Department of Economics **UN3412**
Columbia University **Fall 2021**

SOLUTIONS TO
Problem Set 4
Introduction to Econometrics
Seyhan Erden and Tamrat Gashaw

Q#1: [30 points] Use the data in **hprice1.dta**. to estimate the following model (description of the variables in the data set is listed below in Table 1:

$$price = \beta 0 + \beta 1 sqrft + \beta 2 bdrms + u$$

where price = the (selling) price of the house (in 1000 dollars), sqrft = size of house (square feet) and bdrms = number of bedrooms in the house.

- **(a) [3 point]** Write out the estimation result in equation form.
- **(b) [3 point]** What is the estimated increase in price for a house with one more bedroom keeping square footage constant?
- **(c) [6 point]** What is the estimated increase in price for a house with an additional 1400-square-foot bedroom added? Compare this to your answer in (b).
- **(d) [6 point]** What percentage of the variation in price is explained by square footage and number of bedrooms? Compare your answer to the adjusted R^2. Explain the difference.
- **(e) [6 point]** Consider the first house in the sample. Report the square footage and number of bedrooms for this house. Find the predicted selling price for this house from the OLS regression line.
- **(f) [6 point]** What is the actual selling price of the first house in the sample? Find the residual of this house. Does it suggest that the buyer underpaid or overpaid for the house? Explain.

Table 1: DATA DESCRIPTION, FILE: hprice1.dta

Variable	Definition
price	House price, in $1000.
Assess	Assessed value in $1000.
bdrms	Average number bedrooms.
Lotsize	Size of lot in square feet.
Sqft	Size of house in square feet
colonial	= 1 if house is in Colonial style. = 0 otherwise.
Lprice	Log(price)
lassess	Log(assess)
llotsize	Log(lotsize)
lsqft	Log(sqft)

Solution:

(a) The estimated regression equation is:

$$\widehat{price} = -19.32 + 0.128\ sqrft + 15.20\ bdrms$$

(b) Holding square footage constant, $\Delta price$ = 15.20 $\Delta bdrms$; and so price increases by 15.20 for each additional bedroom. Since the unit of price is in thousands this means, $15,200.

(c) Now, $\Delta price$ = 0.128 $\Delta sqrft$ + 15.20 $\Delta bdrms$ = 194.4 Since unit of price is in thousands this means $194,400 Because the house's size is increasing as well, the total effect is much larger in (c). In part (b) the additional bedroom is obtained by converting existing rooms in the house so square footage remains unchanged. In (c), the added bedroom increases the square footage so the effect on price is much larger.

(d) R^2 = 0.6319 so about 63.19%. On the other hand, adjusted_R^2 = 0.6233 which is smaller. By construction, adjusted_R^2 is always smaller than R^2; this is due to the fact that it takes into account the presence of k = 2 regressors in the equation.

(e) We see that sqrft = 2,438 and bdrms = 4. The predicted price is:

$$\widehat{price} = -19.32 + 0.128\ x\ 2{,}438 + 15.20\ x\ 4 = 353.544.$$

The unit of price is in thousands, 353.544.× 1000 = 353,544, thus, we expect the house price to worth $353,544.

(f) The actual selling price was 300,000. Thus, the residual: $\hat{u}$ = 300,000 – 353,544 = −53,544.
This could suggest that the buyer underpaid by some margin. However, there are many other features of a house (some that we cannot even measure) that affect price, and we have not controlled for these. Thus, the negative residual could simply be a consequence of those other features made the house less attractive/valuable.

Q#2: [40 Points] Consider the following Population Linear Regression Function (PLRF):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i \qquad (1)$$

where, Y_i = average hourly earnings/wage in $, X_1= years of education, X_2 = years of potential experience, X_3 = years with current employer (tenure), X_4 = 1 if female, X_5 = 1 if nonwhite, and u_i= the usual error term of the model.

For this question, use the accompanying WAGE dataset with this problem set (i.e., a different dataset than you used in PS#2). Here is the description of the variables in the dataset for your consumption. We might be using this data set for the coming problem sets too.

```
Obs:   526

1. wage                     average hourly earnings
2. educ                     years of education
3. exper                    years potential experience
4. tenure                   years with current employer
5. nonwhite                 =1 if nonwhite
6. female                   =1 if female
```

7.	married	=1 if married
8.	numdep	number of dependents
9.	smsa	=1 if live in SMSA
10.	northcen	=1 if live in north central U.S
11.	south	=1 if live in southern region
12.	west	=1 if live in western region
13.	construc	=1 if work in construc. Indus.
14.	ndurman	=1 if in nondur. Manuf. Indus.
15.	trcommpu	=1 if in trans, commun, pub ut
16.	trade	=1 if in wholesale or retail
17.	services	=1 if in services indus.
18.	profserv	=1 if in prof. serv. Indus.
19.	profocc	=1 if in profess. Occupation
20.	clerocc	=1 if in clerical occupation
21.	servocc	=1 if in service occupation
22.	lwage	log(wage)
23.	expersq	exper^2
24.	tenursq	tenure^2

(a) **[7 Points]** Consider the following restricted version of equation (1) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Suppose that X_2 is omitted from the model by the researcher. For X_2 to cause omitted variable bias (OVB), what conditions should it satisfy? Show mathematically that the OLS estimator β_1 is biased if X_2 is omitted from the model.

(b) **[6 Points]** Run a regression of $Y_i = \beta_0 + \beta_4 X_4 + u_i$ and interpret the slope coefficient β_4. (Hint: X_4 is a binary explanatory variable.)

(c) **[7 Points]** First generate a dummy variable D_i such that $D_i = 1$ if male and $D_i = 0$ if female. Then run a regression of $Y_i = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_6 D_i + u_i$. What do you notice in the result? Explain why? Show mathematically that if X_4 and D_i are related, this result is inevitable.

(d) **[7 Point]** Run, first, a simple regression of $Y_i = \beta_0 + \beta_1 X_1 + u_i$ then $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$. Explain what happened to β_1(before and after) and why it happened.

(e) **[7 Points]** Now run the full model (1), using both homoscedastic-only and heteroskedasticity-robust standard errors, and interpret and compare the results of both regressions. Why do we care about heteroskedasticity problem that might exist in the data?

(f) **[6 Point]** Based on the regression result of the later (i.e., heteroskedasticity-robust standard errors), conduct the following hypothesis testing:

i. $H_0\colon\ \beta_i = 0\ \ vs\ H_1\colon \beta_i \neq 0$ where $i = 1, 2, \ldots, 5$

ii. $H_0\colon\ \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0\ \ vs\ H_1\colon At\ least\ one\ \beta_i \neq 0$

Solution

1 (a) For the omitted variable, X_2, to cause omitted variable bias (OVB), it should satisfy the following to conditions:

i) Years of potential experience (X_2) should be a determinant factor for average hourly earnings/wage (Y_i). That is, $Y_i = f(X_2)$ and hence X_2 is part of the error term u_i.

ii) Years of potential experience (X_2) should be correlated with years of education (X_1). Mathematically, it means that, $Corr(X_1, X_2) \neq 0$. Intuitively, this implies that more years of experience is correlated or sometimes affects years of education. The more you spent your years in acquiring work experience, the less time you are left with to spend in (formal) education or the number of years of education would increase as you might be taking few courses (i.e., you are part time student) as you are on the job.

Show mathematically that the OLS estimator $\boldsymbol{\beta_1}$ is biased if $\boldsymbol{X_2}$ is omitted from the model.

Start with the OLS equation for $\hat{\beta}_1$:

$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$ where small letter $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$.

Since $k_i = (X_i - \overline{X})/\sum_{i=1}^{n}(X_i - \bar{X})^2 = x_i/\sum_{i=1}^{n} x_i{}^2$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \sum k_i y_i = \sum k_i(Y_i - \bar{Y}) = \sum k_i Y_i - \sum k_i \bar{Y}$$

$\hat{\beta}_1 = \sum k_i Y_i \quad (do\ you\ wee\ why?)$ (2)

Now replace Y_i by the linear model given in (1) and get:

$$\hat{\beta}_1 = \sum k_i(\beta_0 + \beta_1 X_i + u_i) = \sum k_i \beta_0 + \sum k_i \beta_1 X_i + \sum k_i u_i$$

Using the properties of k_i, that are, $\sum_{i=1}^{n} k_i = 0$ and $\sum_{i=1}^{n} k_i X_i = 1$ in the above expression, we get:

$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} k_i u_i$ (3)

(i) If LSA#1 is true, then we can prove that $\hat{\beta}_1$ is unbiased estimator of β_1 by using (3) and take expectation of both sides.

$$E(\hat{\beta}_1) = E(\beta_1) + E(\sum_{i=1}^{n} k_i u_i)$$

$$E(\hat{\beta}_1) = \beta_1 \quad \text{(Unbiased)}$$

However, if LSA#1 is violated and the two conditions are in place, then we can show that:

$$E(\hat{\beta}_1) = E(\beta_1) + E(\sum_{i=1}^{n} k_i u_i)$$

Since $Corr(X_1, X_2) \neq 0$ (condition #2) and hence X_2 is part of the error term u_i, $E(\sum_{i=1}^{n} k_i u_i) \neq 0$.

$$E(\hat{\beta}_1) \neq \beta_1 \quad \text{(Biased)}$$

The direction of the bias depends on the sign of the correlation between $k_i = f(X_1)$ and u_i. That is if $\rho_{xu} > 0$, then $\hat{\beta}_1$overestimates β_1 and if $\rho_{xu} < 0$, then $\hat{\beta}_1$underestimates β_1. You can also show that $\hat{\beta}_1$is inconsistent estimator by using equation (3) and apply the large sample property test as the book shows you on page214 Appendix 6.1.

1(b)

```
. reg wage female, r

       Linear regression                                      Number of obs =     526
                                                              F(  1,    524) =   71.23
                                                              Prob > F      =  0.0000
                                                              R-squared     =  0.1157
                                                              Root MSE      =  3.4763

       ------------------------------------------------------------------------------
                    |               Robust
               wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
       -------------+----------------------------------------------------------------
             female |   -2.51183   .2976209    -8.44   0.000    -3.096507   -1.927154
              _cons |   7.099489   .2513859    28.24   0.000     6.605641    7.593337
       ------------------------------------------------------------------------------
```

Since X_2=gender dummy (binary) variable that takes the value of 1 if female and 0 otherwise, the slope coefficient is interpreted as the difference-in-group mean. That is, average hourly earnings declines by $2.51 if the individual is female. Mathematically,

$$\beta_4 = E(Y_i|X_4 = 1) - \ E(Y_i|X_4 = 0) = -2.512$$

1 (c)

gen D = 1 – female

Since female = 0 are male individuals, this generate command would give you D = 1 for male and D = 0 for female. In other words, D and female are dummy variables that takes opposite values. D = 1 is the same as female = 0.

```
. reg wage educ female D, r
note: D omitted because of collinearity

       Linear regression                                      Number of obs =     526
                                                              F(  2,    523) =   69.10
                                                              Prob > F      =  0.0000
                                                              R-squared     =  0.2588
                                                              Root MSE      =  3.1855

       ------------------------------------------------------------------------------
                    |               Robust
               wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
       -------------+----------------------------------------------------------------
               educ |   .5064521   .0598956     8.46   0.000     .3887867    .6241176
             female |  -2.273362   .2702033    -8.41   0.000    -2.804179   -1.742545
                  D |  (omitted)
              _cons |   .6228168   .7286843     0.85   0.393    -.8086909    2.054324
       ------------------------------------------------------------------------------
```

As you can see it in the above regression output, our generated variable D is dropped out of the model. This is because of perfect multicollinearity between D and female. (see the formula we use to generate D above.

To show this result mathematically, start with the equation we are estimating:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_6 D_i + u_i$$

Since $D_i = 1 - X_4$

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_6 (1 - X_4) + u_i$$

Rearranging it will give us:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_6(1) - \beta_6(X_4) + u_i$$

$$Y_i = (\beta_0 + \beta_6) + \beta_1 X_1 + (\beta_4 - \beta_6) X_4 + u_i$$

Letting $\alpha_0 = \beta_0 + \beta_6$ and $\alpha_1 = \beta_4 - \beta_6$, we get,

$$Y_i = (\alpha_0) + \beta_1 X_1 + (\alpha_1) X_4 + u_i$$

As you can see it from this last expression, the model doesn't contain D_i any more. Using the above regression output, our estimated model parameters are $\alpha_0 = 0.6228$, $\beta_1 = 0.5065$, and $\alpha_1 = -2.2734$.

1 (d)

```
. reg  wage educ, r

       Linear regression                                      Number of obs =     526
                                                              F(  1,    524) =   78.09
```

```
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1648
                                                      Root MSE      =  3.3784

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593   .0612596     8.84   0.000     .4210146    .6617039
       _cons |  -.9048516   .7254795    -1.25   0.213    -2.330057    .5203539
------------------------------------------------------------------------------

. reg  wage educ  exper, r

      Linear regression                               Number of obs =     526
                                                      F(  2,   523) =   50.32
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.2252
                                                      Root MSE      =   3.257

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6442721   .0651869     9.88   0.000     .5162117    .7723324
       exper |   .0700954   .0109943     6.38   0.000      .048497    .0916938
       _cons |  -3.390539   .8648747    -3.92   0.000    -5.089595   -1.691484
------------------------------------------------------------------------------
```

As can be seen from the above two tables, the coefficient on education has increased from 0.54 to 0.64. The reason for this increment is the addition of one of the omitted variable, namely, experience. The fact that it is also statistically significant suggests that it is one of the determinant variable for our dependent variable (condition #1). This result is similar to the test score example that we are using in the text that when we add percentage of English language learner in the model, the coefficient on class size has changed.

1 (e)

```
. reg  wage educ exper tenure female nonwhite

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  5,   520) =   59.43
       Model |  2603.75212     5  520.750424           Prob > F      =  0.0000
    Residual |  4556.66217   520  8.76281186           R-squared     =  0.3636
-------------+------------------------------           Adj R-squared =  0.3575
       Total |  7160.41429   525  13.6388844           Root MSE      =  2.9602

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5703422   .0495667    11.51   0.000     .4729667    .6677177
       exper |    .025343   .0115814     2.19   0.029      .002591    .0480951
      tenure |   .1410697   .0211819     6.66   0.000     .0994572    .1826823
      female |  -1.812043   .2650973    -6.84   0.000    -2.332836    -1.29125
    nonwhite |   -.115874   .4269179    -0.27   0.786    -.9545699    .7228218
       _cons |  -1.540298   .7323115    -2.10   0.036    -2.978951   -.1016454
------------------------------------------------------------------------------

. reg  wage educ exper tenure female nonwhite, r

      Linear regression                               Number of obs =     526
                                                      F(  5,   520) =   35.62
```

```
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3636
                                                       Root MSE      =  2.9602

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5703422   .0613282     9.30   0.000     .4498608    .6908236
       exper |    .025343   .0098146     2.58   0.010     .0060619    .0446242
      tenure |   .1410697   .0279893     5.04   0.000     .0860837    .1960558
      female |  -1.812043   .2545377    -7.12   0.000    -2.312092   -1.311994
    nonwhite |   -.115874    .392651    -0.30   0.768    -.8872512    .6555031
       _cons |  -1.540298   .8306617    -1.85   0.064    -3.172163     .091567
------------------------------------------------------------------------------
```

Here the first table provides a regression result based on homoscedasticity-only standard error and the second one is based on heteroscedasticity-robust standard errors. As it can be seen from these two tables, the coefficients are the same in both cases but the corresponding standard errors are different for each coefficient. Since, the remaining t-statistics, p-values, and the resulting confidence intervals in the two tables are different as all of them are dependent of the standard errors. The interpretation will proceed as usual.

We care about the presence of heteroscedasticity in the data because, if indeed there is the problem of heteroscedasticity, the homoscedasticity-only standard errors will be wrong. As mentioned above, if the standard errors are wrong, then everything else that depends on these wrong standard errors will result in misleading and incorrect statistical inference. It is advisable to use heteroscedasticity-robust standard errors whenever possible even if there is no heteroscedasticity. This is because, if there is no heteroscedasticity in the data, both will give us the correct standard errors. (see page 163 of the text on this issue.)

1 (f)

i) For individual null hypothesis of these coefficients, you can directly use the reported t-statistics and the corresponding p-values and confidence intervals.

ii) For the joint-hypothesis testing, we use the following command and interpret the resulting F-statistic.

```
. test  educ exper tenure female nonwhite

        ( 1)  educ = 0
        ( 2)  exper = 0
        ( 3)  tenure = 0
        ( 4)  female = 0
        ( 5)  nonwhite = 0

              F(  5,   520) =   35.62
                   Prob > F =    0.0000
```

As you can see, the computed F-stat is 35.62 and from the F-distribution table we know that the 1%, 5%, and 10% critical values for q =5 are 3.02, 2.21 and 1.85, respectively. This implies that we can reject the null hypothesis of all slope coefficients are zero. In fact, STATA has already computed the p-value for you and it is **Prob > F =   0.0000.** This implies that we can reject the null at 1% significance level.

Q#3: [10points] Suppose that you have been asked to assess a research paper that is based on your regression results in 2(e). Evaluate the external and internal validity issues of your regression in 2 (e) above.

Solution:

- A useful framework to assess statistical studies is to examine the internal and external validity issues with the paper.

Internal validity: the statistical inferences about causal effects are valid for the population being studied. There are five treats of internal validities. These are:

- **Omitted variable bias**
  - The students can list many omitted variables and explain how that omitted variable can cause a bias.
- **Wrong functional form**
  - The estimated model in 2 (e) is a linear model using an OLS estimation method and it may have wrong functional form as it is linear.
- **Errors-in-variables bias**
  - There is a possibility for measurement error both in the Y-variable (i.e., wage) if the respondent is responding a wrong information due to a number of reasons; as well as in one of the regressors. Depending on the type of measurement error the estimated slope coefficient may be biased and inconsistent.
- **Sample selection bias**
  - In our sample for Q#2, we may have selected the observations in part based on the Y-variable itself or based on the error term as in the basketball example given in class.
- **Simultaneous causality bias**
  - This can happen in this dataset's context if the Y-variable, the average hourly earnings, is causal factor to one or more of the X-variables, say years of education or experience. That is, large wage (i.e., due to factors in the error term) leads to large values on one of the X-variables, then there can be simultaneous causality due to the violation of LSA#1.

External validity: the statistical inferences can be generalized from the population and setting studied to other populations and settings, where the "setting" refers to the legal, policy, and physical environment and related salient features.

- Specifically, the external validity concerns for paper based on this dataset are related to these types of questions. Can we use the findings of this paper to apply it to all states in the US? Can we extrapolate the results to use it to countries in Europe, Asia, Africa, or Canada? Can one extend the results for out of sample periods, say now in 2021?

Q#4. [20 Points] Consider the following model, known as the exponential regression model:

$$Y_i = \beta_0 e^{\beta_1 X_i} + u_i$$

(a) [**6 points**] Do you think that you can estimate the model parameters using OLS? Explain.
(b) [**7 points**] What do you suggest how to estimate the model parameters?
(c) [**7 points**] How do you think one can proceed estimating these by trial-and-error, or iterative, process?

SOLUTIONS

(a) No. Because the OLS technique that tries to minimize the error sum of squares, would result in nonlinear and non-solvable normal equations. That is:

$$u_i = Y_i - \beta_0 e^{\beta_1 X_i}$$
$$\sum u_i^2 = \sum (Y_i - \beta_0 e^{\beta_1 X_i})^2$$

Therefore, to minimize the error sum of squares, we have to differentiate it with respect to the two unknowns, which gives us:

$$\frac{\partial \sum u_i^2}{\partial \beta_0} = 2\sum (Y_i - \beta_0 e^{\beta_1 X_i})\left(-1e^{\beta_1 X_i}\right) = 0$$

$$\frac{\partial \sum u_i^2}{\partial \beta_1} = 2\sum (Y_i - \beta_0 e^{\beta_1 X_i})\left(-\beta_0 e^{\beta_1 X_i} X_i\right) = 0$$

Unlike the normal equations in the case of the linear regression model, the normal equations for nonlinear regression have the unknowns (the $\hat{\beta}_i$'s) both on the left- and right-hand sides of the equations. As a consequence, we cannot obtain explicit solutions of the unknowns in terms of the known quantities. To put it differently, the unknowns are expressed in terms of themselves and the data.

(b) To estimate these parameters, one can try two approaches. These are:

i. Transform the model into natural logarithmic linear format and apply OLS. That is, taking natural log on the model will give us:

$$\ln(Y_i) = \ln(\beta_0) + \beta_1 X_i + u_i$$

This is a log-linear model that can be estimated using OLS (i.e., after transforming the Y variable into natural log and the estimated intercept term is $\ln(\beta_0)$.

ii. The second approach is to use trial-and-error, or iterative, process as shown below (c).

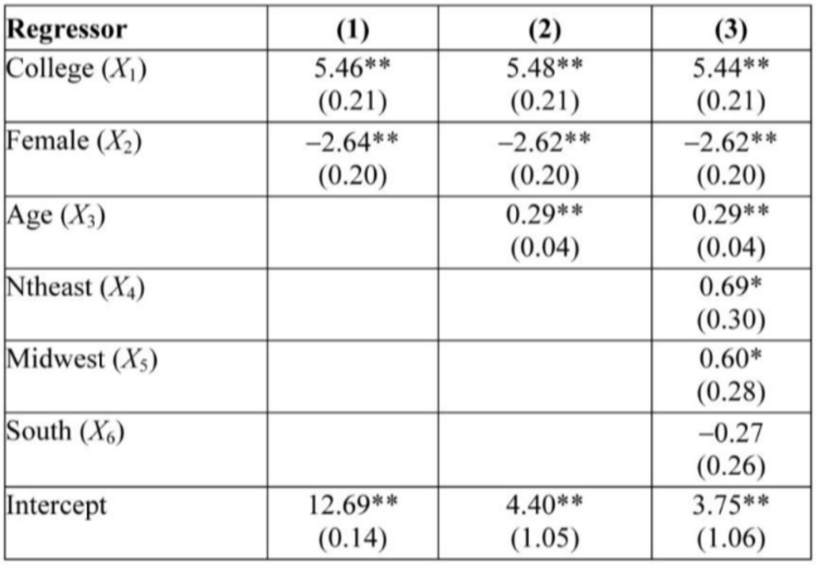
(c) The use trial-and-error, or iterative, process is implemented as follows. Suppose we assume that initially β_0 = 0.45 (or a) and β_1 = 0.01 (or b). These are pure guesses, sometimes based on prior experience or prior empirical work or obtained by just fitting a linear regression model even though it may not be appropriate. At this stage do not worry about how these values are obtained. Then, use these values in

$$\sum u_i^2 = \sum (Y_i - \beta_0 e^{\beta_1 X_i})^2$$

And see if the error sum of squares is the smallest possible values. If not keep on trying other values of the parameters until you get the minimum value for the error sum of squares. This is what STATA doing when it was estimating parameters using Nonlinear Least Square technique.

Following questions will not be graded, they are for you to practice and will be discussed at the recitation:

1. *SW Exercise 7.1*



Regressor	**(1)**	**(2)**	**(3)**
College (X_1)	5.46** (0.21)	5.48** (0.21)	5.44** (0.21)
Female (X_2)	–2.64** (0.20)	–2.62** (0.20)	–2.62** (0.20)
Age (X_3)		0.29** (0.04)	0.29** (0.04)
Ntheast (X_4)			0.69* (0.30)
Midwest (X_5)			0.60* (0.28)
South (X_6)			–0.27 (0.26)
Intercept	12.69** (0.14)	4.40** (1.05)	3.75** (1.06)

2. *SW Exercise 7.4*

(a) The F-statistic testing the coefficients on the regional regressors are zero is 6.10. The 1% critical value (from the $F_{3,\circ}$ distribution) is 3.78. Because 6.10 > 3.78, the regional effects are significant at the 1% level.

(bi) The expected difference between Juanita and Molly is $(X_{6,\text{Juanita}}\quad X_{6,\text{Molly}}) \cdot ®_6 = ®_6$. Thus a 95% confidence interval is 0.27 ± 1.96 · 0.26.

(b ii) The expected difference between Juanita and Jennifer is $(X_{5,\text{Juanita}}\quad X_{5,\text{Jennifer}}) \cdot ®_5 + (X_{6,\text{Juanita}}\quad X_{6,\text{Jennifer}}) \cdot ®_6 = \quad ®_5 + ®_6$. A 95% confidence interval could be constructed using the general methods discussed in Section 7.3. In this case, an easy way to do this is to omit *Midwest* from the regression and replace it with $X_5 =$ *West*. In this new regression the coefficient on *South* measures the difference in wages between the *South* and the *Midwest*, and a 95% confidence interval can be computed directly.

3. SW Empirical Exercises 7.1

	Model	
Regressor	**a**	**b**
Age	0.60 (0.04)	0.59 (0.04)
Female		–3.66 (0.21)
Bachelor		8.08 (0.21)
Intercept	1.08 (1.17)	–0.63 (1.08)
SER	9.99	9.07
R^2	0.029	0.200
$\bar{R}^2$	0.029	0.199

(a) The estimated slope is 0.60. The estimated intercept is 1.08.

(b) The estimated marginal effect of *Age* on *AHE* is 0.59 dollars per year. The 95% confidence interval is 0.59 ± 1.96 × 0.04 or 0.51 to 0.66.

(c) The results are quite similar. Evidently the regression in (a) does not suffer from important omitted variable bias.

(d) Bob's predicted average hourly earnings = $(0.59 \times 26) + (-3.66 \times 0) + (8.08 \times 0) - 0.63$ = \$14.17. Alexis's predicted average hourly earnings = $(0.59 \times 30) + (-3.66 \times 1) + (8.08 \times 1) - 0.63$ = \$21.49.

(e) The regression in (b) fits the data much better. Gender and education are important predictors of earnings. The R^2 and $\bar{R}^2$ are similar because the sample size is large (n = 7711).

(f) Gender and education are important. The *F*-statistic is 781, which is (much) larger than the 1% critical value of 4.61.

(g) The omitted variables must have non-zero coefficients and must correlated with the included regressor. From (f) *Female* and *Bachelor* have non-zero coefficients; yet there does not seem to be important omitted variable bias, suggesting that the correlation of *Age* and *Female* and *Age* and *Bachelor* is small. (The sample correlations are *Cor* (*Age, Female*) = –0.03 and *Cor* (*Age,Bachelor*) = 0.00).

4. SW exercise 8.2

a) According to the regression results in column (1), the house price is expected to increase by 21% (= 100% × 0.00042 × 500) with an additional 500 square feet and other factors held constant. The 95% confidence interval for the percentage change is 100% × 500 × (0.00042 ± 1.96 × 0.000038) = [17.276% to 24.724]

b) Because the regressions in columns (1) and (2) have the same dependent variable, $\bar{R}^2$ can be used to compare the fit of these two regressions. The log-log regression in column (2) has the higher $\bar{R}^2$, so it is better so use ln(*Size*) to explain house prices.

c) The house price is expected to increase by 7.1% (= 100% × 0.071 × 1). The 95% confidence interval for this effect is 100% × (0.071 ± 1.96 × 0.034) = [0.436% to 13.764%].

d) The house price is expected to increase by 0.36% (100% × 0.0036 × 1 = 0.36%) with an additional bedroom while other factors are held constant. The effect is not statistically significant at a 5% significance level: $|t| = \frac{0.0036}{0.037} = 0.09730 < 1.96.$ Note that this coefficient measures the effect of an additional bedroom holding the size of the house constant.

e) The quadratic term ln(*Size*)^2 is not important. The coefficient estimate is not statistically significant at a 5% significance level: $|t| = \frac{0.0078}{0.14} = 0.05571 < 1.96.$

f) The house price is expected to increase by 7.1% (= 100% × 0.071 × 1) when a swimming pool is added to a house without a view and other factors are held constant. The house price is expected to increase by 7.32% (= 100% × (0.071 × 1 + 0.0022 × 1)) when a swimming pool is added to a house with a view and other factors are held constant. The difference in the expected percentage change in price is 0.22%. The difference is not statistically significant at a 5% significance level: $|t| = \frac{0.0022}{0.10} = 0.022 < 1.96.$

5. SW Exercise 8.10

(a) $\Delta Y = f(X_1 + \Delta X_1, X_2) - f(X_1, X_2) = \beta_1 \Delta X_1 + \beta_3 \Delta X_1 \times X_2,$ so $\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2.$

(b) $\Delta Y = f(X_1, X_2 + \Delta X_2) - f(X_1, X_2) = \beta_2 \Delta X_2 + \beta_3 X_1 \times \Delta X_2,$ so $\frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1.$

(c)

$$\begin{aligned}
\Delta Y &= f(X_1 + \Delta X_1, X_2 + \Delta X_2) - f(X_1, X_2) \\
&= \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2(X_2 + \Delta X_2) + \beta_3(X_1 + \Delta X_1)(X_2 + \Delta X_2) \\
&\quad - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2) \\
&= (\beta_1 + \beta_3 X_2)\Delta X_1 + (\beta_2 + \beta_3 X_1)\Delta X_2 + \beta_3 \Delta X_1 \Delta X_2.
\end{aligned}$$

6. SW Exercise 9.6

(a) The parameter estimates do not change. Nor does the the R^2. The sum of squared residuals from the 100 observation regression is $SER_{200} = (100 - 2) \times 15.1^2 = 22,344.98,$ and the sum of squared residuals from the 200 observation regression is twice this value: $SSR_{200} = 2 \times 22,344.98.$ Thus, the *SER* from the 200 observation regression is $SER_{200} = \sqrt{\frac{1}{200-2} SSR_{200}} = 15.02.$ The standard errors for the regression coefficients are now computed using equation (5.4) where $\sum_{i=1}^{200}(X_i - \bar{X})^2 \hat{u}_i^2$ and $\sum_{i=1}^{200}(X_i - \bar{X})^2$ are twice their value from the 100 observation regression. Thus the standard errors for the 200 observation regression are the standard errors in the 100 observation regression multiplied by $\sqrt{\frac{100-2}{200-2}} = 0.704.$ In summary, the results for the 200 observation regression are

$$\underset{(10.63)}{\hat{Y} = 32.1} + \underset{(8.59)}{66.8X}, \; SER = 15.02, \; R^2 = 0.81$$

(b) The observations are not *i.i.d.*: half of the observations are identical to the other half, so that the observations are not *independent*.

7. SW Empirical Exercise 8.2

Using the data set **TeachingRatings** described in Empirical Exercise AEE4.2, carry out the following exercises.

a) Estimate a regression of Course_*Eval* on *Beauty, Intro, OneCredit, Female, Minority*, and *NNEnglish*.

b) Add *Age* and *Age*^{2} to the regression. Is there evidence that *Age* has a nonlinear effect on *Course_Eval*? Is there evidence that *Age* has any effect on *Course_Eval*?

c) Modify the regression in (a) so that the effect of *Beauty* on *Course_Eval* is different for men and women. Is the male–female difference in the effect of *Beauty* statistically significant?

d) Professor Smith is a man. He has cosmetic surgery that increases his beauty index from one standard deviation below the average to one standard deviation above the average.

What is his value of *Beauty* before the surgery? After the surgery? Using the regression in (c), construct a 95% confidence for the increase in his course evaluation.

e) Repeat (d) for Professor Jones, who is a woman.

Solution:

Dependent Variable = *Course_Eval*

Regressor	**(1)**	**(2)**	**(3)**	**(4)**
Beauty	0.166** (0.032)	0.160** (0.030)	0.231** (0.048)	0.090* (0.040)
Intro	0.011 (0.056)	0.002 (0.056)	–0.001 (0.056)	–0.001 (0.056)
OneCredit	0.635** (0.108)	0.620** (0.109)	0.657** (0.109)	0.657** (0.109)
Female	–0.173** (0.049)	–0.188** (0.052)	–0.173** (0.050)	–0.173** (0.050)
Minority	–0.167* (0.067)	–0.180** (0.069)	–0.135 (0.070)	–0.135 (0.070)
NNEnglish	–0.244** (0.094)	–0.243* (0.096)	–0.268** (0.093)	–0.268** (0.093)
Age		0.020 (0.023)		
Age^2		–0.0002 (0.0002)		
Female × *Beauty*			–0.141* (0.063)	
Male × *Beauty*				0.141 (0.063)
Intercept	4.068** (0.037)	3.677** (0.550)	4.075** (0.037)	4.075** (0.037)
***F*-statistic and *p*-values on joint hypotheses**				
Age and *Age*^{2}		0.63 (0.53)		
SER	0.514	0.514	0.511	0.511
$\bar{R}^2$	0.144	0.142	0.151	0.151

Significant at the *5% and **1% significance level.

(a) See Table

(b) The coefficient on Age^2 is not statistically significant, so there is no evidence of a nonlinear effect. The coefficient on *Age* is not statistically significant and the *F*-statistic testing whether the coefficients on *Age* and Age^2 are zero does not reject the null hypothesis that the coefficients are zero. Thus, *Age* does not seem to be an important determinant of course evaluations.

(c) See the regression (3) which adds the interaction term *Female* × *Beauty* to the base specification in (1). The coefficient on the interaction term is statistically significant at the 5% level. The magnitude of the coefficient in investigated in parts (d) and (e).

(d) Recall that the standard deviation of *Beauty* is 0.79. Thus Professor Smith's course rating is expected to increase by 0.231 × (2 × 0.79) = 0.37. The 95% confidence interval for the increase is (0.231 ± 1.96 × 0.048) × (2 × 0.79) or 0.22 to 0.51.

(e) Professor Smith's course rating is expected to increase by (0.231 – 0.173) × (2 × 0.79) = 0.09. To construct the 95% confidence interval, we need the standard error for the sum of coefficients $\beta_{\text{Beauty}} + \beta_{\text{Female}\times\text{Beauty}}$. How to get the standard error depends on the software that you are using. An easy way is re-specify the regression replacing *Female* × *Beauty* with *Male* × *Beauty*. The resulting regression is shown in (4) in the table. Now, the coefficient on *Beauty* is the effect of *Beauty* for females and the standard error is given in the table. The 95% confidence interval is (0.090 ± 1.96 × 0.040) × (2 × 0.79) or 0.02 to 0.27

Do file for the question above:

```
use "TeachingRatings.dta", clear
cap log close
log using "PS5_Q13.log", replace
reg course_eval beauty intro onecredit female minority nnenglish, robust
ereturn list r2_a
gen agesquared = age^2
reg course_eval beauty intro onecredit female minority nnenglish age agesquared,
robust
test age agesquared
ereturn list r2_a
gen fembeauty = female*beauty
reg course_eval beauty intro onecredit female minority nnenglish fembeauty, robust
ereturn list r2_a
gen male = 1 - female
gen malbeauty = male*beauty
reg course_eval beauty intro onecredit female minority nnenglish malbeauty, robust
ereturn list r2_a
log close
```

8. SW Empirical Exercise 9.2

A committee on improving undergraduate teaching at your college needs your help before reporting to the dean. The committee seeks your advice, as an econometric expert, about whether your college should take physical appearance into account when hiring teaching faculty. (This is legal as long as doing so is blind to race, religion, age, and gender.) You do not have time to collect your own data, so you must base your recommendations on the analysis of the data set **TeachingRatings** described in Empirical Exercise 4.2 that has served as the basis for several empirical exercises in Part II of the text. Based on your analysis of these data, what is your advice? Justify your advice based on a careful and complete assessment of the internal and external validity of the regressions that you carried out to answer the empirical exercises using these data in earlier chapters.

Solution:

We begin by discussing the internal and external validity of the results summarized in Empirical Exercise 8.2.

Internal Validity

1. ***Omitted variable bias.*** It is always possible to think of omitted variables, but the relevant question is whether they are likely to lead to substantial omitted variable bias. Standard examples like instructor diligence, are likely to be major sources of bias, although this is speculation and the next study on this topic should address these issues (both can be measured). One possible source of OV bias is the omission of the department. French instructors could well be more attractive than chemists, and if French is more fun (or better taught) than chemistry then the department would belong in the regression, and its omission could bias the coefficient on *Beauty*. It is difficult to say whether this is a major problem or not, one approach would be to put in a full set of binary indicators for the department and see if this changed the results. We suspect this is not an important effect, however this must be raised as a caveat.
2. ***Wrong functional form.*** Interactions with *Female* showed some evidence of nonlinearity. It would be useful to see if $Beauty^2$ enters the regression. (We have run the regression, and the *t*-statistic on $Beauty^2$ is –1.15.)
3. ***Measurement error in the regressors.*** The *Beauty* variable is subjectively measured so that it will have measurement error. This is plausibly a case in which the measurement error is more or less random, reflecting the tastes of the six panelists. If so, then the classical measurement error model, in which the measured variable is the true value plus random noise, would apply. But this model implies that the coefficient is biased *down*—so the actual effect of *Beauty* would be greater than is implied by the OLS coefficient. This suggests that the regressions *understate* the effect of *Beauty*.
4. ***Sample selection bias.*** The only information given in this exam about the sample selection method is that the instructors have their photos on their Web site. Suppose

instructors who get evaluations below 3.5 are so embarrassed that they don't put up their photos, and suppose there is a large effect of *Beauty*. Then, of the least attractive instructors, the only ones that will put up their photos are those with particular teaching talent and commitment, sufficient to overcome their physical appearance. Thus the effect of physical appearance will be attenuated because the error term will be correlated with *Beauty* (low values of *Beauty* means there must be a large value of *u*, else the photo wouldn't be posted.) This story, while logically possible, seems a bit far-fetched, and whether an instructor puts up his or her photo is more likely to be a matter of departmental policy, whether the department has a helpful webmaster and someone to take their photo, etc. So sample selection bias does not seem (in my judgment) to be a potentially major threat.

5. ***Simultaneous causality bias.*** There is an interesting possible channel of simultaneous causality, in which good course evaluations improve an instructor's self-image which in turn means they have a more resonant, open, and appealing appearance—and thus get a higher grade on *Beauty*. Against this, the panelists were looking at the Web photos, not their conduct in class, and were instructed to focus on physical features. So for the *Beauty* variable as measured, this effect is plausibly large.

External Validity

The question of external validity is whether the results for UT-Austin in 2000–2002 can be generalized to, say, Harvard or Cal-State University Northridge (CSUN) in 2005. The years are close, so the question must focus on differences between students and the instructional setting.

1. Are UT-Austin students like Harvard (or CSUN) students? Perhaps Beauty matters more or less to Harvard (CSUN) students?

2. Do the methods of instruction differ? For example, if beauty matters more in small classes (where you can see the instructor better) and if the distribution of class size at UT-Austin and Harvard (CSUN) were substantially different, then this would be a threat to external validity.

Policy Advice

As an econometric consultant, the question is whether this represents an internally and externally valid estimate of the causal effect of *Beauty*, or whether the threats to internal and/or external validity are sufficiently severe that the results should be dismissed as unreliable for the purposes of the Dean. A correct conclusion is one that follows logically from the systematic discussion of internal and external validity.

We would be surprised if the threats to internal and external validity above are sufficiently important, in a quantitative sense, to change the main finding from E8.2 that the effect of *Beauty* is positive and quantitatively large. So our advice, solely econometric consultants, would be that implementing a policy of affirmative action for attractive people (all else equal, higher the better-looking) would, in expectation, improve course evaluations.

This said, a good econometric policy advisor always has some advice about the next study. One thing that next study could do is focus on institutions like your's. (UT-Austin students and professors might be different from students at Harvard or CSUN), and collect data on some potential omitted variables (department offering the course, etc.).

A very different study would be to do a randomized controlled experiment that would get directly at the policy question. Some department heads would be instructed to assign their most attractive teachers to the largest introductory courses (treatment group), others would be instructed to maintain the status quo (control group). The study would assess whether there is an improvement in evaluation scores (weighted by class size) in the treatment group. A positive result would indicate that this treatment results in an increase in customer satisfaction.

Finally, some thoughts that were out of bounds for this question, but would be relevant and important to raise in the report of an econometric consultant to the Dean. First, the *Course Evaluation* score is just a student evaluation, not a measure of what students actually learned or how valuable the course was; perhaps an assessment of the value of the course, five years hence, would produce a very different effect of *Beauty*, and that is arguably a more important outcome than the end-of-semester evaluation (this could be thought of as a threat to external validity, depending on how one defines the Dean's goal). Second, academic output is not solely teaching, and there is no reason at all that the results here would carry over to an analysis of research output, or even graduate student advising and teaching (the data are only for undergrad courses); indeed, the sign might be the opposite for research. Third, the econometric consultant could raise the question of whether *Beauty* has the same *moral* status as gender or race, even if it does not have the same *legal* status as a legally protected class; answering this question is outside the econometric consultant's area of expertise, but it is a legitimate question to raise and to frame so that others can address it.