditures:

$$voteA = \beta_0 + \beta_1 * prtystrA + \beta_2 * expendA + \beta_3 * expendB + \beta_4 * expendA * expendB + u$$

(a) What is the partial effect of *expendB* on *voteA*, holding *prtystrA* and *expendA* fixed? What is the partial effect of *expendA* on *voteA*? Estimate the equation and report the results in the usual form. Is the interaction term statistically significant?

Solution:

For the model

$$voteA = \beta_0 + \beta_1 * prtystrA + \beta_2 * expendA + \beta_3 * expendB + \beta_4 * expendA * expendB + u$$

the ceteris paribus effect of *expendB* on *voteA* is obtained by taking changes and holding *prtystrA*,*expendA*, and u fixed:

$$\Delta voteA = \beta_3 * \Delta expendB + \beta_4 * expendA * (\Delta expendB) = (\beta_3 + \beta_4 * expendA)\Delta expendB$$

or

$$\frac{\Delta voteA}{\Delta expendB} = (\beta_3 + \beta_4 * expendA)$$

To estimate the model type the following command on STATA:

```
g expendAB=expendA*expendB
```

```
. reg voteA prtystrA expendA expendB expendAB, r

Linear regression                                      Number of obs =     173
                                                       F(  4,   168) =   31.86
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5708
                                                       Root MSE      =  11.126

------------------------------------------------------------------------------
             |               Robust
       voteA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    prtystrA |   .3419408   .0890764     3.84   0.000     .1660875    .5177941
     expendA |    .038281   .0064561     5.93   0.000     .0255355    .0510265
     expendB |  -.0317238    .005876    -5.40   0.000    -.0433241   -.0201235
    expendAB |  -6.63e-06   8.19e-06    -0.81   0.419    -.0000228    9.54e-06
       _cons |   32.11731    4.97646     6.45   0.000     22.29285    41.94176
------------------------------------------------------------------------------
```

The interaction term is not statistically significant, as its *t-statistic* is less than one in absolute value.

3. The dataset for this exercise is GPA2.RAW that contains the following data:

```
  obs:         4,137
 vars:            12                          14 Nov 2010 01:15
 size:       215,124 (97.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
sat             float  %9.0g                  combined SAT score
tothrs          float  %9.0g                  total hours through fall semest
colgpa          float  %9.0g                  GPA after fall semester
athlete         float  %9.0g                  =1 if athlete
verbmath        float  %9.0g                  verbal/math SAT score
hsize           float  %9.0g                  size graduating class, 100s
hsrank          float  %9.0g                  rank in graduating class
hsperc          float  %9.0g                  100*(hsrank/hssize)
female          float  %9.0g                  =1 if female
white           float  %9.0g                  =1 if white
black           float  %9.0g                  =1 if black
hsizesq         float  %9.0g                  hsize^2
-------------------------------------------------------------------------------
```

Consider the following model

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2$$

where SAT is student's standardized test score: students take the SAT to get into college. HSIZE is size of graduating class (in hundreds).

(a) Write the results in the usual form. Is the quadratic term statistically significant?

Solution:

To estimate the model type the following command on STATA:

```
reg sat hsize hsizesq
```

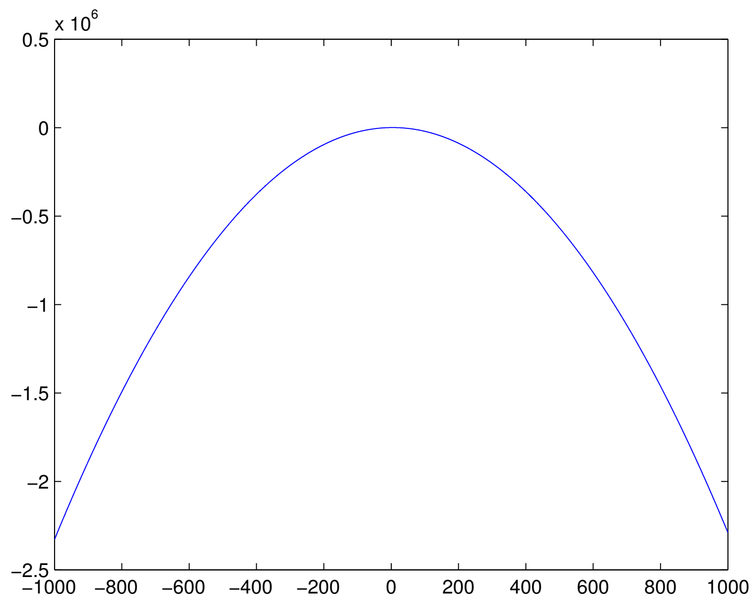
Estimating this model with the given data, we have:

$$sat = 997.9805 + 19.81446hsize - 2.130606hsize^2$$

The t-statistic for the hypothesis test that is -3.91, so the quadratic term is not statistically significant at the standard significance levels (1%, 5%, or 10%).

(b) Using the estimated equation from first part, what is the optimal high school size? Justify your answer. *(Hint: The geometric interpretation could be helpful)*

Solution: Looking at the estimated coefficients, we see that the model predicts a relationship between graduating class size and SAT score that looks like this parabola:



We know that for any quadratic equation,$y = ax + bx^2 + c$, the maximum or minimum of y is at $x = -\frac{a}{2b}$.

Applying this to the regression results, SAT score peaks at $-\frac{\hat{\beta}_1}{2\hat{\beta}_2} = -\frac{19.81446}{2(-2.130606)} =$ $4,65$, which is measured in hundreds of students, so 465 students is the optimal high school graduating class size. We can see that the second order derivatives is negative, hence the point found is a maximum.

(c) Is this analysis representative of the academic performance of all high school seniors? Explain.

Solution: No, because only students who have some intention or interest in attending college take the SAT. So this regression only gives us information about the performance of these potentially college-bound students.

(d) Find the estimated optimal high school size, using $log(sat)$ as the dependent variable. Is it much different from what you obtained in part b?

Solution:

We estimate the same equation but with as the dependent variable:

$$log(sat) = \beta_0 + \beta_1 hsize + \beta_2 hsize^2$$

In STATA you have to generate a new variable

```
g lsat=log(sat)
```

Here we are still trying to maximize the value of the dependent variable, so we can use the same formula to find the optimal class size: $-\frac{\hat{\beta}_1}{2\hat{\beta}_2} = -\frac{.0196029}{2(-.0020872)}$, so 475 students, pretty close to the previous model's result.