WRITTEN PAPER I (ECON 4135)

September 16, 2009

Due to budget constraint can we not correct or comment on individual papers. Students are however encouraged to do the work and write a paper, and to discuss it in their colloquium group. The problem set will be discussed in the seminar in week 42.

The main purpose of this exercise is to practice linear regression with one regressor and to get acquainted with STATA. *For detailed instructions and examples about running STATA see the home page of the Statistics Department at UCLA at* http://www.ats.ucla.edu/STAT/STATA/ *. Here you'll even find movies! (Of course, STATA's Help menu may also be useful).*

We will examine the relationship between log earnings, $\ln Y$, where Y is total annual taxable labor income, and some potential explanatory variables. Two STATA data sets accompany this problem set: earningsdata_males.dta and earningsdata_females.dta. The data sets contain individual level information about log earnings, years of schooling, labor market experience, type of education, sector of occupation and work place region ("fylke"). The observation units in earningsdata_males.dta are 5859 native-born males who lived in Norway in 1970, was born between 1952 and 1970, and who still lived in Norway in 1997. The data set earningsdata_females.dta contains the same variables for 3247 females.

The samples (males and females) are randomly drawn from the Norwegian system of register data in 1997 and is restricted to full-time wage-earners, defined as individuals working 30 hours or more per week. The measure of years of schooling is defined as the standard number of years necessary to complete this level. Labor market experience is represented as age minus years of schooling minus 7 years, i.e. *potential* (not *actual)* experience. The earnings measure reflects annual earnings, observations whose employment relationships started or terminated within the actual year were excluded. Holders of multiple jobs and individuals that received labor market compensation or have participated in active labor market programs have been excluded. A full list of variables is given in Table 1, although you may use only some of them in this exercise.

1 . Compute the sample mean $\overline{\ln Y}$ for average log of earnings ($\ln Y$) for both men and women. Compute 90% and 95% confidence intervals for the population mean of $\ln Y$. Formulate a statistical test of whether the mean of $\ln Y$ are different for men and women. Calculate the p-value of the test. Do you reject H_0 at the 1% level of significance?

2. Now compute the sample mean for earnings ($\overline{Y}$). Do you find that $\overline{Y} = \exp(\overline{\ln Y})$? explain!

3. Use the data set for males and estimate the regression equation

$$\ln Y_i = \alpha + \beta_1 S_i + u_i. \tag{1}$$

Specify your assumptions about the error term u_i: What assumptions are need for the OLS estimator to be i) unbiased, ii) consistent and iii) asymptotically normally distributed?

4. According to your estimates, what is the conditional expectation $E(\ln Y_i|S)$ for men and women, respectively? Calculate the expected log earnings for a man and woman, respectively, with 12 year schooling. Does schooling account for a large fraction of the variance in earnings across individuals? Explain.

5. Give a 95% confidence interval for β_1. What is your estimated expected marginal return to schooling, $\partial E(\ln Y_i|S_i)/\partial S_i$? What is the estimated percentage increase in income of one additional year of schooling? Do these calculations both for men and women.

6. We want to test whether the returns to schooling is different for men and women. Set up an appropriate test statistic and carry out a formal statistical test. Discuss the results.

7. If the main interest parameter is β_1, i.e. the marginal returns to schooling, do you see any problems with running a univariate regression? Discuss possible sources of bias of your estimates.

8. Now try out for yourself some of the other possible explanatory variables in your data set. Choose one of the data sets. Compare different univariate regressions and compare their R^2. Comment on the results. What does R^2 say? Which of the regressors you try appear to explain most of the variation in the earnings data.

Table 1: **Variable list**

Variable name	**Description**
lnY	Log of earnings
S	Years of schooling exceeding 7 years
E	Years of experience
E^2	Years of experience squared
public	Sector of occupation= public services (1 if true, 0 else)
servi	Sector of occupation= private services
unsp	Type of education=unspecified (1 if true, 0 else)
gen	Type of education=general
hum	Type of education=humanities
teach	Type of education=teaching
admin	Type of education=business/adminstrative
transp	Type of education=transport
health	Type of education=health
farm	Type of education=farming/fisheries
serv	Type of education=services/military
ostf	Region=Østfold (1 if true, 0 else)
akershus	Region=Akershus
hedemark	⋮
oppland	⋮
buskerud	⋮
vestfold	⋮
telemark	⋮
a-agder	⋮
v-agder	⋮
rogaland	⋮
hordaland	⋮
sognfj	⋮
moreroms	⋮
s-tr	⋮
n-tr	⋮
nordland	⋮
troms	⋮
finmark	Region=Finmark