PUBLIC POLICY 639: QUIZ 4 SOLUTIONS
Winter 2011

On pages 3 and 4 you will see variable definitions and regression output from regressions of monthly wages on education and several other controls.

In Model 1, a researcher is evaluating the returns to education. Using the information from Model 1, answer the questions below.

1. (1 point) Interpret the coefficient on the variable male. Be specific and precise.

On average, males earn $768.31 more per month than females, holding education, age, and residential location constant.

2. (3 points) Multicollinearity can be a potential issue when using multiple independent variables. Describe the difference between perfect and imperfect multicollinearity and the impact of each.

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors. The result of this is the exclusion of one of regressors (dropped from estimation procedure). Imperfect multicollinearity occurs when two or more regressors are very highly correlated. The result of this leads to one or more of the regression coefficients being imprecisely estimated.

Unsure how to measure education, the researcher uses a categorical form of education in the second model. Using the information from Model 2, answer the questions below.

3. (1 point). What is the value of $\hat{\beta}_0$ in the regression we have run? Interpret $\hat{\beta}_0$ in words.

The constant term in model 2 is equal to $1,027.408. This value is the average monthly wage for urban females with less than a high school education with an age of zero.

4. (1 point) From the results the researcher is unsure which measurement of education is superior. What information or technique(s) should be used to make this decision.

Economic theory should be the primary influence in determining which specification to utilize. Additionally, the model fit statistics (R^2 ; adjusted R^2) and joint hypotheses tests can be used to evaluate alternative specifications of distributions.

It has been suggested to the researcher that a squared age term (quadratic term) might be useful, and so it is included in Model 3. Answer the questions below using the Model 3 output.

5. (1 point) Do you agree that a squared age term is called for given the model? In your discussion, include why a researcher would include a squared term.

A squared term is included to allow for non-linear association between x and y. In this case, theory suggests that the relationship between age and wages is non-linear, and so adding a squared term makes good sense theoretically. Looking at the regression output, the R^2 value does increase with the inclusion of the squared term (although evaluating the adjusted R^2 would be preferred) and more importantly the squared term coefficient is statistically significant at .01 alpha level, which also suggests the term is necessary.

6. (2 points) Interpret the coefficient on the age squared variable. How, if at all, does this change the interpretation of the age variable?

Given a positive age coefficient and a negative age squared coefficient it is clear that the change in monthly wages associated with a one year increase in age will be positive and increasingly larger as respondents get older and older until at some point in the age scale (near middle of age range) will have negative association that will increasingly get larger.

It is difficult to directly interpret age squared coefficient in terms of association with wages. This is because the association is now dependent on the level of age (as now we are specifying a non-linear relationship). So to interpret the association of age on wages we need to consider both the age and age squared coefficient.

Can directly calculate the predicted change in monthly wages holding all else constant. For example:

*1-year increase from 25 to 26 => ((105.673*26) + (-1.205*26*26)) - ((105.673*25) + (-1.205*25*25)) = 44.218*
*1-year increase from 45 to 46 => ((105.673*46) + (-1.205*46*46)) - ((105.673*45) + (-1.205*45*45)) = -3.98*
*1-year increase from 65 to 66 => ((105.673*66) + (-1.205*66*66)) - ((105.673*65) + (-1.205*65*65)) = -52.18*

7. (1 point) In evaluating Model 3, it is suggested that the researcher run the fully saturated model. What are the possible benefits of this? Note: There is no need to write the formula for the saturated model.

The most flexible way of allowing a non-linear relationship is to utilize the saturated model. So given we have suggested that a non-linear relationship may exist, a saturated model would be a superior method of capturing this relationship. More specifically, a saturated model allows for the prediction of the outcome for every possible respondent specification in the sample of interest. Although practically, interpretation of all the output can be tedious and difficult, further point estimates are typically very imprecise.

STATA OUTPUT

Variable Description

variable name	storage type	display format	value label	variable label
wage	float	%9.0g		monthly wages (measured in dollars)
educ	byte	%9.0g		years of education completed
lths	float	%9.0g		less than high school indicator variable
hs	float	%9.0g		high school indicator variable
col	float	%9.0g		college indicator variable
male	byte	%9.0g		male indicator (1=male, 0=female)
age	byte	%9.0g		age of respondent
age2	float	%9.0g		age squared (age*age)
rural	byte	%9.0g		rural indicator (1=rural, 0=urban)

Regression Output

*** Model 1 ***

```
. reg wage educ male age rural, robust

Linear regression                               Number of obs =     509
                                                F(  4,   504) =   43.25
                                                Prob > F      =  0.0000
                                                R-squared     =  0.4007
                                                Root MSE      =  1537.3

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   241.8175   23.95298    10.10   0.000     194.7575    288.8775
        male |   768.3144   132.0505     5.82   0.000      508.877    1027.752
         age |   25.77636   5.857708     4.40   0.000     14.26782    37.28489
       rural |  -917.1919   126.4033    -7.26   0.000    -1165.534   -668.8495
       _cons |  -886.7967   325.6955    -2.72   0.007    -1526.685   -246.9087
------------------------------------------------------------------------------
```

*** Model 2 ***

```
. reg wage hs col male age rural, robust

Linear regression                               Number of obs =     509
                                                F(  5,   503) =   38.89
                                                Prob > F      =  0.0000
                                                R-squared     =  0.4388
                                                Root MSE      =  1489.2

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          hs |   2308.537   334.3528     6.90   0.000     1651.637    2965.437
         col |   4587.894   733.4853     6.25   0.000     3146.822    6028.966
        male |   726.0526   129.3898     5.61   0.000     471.8416    980.2635
         age |   12.24378   5.299811     2.31   0.021     1.831285    22.65627
       rural |  -1347.282    131.393   -10.25   0.000    -1605.428   -1089.135
       _cons |   1027.408   209.2272     4.91   0.000     616.3407    1438.475
------------------------------------------------------------------------------
```

*** Model 3 ***

```
. reg wage hs col male age age2 rural, robust

Linear regression                               Number of obs =     509
                                                F(  6,   502) =   33.80
                                                Prob > F      =  0.0000
                                                R-squared     =  0.4458
                                                Root MSE      =  1481.2

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          hs |   2291.634   331.9414     6.90   0.000     1639.468    2943.799
         col |   4516.932   724.4331     6.24   0.000     3093.637    5940.226
        male |   709.4947    127.789     5.55   0.000     458.4276    960.5619
         age |    105.673   29.56296     3.57   0.000     47.59059    163.7553
        age2 |  -1.205561   .3891506    -3.10   0.002    -1.970126   -.4409968
       rural |   -1340.86   130.7386   -10.26   0.000    -1597.722   -1083.998
       _cons |  -613.0723   527.4342    -1.16   0.246    -1649.323    423.1782
------------------------------------------------------------------------------
```