Econometrics

Summer term 2018, Berlin School of Economics and Law

Assignment 1

I (8 Points)

Use the HPRICE1-data (HPRICE1.DTA) in order to analyze house prices.

(a) What is the sample mean of *price*? Provide a table of summary statistics for *price*, *bdrms* and *lotsize*

(b) Consider the model

$$\log(price) = \beta_0 + \beta_1 bdrms + \beta_2 \log(lotsize) + v. \quad (1)$$

Estimate the model and provide the results in the usual form, including n and R^2. Interpret the coefficients, i.e. explain the meaning.

(c) Now consider the model

$$\log(price) = \beta_0 + \beta_1 bdrms + \beta_2 \log(lotsize) + \beta_3 \left[\log(lotsize)\right]^2 + u. \quad (2)$$

(i) Estimate the model and provide the results in the usual form and interpret the coefficients.

(ii) Compare point estimates and standard errors of the *log(lotsize)*- and *bdrms*-coefficients in the two models (equations (1) and (2)). Explain why this pattern occurs.

(d) A possible estimator for the variance of u in model (2) is to calculate the sample variance of the residuals ($SSR/(n-1)$). We know that this estimator is biased. Which (unbiased) estimator is usually implemented in regression packages (provide the formula)? Explain what unbiasedness means. Which assumptions are required for this property? Obtain the two suggested variance estimates for model (2). Are there large differences? Why (not)?

(e) Suppose that the regression output of model (2) looks like this:

```
. reg lprice bdrms llotsize llotsize_sq

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   18.81
       Model |  3.22135859     3  1.0737862            Prob > F      =  0.0000
    Residual |  4.79624493    84  .057098154           R-squared     =  0.4018
-------------+------------------------------           Adj R-squared =  0.3804
       Total |  8.01760352    87  .092156362           Root MSE      =  .23895

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       bdrms |   .1403251
    llotsize |   .0875829
 llotsize_sq |   .0085262
       _cons |   3.673906
------------------------------------------------------------------------------
```

Here is the output from a regression of log(*lotsize*) on $[\log(lotsize)]^2$ and *bdrms*:

```
. reg llotsize llotsize_sq bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) = 6960.06
       Model |  25.5958243     2  12.7979122           Prob > F      =  0.0000
    Residual |  .156294979    85  .001838764           R-squared     =  0.9939
-------------+------------------------------           Adj R-squared =  0.9938
       Total |  25.7521193    87  .296001371           Root MSE      =  .04288

------------------------------------------------------------------------------
    llotsize |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 llotsize_sq |   .0539998   .0004645   116.27   0.000     .0530764    .0549233
       bdrms |   .0000134   .0055446     0.00   0.998    -.0110108    .0110376
       _cons |   4.607022   .0391183   117.77   0.000     4.529244      4.6848
------------------------------------------------------------------------------
```

Calculate the standard error of $\hat{\beta}_2$ in model (2). Similarly, calculate the standard error of $\hat{\beta}_1$ using this output:

```
. reg bdrms llotsize_sq llotsize
```

```
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =    1.26
       Model |   1.7797034     2  .889851702           Prob > F      =  0.2876
    Residual |  59.8112057    85  .703661243           R-squared     =  0.0289
-------------+------------------------------           Adj R-squared =  0.0060
       Total |  61.5909091    87  .707941484           Root MSE      =  .83885

------------------------------------------------------------------------------
       bdrms |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 llotsize_sq |   .0139622   .1149278     0.12   0.904     -.214545    .2424694
    llotsize |   .0051277   2.121824     0.00   0.998    -4.213626    4.223882
       _cons |   2.411218   9.801707     0.25   0.806     -17.0772    21.89964
------------------------------------------------------------------------------
```

Relate these calculations to your answer of part (cii).

II (7 Points)

Use the data in WAGE2.RAW for this exercise.

(a) Estimate the model

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u \quad (3)$$

and report the results in the usual form. Holding other factors fixed, what is the approximate difference in monthly salary between blacks and nonblacks? What is the corresponding 90%- confidence interval for this salary difference?

(b) Adding the variables $exper^2$ and $tenure^2$ yields the following output:

```
. gen exper_sq = exper^2

. gen tenure_sq = tenure^2

.
. reg lwage educ exper tenure married black south urban exper_sq tenure_sq

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  9,   925) =   35.17
       Model |  42.2353257     9  4.69281397           Prob > F      =  0.0000
    Residual |  123.420958   925  .133428062           R-squared     =  0.2550
-------------+------------------------------           Adj R-squared =  0.2477
       Total |  165.656283   934  .177362188           Root MSE      =  .36528
```

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0642761	.0063115	10.18	0.000	.0518896	.0766625
exper	.0172146	.0126138	1.36	0.173	-.0075403	.0419695
tenure	.0249291	.0081297	3.07	0.002	.0089743	.0408838
married	.198547	.0391103	5.08	0.000	.1217917	.2753023
black	-.1906636	.0377011	-5.06	0.000	-.2646533	-.116674
south	-.0912153	.0262356	-3.48	0.001	-.1427035	-.0397271
urban	.1854241	.0269585	6.88	0.000	.1325171	.2383311
exper_sq	-.0001138	.0005319	-0.21	0.831	-.0011576	.00093
tenure_sq	-.0007964	.000471	-1.69	0.091	-.0017208	.0001279
_cons	5.358676	.1259143	42.56	0.000	5.111565	5.605787

Use an F-Test for choosing between this model and the more parsimonious model of part (a), i.e. test whether the additional variables provide 'enough' additional explanatory power. Provide the test statistics, the rejection rule ($\alpha = 0.05$) and the test decision.

(c) Extend the original model to allow the return to education to depend on race. Provide the model equation and then run the regression and provide the results. Test whether the return to education does depend on race ($\alpha = 0.1$).

   (i) Interpret the *black*-coefficient of this model and compare it to part (a).

   (ii) Modify your model by replacing the interaction term with $black \cdot (educ - c)$ where c is an appropriate value. Which value of c do you suggest? Run the regression and interpret the *black*-coefficient of this regression.

(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?

III (8 Points)

There has been much interest in the question whether the presence of 401(k) pension plans, available to many U.S. workers, increases net savings. The data set 401KSUBS.RAW contains information on net financial assets (*nettfa*), family income (*inc*), a binary variable for eligibility in a 401(k) plan (*e401k*), and several other variables. In the following, you are asked to run a regression that predicts eligibility.

(i) How many families are eligible and how many are not eligible for participation in a 401(k) plan? Present the absolute numbers and the respective fractions.

(ii) Estimate a linear probability model explaining 401(k) eligibility in terms of income, age, and gender. Include income and age in quadratic form, and report the results in the usual form.

(iii) Interpret the coefficients.

(iv) Would you say that 401(k) eligibility is independent of income? What about age? What about gender? Explain.

(v) Obtain the fitted values from the linear probability model estimated in part (ii). Are any fitted values negative or greater than one?

(vi) Using the fitted values $\widehat{e401k_i}$ from part (iv), define $\widetilde{e401k_i} = 1$ if $\widehat{e401k_i} \geq 0.5$ and $\widetilde{e401k_i} = 0$ if $\widehat{e401k_i} < 0.5$. Out of 9,275 families, how many are predicted to be eligible for a 401(k) plan?

(vii) Use the variable $\widetilde{e401k_i}$ to compute the overall percent of correctly predicted/classified observations (families).

(viii) Now compute the percent correctly predicted/classified for both eligible and non-eligible families. What does this suggest regarding your previously computed (part vii) measure of model fit?

This assignment is due Wednesday, May 23, 12AM

Save all your commands in a do-file. Submit

1. a well-formatted document containing your answers and your results. Please submit a single pdf-file.
2. the corresponding do-file (or Rmd/R-file) that generates your results and the log-file.

Upload the files to moodle.