

EC220**Problem Sets 3 – 9,
Michaelmas Term 2011****2011/2012**

Note: Problem Sets 1 and 2 were distributed with the initial hand-out.

For the class in the week beginning October 31

PROBLEM SET 3

Note: For Items 1 and 2 you will use Stata and an *EAEF* data set derived from the US National Longitudinal Survey of Youth. For a description of the *EAEF* data sets, see Appendix B at the end of the text. You should download your data set from the website at

<http://econ.lse.ac.uk/courses/ec220/G/iedata/eecs/stata/>

To download, double click on the data set, choose ‘save this file to disk’, and then decide where you wish to save the file in your H:space. You will download the data set only once. Thereafter you will work with it from your H:space.

You may save Stata output by blocking it with the cursor and copying directly to a word-processing application such as Word. You should then change the typeface to Courier to ensure that the output lines up correctly. Alternatively, you may open a Stata output file. If you do this, you should choose ‘.log’ file, rather than the default ‘.smcl’ file, in the ‘save as file type’ window.

You will probably find it convenient to do Items 1 and 2 in the same Stata session. In principle, downloading the data set, launching Stata, running the regressions, and saving to a word processor should take no more than five minutes. Obviously, since this is your first time, it will take longer.

(1)* Exercise 1.6 in the text

Does educational attainment depend on intellectual ability? In the United States, as in most countries, there is a positive correlation between educational attainment and cognitive ability. *S* (highest grade completed by 2002) is the number of years of schooling of the respondent. *ASVABC* is a composite measure of numerical and verbal ability with mean 50 and standard deviation 10 (both approximately; for further details of the measure, see Appendix B). Perform a regression of *S* on *ASVABC* and interpret the regression results. Comment on the value of R^2 .

Stata command

```
reg S ASVABC
```

(2)* Exercise 1.7 in the text

Do earnings depend on education? Using your *EAEF* data set, fit an earnings function parallel to that in Table 1.2, regressing *EARNINGS* on *S*, and give an interpretation of the coefficients. Comment on the value of R^2 .

Stata command

```
reg EARNINGS S
```

(3) Exercise 1.16 in the text.

Come to class prepared to present an answer to this exercise. A written answer is not required.

Two individuals fit earnings functions relating *EARNINGS* to *S* using *EAEF* Data Set 21. The first individual does it correctly and obtains the result found in Section 1.6:

$$\widehat{EARNINGS} = -13.93 + 2.46 S$$

The second individual makes a mistake and regresses *S* on *EARNINGS*, obtaining the following result:

$$\hat{S} = 12.29 + 0.070EARNINGS$$

From this result the second individual derives

$$\widehat{EARNINGS} = -175.57 + 14.29S$$

Explain why this equation is different from that fitted by the first individual.

(4) Exercise 2.2 in the text.

For the model $Y_i = \beta_2 X_i + u_i$, the OLS estimator of β_2 is $b_2 = \sum_{i=1}^n X_i Y_i / \sum_{j=1}^n X_j^2$. Demonstrate that b_2 may be decomposed as

$$b_2 = \beta_2 + \sum_{i=1}^n d_i u_i$$

where $d_i = \frac{X_i}{\sum_{j=1}^n X_j^2}$ and hence demonstrate that it is an unbiased estimator of β_2 .

(5)* Exercise 2.4 in the text.

An investigator correctly believes that the relationship between two variables *X* and *Y* is given by

$$Y = \beta_1 + \beta_2 X + u$$

Given a sample of *n* observations, the investigator estimates β_2 by calculating it as the average value of *Y* divided by the average value of *X*. Discuss the properties of this estimator. What difference would it make if it could be assumed that $\beta_1 = 0$?

For the class in the week beginning November 7

PROBLEM SET 4

(1)* Exercise 2.12 in the text

Suppose that the true relationship between *Y* and *X* is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. In Section 1.4, it was shown that if $Y_i^* = \lambda_1 + \lambda_2 Y_i$, and Y^* is regressed on *X*, the slope coefficient $b_2^* = \lambda_2 b_2$. How will the standard error of b_2^* be related to the standard error of b_2 ?

(2) Exercises 2.17 and 2.22 in the text

Perform a *t* test on the slope coefficient and the intercept of the earnings function fitted using your *EAEF* data set in Exercise 1.7, and state your conclusions. Explain whether it would be possible to

perform one-sided tests instead of two-sided tests. If you think that one-sided tests are justified, perform them, giving your justification, and state whether the use of a one-sided test makes any difference.

(3) Exercise 2.27 in the text.

Calculate the 95% confidence interval for the slope coefficient of the earnings function fitted with your *EAEF* data set in Exercise 1.7. Check that it agrees with the confidence interval printed as part of the Stata output.

(4) Exercise 2.30 in the text.

Calculate the F statistic from ESS and RSS obtained in the earnings function fitted using your *EAEF* data set and check that it is equal to the value printed in the output. Check that the F statistic derived from R^2 is the same. Perform an appropriate F test.

(5) Exercise 2.31 in the text.

Verify that the F statistic in the earnings function regression run by you using your *EAEF* data set in Exercise 1.7 is equal to the square of the t statistic for the slope coefficient, and that the critical value of F at the 1 percent significance level is equal to the square of the critical value of t .

(6)* Exercise 2.32 in the text.

In Exercise 1.16 both researchers obtained a t statistic of 10.59 for the slope coefficient in their regressions. Was this a coincidence?

For the class in the week beginning November 14

PROBLEM SET 5

(1) Write down the name of your class teacher and his or her office hour (location, time).

(The purpose of this is to enable your class teacher to check that you know.)

(2) Exercise 3.2 in the text.

Does educational attainment depend on parents' education? Using your *EAEF* data set, first regress S on $ASVABC$ and SM and interpret the regression results. Repeat the regression using SF instead of SM , and then again including both SM and SF as regressors. There is a saying that if you educate a male, you educate an individual, while if you educate a female, you educate a nation. The premise is that the education of a future mother has a future beneficial knock-on effect on the educational attainment of her children. Do your regression results support this view?

Stata commands

```
reg S ASVABC SM
reg S ASVABC SF
reg S ASVABC SM SF
```

(3)* Exercise 3.7 in the text

Two researchers are investigating the effects of time spent studying on the examination marks earned by students on a certain course. For a sample of 100 students, they have the examination mark, M , total hours spent studying, H , hours on primary study, P , and hours spent on revision, R

By definition, $H = P + R$. Researcher A decides to regress M on P and R and fits the following regression:

$$\hat{M} = 45.6 + 0.15 P + 0.21 R$$

Researcher B decides to regress M on H and P , with regression output

$$\hat{M} = 45.6 + 0.21 H - 0.05 P$$

Give an interpretation of the coefficients of both regressions. Researcher B thinks that his negative coefficient of P is implausible. Is this correct?

(4)* Exercise 3.13 in the text.

In Exercise 3.7, the sample means of H , P , and R are 100 hours, 95 hours, and 5 hours, respectively and the standard deviations of the distributions of H , P , and R are 10.1, 10.1, and 2.1, respectively. The standard errors of the coefficients of the regression of Researcher A are shown in parentheses under the coefficients.

$$\hat{M} = 45.6 + 0.15 P + 0.21 R$$

(2.8) (0.03) (0.14)

Perform t tests of the significance of the coefficients of P and R . The researcher says that the insignificant coefficient of R is to be expected because the students, on average, spent much less time on revision than on primary study. Explain whether this assertion is correct.

(5) Exercise 3.10 in the text.

The following earnings functions were fitted separately for males and females, using *EAEF* Data Set 21 (standard errors in parentheses):

males

$$\widehat{EARNINGS} = -31.5168 + 3.1408 S + 0.6453 EXP$$

(7.8708) (0.3693) (0.2382)

females

$$\widehat{EARNINGS} = -17.2028 + 2.0772 S + 0.3179 EXP$$

(4.5797) (0.2805) (0.1388)

Using equation (3.36) in the text, explain why the standard errors of the coefficients of S and EXP are greater for the male subsample than for the female subsample, and why the difference in the standard errors is relatively large for EXP .

Further data:

	<i>males</i>	<i>females</i>
s_u	14.278	10.548
n	270	270
$r_{S,EXP}$	-0.4029	-0.0632
MSD(S)	6.6080	5.2573
MSD(EXP)	15.8858	21.4628

(6)* What kind of ability is important for educational attainment? Using your *EAEF* data set, regress S on SM , SF , $ASVAB02$, $ASVAB03$, and $ASVAB04$, the three components of the *ASVABC* composite score. Calculate correlation coefficients for the three *ASVAB* components. The two verbal components,

ASVAB03 (word knowledge) and *ASVAB04* (paragraph comprehension) are likely to be particularly highly correlated. Generate a composite, *VERBAL*, as their sum, and regress *S* on *SM*, *SF*, *ASVAB02*, and *VERBAL*. Compare the regression results.

Stata commands

```
reg S SM SF ASVAB02 ASVAB03 ASVAB04
cor ASVAB02 ASVAB03 ASVAB04
gen VERBAL = ASVAB03 + ASVAB04
reg S SM SF ASVAB02 VERBAL
```

(7) Exercise 3.18 in the text

Using your *EAEF* data set, fit an educational attainment function, regressing *S* on *ASVABC*, *SM*, and *SF*. Calculate the *F* statistic using the explained sum of squares and the residual sum of squares in the regression output, verify that it matches the *F* statistic in the output, and perform a test of the explanatory power of the equation as a whole. Also calculate the *F* statistic using R^2 and verify that it is the same.

Stata command

```
reg S ASVABC SM SF
```

(8)* Exercise 3.19 in the text.

Fit an educational attainment function using the specification in Exercise 3.18, adding the *ASVAB* speed test scores *ASVAB05* and *ASVAB06*. Perform an *F* test of the joint explanatory power of *ASVAB05* and *ASVAB06*, using the results of this regression and that in Exercise 3.18.

Stata command

```
reg S SM SF ASVABC ASVAB05 ASVAB06
```

For the class in the week beginning November 21

PROBLEM SET 6

(1)* Exercise 4.6 in the text.

What is the relationship between weight and height? Using your *EAEF* data set, regress the (natural) logarithm of *WEIGHT85* on the logarithm of *HEIGHT*. Interpret the regression results and perform appropriate tests.

Stata commands

```
gen LGWT85 = ln(WEIGHT85)
gen LGHEIGHT = ln(HEIGHT)
reg LGWT85 LGHEIGHT
```

(2)* Exercise 4.7 in the text

Suppose that the logarithm of *Y* is regressed on the logarithm of *X*, the fitted regression being

$$\log Y = b_1 + b_2 \log X$$

Suppose $Y^* = \lambda Y$, where λ is a constant, and suppose that $\log Y^*$ is regressed on $\log X$. Determine how the regression coefficients are related to those of the original regression. Determine also how the *t* statistic for b_2 and R^2 for the equation are related to those in the original regression.

(3)* Exercise 4.9 in the text.

Using your *EAEF* data set, regress the logarithm of earnings on *S* and *EXP*. Interpret the regression results and perform appropriate tests.

Stata commands

```
gen LGEARN = ln(EARNINGS)
reg LGEARN S EXP
```

Save the data set after generating *LGEARN* so that you do not have to define it again in the future. Give the new data set the same name as the old one, overwriting the old one.

(4) Exercise 4.10 in the text.

Using your *EAEF* data set, evaluate whether the dependent variable of an earnings function should be linear or logarithmic. Calculate the geometric mean of *EARNINGS* by taking the exponential of the mean of *LGEARN*. Define *EARNSTAR* by dividing *EARNINGS* by this quantity and calculate *LGEARNST* as its logarithm. Regress *EARNSTAR* and *LGEARNST* on *S* and *EXP* and compare the residual sums of squares.

Stata commands

```
sum LGEARN
```

In the next instruction, replace *Z* with the mean of *LGEARN* from the sum command:

```
gen EARNSTAR = EARNINGS/exp(Z)
gen LGEARNST = ln(EARNSTAR)
reg EARNSTAR S EXP
reg LGEARNST S EXP
```

(5) Exercise 4.12 in the text

Come to class prepared to discuss this item. A written answer is not required.

The output shows the result of regression of *S* on *SM* and its square, *SMSQ*. Evaluate the regression results. In particular, explain why the coefficient of *SM* is negative.

```
. gen SMSQ = SM*SM
. reg S SM SMSQ
```

Source	SS	df	MS			
Model	519.131914	2	259.565957	Number of obs =	540	
Residual	2685.85142	537	5.00158551	F(2, 537) =	51.90	
Total	3204.98333	539	5.94616574	Prob > F =	0.0000	
				R-squared =	0.1620	
				Adj R-squared =	0.1589	
				Root MSE =	2.2364	

S	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
SM	-.2564658	.1318583	-1.95	0.052	-.5154872	.0025556
SMSQ	.0271172	.0060632	4.47	0.000	.0152068	.0390277
_cons	12.79121	.7366358	17.36	0.000	11.34416	14.23825

For the class in the week beginning November 28

PROBLEM SET 7

- (1) Exercise 5.1 in the text.

Does the sex of an individual affect educational attainment? Using your *EAEF* data set, regress *S* on *ASVABC*, *SM*, *SF* and *MALE*, a dummy variable that is 1 for male respondents and 0 for female ones. Interpret the coefficients and perform *t* tests. Is there any evidence that the educational attainment of males is different from that of females?

Stata command

```
reg S ASVABC SM SF MALE
```

- (2)* Exercise 5.6 in the text.

Does ethnicity affect educational attainment? In your *EAEF* data set you will find the following ethnic dummy variables:

ETHHISP 1 if hispanic, 0 otherwise
ETHBLACK 1 if black, 0 otherwise
ETHWHITE 1 if not hispanic or black, 0 otherwise.

Regress *S* on *ASVABC*, *MALE*, *SM*, *SF*, *ETHBLACK*, and *ETHHISP*. (In this specification *ETHWHITE* has been chosen as the reference category, and so it is omitted.) Interpret the regression results and perform *t* tests on the coefficients.

Stata command

```
reg S ASVABC MALE SM SF ETHBLACK ETHHISP
```

- (3)* Exercise 5.10 in the text.

Evaluate whether the ethnicity dummies as a group have significant explanatory power for educational attainment by comparing the residual sums of squares in the regressions in Exercises 5.1 and 5.6.

- (4) Exercise 5.17 in the text.

Is the effect of education on earnings different for males and females? Using your *EAEF* data set, define a slope dummy variable *MALES* as the product of *MALE* and *S*:

$$MALES = MALE * S$$

Regress *LGEARN* on *S*, *EXP*, *ETHBLACK*, *ETHHISP*, *MALE*, and *MALES*, interpret the equation and perform appropriate statistical tests.

Stata commands

```
gen MALES = MALE*S
reg LGEARN S EXP ETHBLACK ETHHISP MALE MALES
```

- (5)* Exercise 5.19 in the text.

Are educational attainment functions are different for males and females? Using your *EAEF* data set, regress *S* on *ASVABC*, *ETHBLACK*, *ETHHISP*, *SM*, and *SF* (do not include *MALE*). Repeat the regression using only the male respondents. Repeat it again using only the female respondents. Perform a Chow test.

Stata commands

```
reg S ASVABC ETHBLACK ETHHISP SM SF
reg S ASVABC ETHBLACK ETHHISP SM SF if MALE==1
reg S ASVABC ETHBLACK ETHHISP SM SF if MALE==0
```

(6)* Exercise 5.21 in the text.

Are there differences in male and female educational attainment functions? This question has been answered by Item 5 but nevertheless it is instructive to investigate the issue using the dummy variable approach. Using your *EAEF* data set, define the following slope dummies combining *MALE* with the parental education, cognitive ability, and ethnicity variables:

$$\begin{aligned} MALESM &= MALE * SM \\ MALESF &= MALE * SF \\ MALEASVC &= MALE * ASVABC \\ MALEBLAC &= MALE * ETHBLACK \\ MALEHISP &= MALE * ETHHISP \end{aligned}$$

and regress *S* on *ETHBLACK*, *ETHHISP*, *ASVABC*, *SM*, *SF*, *MALE*, *MALEBLAC*, *MALEHISP*, *MALEASVC*, *MALESM*, and *MALESF*. Next regress *S* on *ETHBLACK*, *ETHHISP*, *ASVABC*, *SM*, and *SF* only. Examine the correlations among the male intercept and slope dummy variables. Perform an *F* test of the joint explanatory power of *MALE* and the slope dummy variables as a group (verify that the *F* statistic is the same as in Exercise 5.19) and perform *t* tests on the coefficients of the slope dummy variables in the first regression. What was the point of computing the correlations between the male intercept and slope dummy variables?

Stata commands

```
gen MALESM = MALE * SM
gen MALESF = MALE * SF
gen MALEASVC = MALE * ASVABC
gen MALEBLAC = MALE * ETHBLACK
gen MALEHISP = MALE * ETHHISP
reg S ASVABC ETHBLACK ETHHISP SM SF MALE MALESM MALESF MALEASVC
    MALEBLAC MALEHISP
reg S ETHBLACK ETHHISP ASVABC SM SF
cor MALE MALESM MALESF MALEASVC MALEBLAC MALEHISP
```

For the class in the week beginning December 5

PROBLEM SET 8**(1)*** Exercise 6.2 in the text.

Using your *EAEF* data set, regress *LGEARN* (1) on *S* and *EXP*, (2) on *S* only, and (3) on *EXP* only. Calculate the correlation between *S* and *EXP*. Compare the coefficients of *S* in regressions (1) and (2). Give both mathematical and intuitive explanations of the direction of the change. Also compare the coefficients of *EXP* in regressions (1) and (3) and explain the direction of the change.

Stata commands

```
reg LGEARN S EXP
reg LGEARN S
reg LGEARN EXP
cor S EXP
```

(2)* Exercise 6.3 in the text

Using your *EAEF* data set, regress *LGEARN* (1) on *S*, *EXP*, *MALE*, *ETHBLACK*, and *ETHHISP*, and (2) on *S*, *EXP*, *MALE*, *ETHBLACK*, *ETHHISP*, and *ASVABC*. Calculate the correlation coefficients for the explanatory variables and discuss the differences in the regression results. (A detailed mathematical analysis is not expected.)

Stata commands

```
reg LGEARN S EXP MALE ETHBLACK ETHHISP
reg LGEARN S EXP MALE ETHBLACK ETHHISP ASVABC
cor S EXP MALE ETHBLACK ETHHISP ASVABC
```

(3)* Variation on Exercise 6.8 in the text.

Using your *EAEF* data set, regress *LGEARN* on *S*, *EXP*, *ASVABC*, *MALE*, *ETHBLACK*, and *ETHHISP*. Repeat the regression, adding *SM* and *SF*. Calculate the correlations between *SM* and *SF* and the other explanatory variables. Compare the results of the two regressions.

Stata commands

```
reg LGEARN S EXP ASVABC MALE ETHBLACK ETHHISP
reg LGEARN S EXP ASVABC MALE ETHBLACK ETHHISP SM SF
cor SM SF S EXP ASVABC MALE ETHBLACK ETHHISP
```

(4) A Monte Carlo experiment uses the following model:

$$Y = -7,125 + 1,500X_2 + 25X_3 + u$$

There is a positive correlation between X_2 and X_3 . u is generated as a multiple of a random variable drawn from a normal distribution with zero mean and unit variance. In one replication of the experiment, when Y is regressed against X_2 , omitting X_3 , the coefficient of X_2 is 1,478. Is this what you would expect? If not, explain what has happened.

For the class in the week beginning January 9

PROBLEM SET 9**(0)** Write down the name and office hour of your class teacher.**(1)*** Exercise 6.9 in the text.

Is potential work experience a satisfactory proxy for actual work experience? Length of work experience is generally found to be an important determinant of earnings. Many data sets do not contain this variable. To avoid the problem of omitted variable bias, a standard practice is to use *PWE*, potential years of work experience, as a proxy. *PWE* is defined as *AGE*, less age at completion of full-time education (years of schooling plus 5, assuming that schooling begins at the age of 6):

$$PWE = AGE - S - 5.$$

Using your *EAEF* data set, regress *LGEARN* (1) on *S*, *ASVABC*, *MALE*, *ETHBLACK*, *ETHHISP*, (2) on *S*, *ASVABC*, *MALE*, *ETHBLACK*, *ETHHISP*, and *PWE* and (3) on *S*, *ASVABC*, *MALE*, *ETHBLACK*, *ETHHISP*, and *EXP*. Compare the results and evaluate whether *PWE* would have been a satisfactory proxy for *EXP* if data for *EXP* had not been available.

Variation: *PWE* is not likely to be a satisfactory proxy for work experience for females because it

does not take into account time spent out of the labor force rearing children. Investigate this by running the three regressions for the male and female subsamples separately. You must drop the *MALE* dummy from the specification (explain why).

Stata commands

```
gen PWE = AGE - S - 5
reg LG EARN S ASVABC MALE ETHBLACK ETHHISP
reg LG EARN S ASVABC MALE ETHBLACK ETHHISP PWE
reg LG EARN S ASVABC MALE ETHBLACK ETHHISP EXP
```

Variation:

```
reg LG EARN S ASVABC MALE ETHBLACK ETHHISP if MALE==1
reg LG EARN S ASVABC MALE ETHBLACK ETHHISP PWE if MALE==1
reg LG EARN S ASVABC MALE ETHBLACK ETHHISP EXP if MALE==1
reg LG EARN S ASVABC MALE ETHBLACK ETHHISP if MALE==0
reg LG EARN S ASVABC MALE ETHBLACK ETHHISP PWE if MALE==0
reg LG EARN S ASVABC MALE ETHBLACK ETHHISP EXP if MALE==0
```

(2)* Exercise 6.11 in the text.

Is previous work experience as valuable as experience with the current employer? Using your *EAEF* data set, first regress *LG EARN* on *S*, *EXP*, *MALE*, *ETHBLACK*, and *ETHHISP*. Then define

$$PREVEXP = EXP - TENURE.$$

The variable *TENURE* in your data set is the number of years spent working with the current employer. Regress *LG EARN* on *S*, *PREVEXP*, *TENURE*, *MALE*, *ETHBLACK*, and *ETHHISP*. The estimates of the coefficients of *PREVEXP* and *TENURE* will be different. This raises the issue of whether the difference is due to random factors or whether the coefficients are significantly different. Set up the null hypothesis $H_0: \delta_1 = \delta_2$, where δ_1 is the coefficient of *PREVEXP* and δ_2 is the coefficient of *TENURE*. Explain why the regression with *EXP* is the correct specification if H_0 is true, while the regression with *PREVEXP* and *TENURE* should be used if H_0 is false. Perform an *F* test of the restriction using *RSS* for the two regressions. Do this for the combined sample and also for males and females separately.

Stata commands

```
gen PREVEXP = EXP - TENURE
reg LG EARN EXP S MALE ETHBLACK ETHHISP
reg LG EARN PREVEXP TENURE S MALE ETHBLACK ETHHISP
reg LG EARN EXP S MALE ETHBLACK ETHHISP if MALE==1
reg LG EARN PREVEXP TENURE S MALE ETHBLACK ETHHISP if MALE==1
reg LG EARN EXP S MALE ETHBLACK ETHHISP if MALE==0
reg LG EARN PREVEXP TENURE S MALE ETHBLACK ETHHISP if MALE==0
```

(3)* Exercise 6.12 in the text.

Using your *EAEF* data set, regress *LG EARN* on *S*, *EXP*, *MALE*, *ETHBLACK*, *ETHHISP*, and *TENURE*. Demonstrate that a *t* test on the coefficient of *TENURE* is a test of the restriction described in Exercise 6.11. Verify that the same result is obtained. Do this for the combined sample and also for males and females separately.

Stata commands

```
reg LG EARN S EXP MALE ETHBLACK ETHHISP TENURE
reg LG EARN S EXP MALE ETHBLACK ETHHISP TENURE if MALE==1
reg LG EARN S EXP MALE ETHBLACK ETHHISP TENURE if MALE==0
```