

EDUCATIONAL ATTAINMENT AND EARNINGS FUNCTIONS

REGRESSION EXERCISES

c.dougherty@lse.ac.uk
September 2005

TABLE OF CONTENTS

1. **Description of the data sets**
2. **Exercises**
3. **Interpretation of a semilogarithmic regression**
4. **Interpretation of a logarithmic regression**
5. **Explanation of the Oaxaca decomposition**
6. **Further details of the variables**

DOWNLOADING A DATA SET

You will find data sets in three different formats at the Introduction to Econometrics website, URL <http://econ.lse.ac.uk/courses/ec220>. There are special versions for Stata and EViews users, and a further version in ASCII (text) format for everybody else. In each case there are 22 parallel data files, as described in the next section.

The first twenty data sets are intended for use by members of a workshop. At the beginning of the course, the workshop instructor should assign a different data set to each member of the workshop. If you are working on your own, choose any one of the 20. Data Set 21 is used in examples in the text. You can use it to replicate the examples if you so wish. Data Set 22 is intended for use by instructors.

To download, click on the name of the chosen data set and follow the instructions in the dialogue box. A Stata format data set should be ready for use. At the present time, an EViews data set will need renaming. For example, EViews Data Set 5 will download with filename EAEF05_wf1.bin. It needs to be renamed EAEF05.wf1 before EViews will recognize it.

1. DESCRIPTION OF THE DATA SET

In view of its relevance for social policy, it is not surprising that analysis of the closely related topics of the determinants of educational attainment and the determinants of earnings has long been a major application of econometrics. Particularly sensitive issues are those relating to differences in educational attainment and earnings attributable to ethnicity, sex, and genetic endowment, to interactions in the effects of these factors, and to changes through time. The data sets described here will allow you to explore some of these issues using a subset of a major US data-base, the National Longitudinal Survey of Youth 1979– (NLSY79).

NLSY79 is a panel survey with repeated interviews of a nationally representative sample of young males and females aged 14 to 21 in 1979. From 1979 to 1994 the interviews took place annually. Since 1994 they have been conducted at two-year intervals. The core sample originally consisted of 3,003 males and 3,108 females. In addition there are special supplementary samples (some now discontinued) of ethnic minorities, those in poverty, and those serving in the armed forces. Extensive background information was obtained in the base-year survey in 1979 and since then information has been updated each year on education, training, employment, marital status, fertility, health, child care and assets and income. In addition special sections have been added from time to time on other topics – for example, drug use. The surveys have been extremely detailed and the quality of the execution of the survey is very high. As a consequence NLSY79 is regarded as one of the most important data bases available to social scientists working with US data.

For the practical work there are 20 parallel data subsets each consisting of 540 observations, 270 drawn randomly from the male respondents in the source data set and the same number drawn randomly from the female respondents. Each subset contains data for each respondent on the following variables (C indicates a continuous variable, D a dummy variable):

Personal variables

<i>FEMALE</i>	D	Sex of respondent (0 if male, 1 if female)
<i>MALE</i>	D	Sex of respondent (1 if male, 0 if female)
<i>Ethnicity:</i>		
<i>ETHBLACK</i>	D	Black
<i>ETHHISP</i>	D	Hispanic
<i>ETHWHITE</i>	D	Non-black, non-hispanic
<i>AGE</i>	C	Age in 2002
<i>S</i>	C	Years of schooling (highest grade completed as of 2002)
<i>Highest educational qualification:</i>		
<i>EDUCPROF</i>	D	Professional degree
<i>EDUCPHD</i>	D	Doctorate
<i>EDUCMAST</i>	D	Master's degree
<i>EDUCBA</i>	D	Bachelor's degree
<i>EDUCAA</i>	D	Associate's (two-year college) degree
<i>EDUCHSD</i>	D	High school diploma or equivalent
<i>EDUCDO</i>	D	High school drop-out

Marital status

<i>SINGLE</i>	D	Single, never married
<i>MARRIED</i>	D	Married, spouse present
<i>DIVORCED</i>	D	Divorced or separated

Scaled score on a component of the ASVAB battery (see Section 6 for further details on the ASVAB variables)

<i>ASVAB2</i>	C	Arithmetic reasoning
<i>ASVAB3</i>	C	Word knowledge
<i>ASVAB4</i>	C	Paragraph comprehension
<i>ASVAB5</i>	C	Numerical operations (speed test)
<i>ASVAB6</i>	C	Coding speed (speed test)
<i>ASVABC</i>	C	Composite of <i>ASVAB2</i> (with double weight), <i>ASVAB3</i> and <i>ASVAB4</i>

Faith:

<i>FAITHN</i>	D	None
<i>FAITHC</i>	D	Catholic
<i>FAITHJ</i>	D	Jewish
<i>FAITHP</i>	D	Protestant
<i>FAITHO</i>	D	Other
<i>HEIGHT</i>	C	Height, in inches, in 1985
<i>WEIGHT85</i>	C	Weight, in pounds, in 1985
<i>WEIGHT02</i>	C	Weight, in pounds, in 2002

Family background variables

<i>SM</i>	C	Years of schooling of respondent's mother
<i>SF</i>	C	Years of schooling of respondent's father
<i>SIBLINGS</i>	C	Number of siblings
<i>Living at age 14:</i>		
<i>L14TOWN</i>	D	in a town or city
<i>L14COUN</i>	D	in the country, not on a farm
<i>L14FARM</i>	D	on a farm
<i>LIBRARY</i>	D	Member of family possessed a library card when respondent was 14
<i>POV78</i>	D	Family living in poverty in 1978

Work-related variables

<i>EARNINGS</i>	C	Current hourly earnings in \$ reported at the 2002 interview
<i>HOURS</i>	C	Usual number of hours worked per week, 2002 interview
<i>TENURE</i>	C	Tenure (years) with current employer at the 2002 interview
<i>EXP</i>	C	Total out-of-school work experience (years) as of the 2002 interview.
<i>COLLBARG</i>	D	Pay set by collective bargaining, 2002

Category of employment

<i>CATGOV</i>	D	Government
<i>CATPRI</i>	D	Private sector
<i>CATSE</i>	D	Self-employment
<i>URBAN</i>	D	Living in an urban area at 2002 interview

Living in 2002 in:

<i>REGNC</i>	D	North central census region
<i>REGNE</i>	D	North eastern
<i>REGS</i>	D	Southern
<i>REGW</i>	D	Western

Not all of the variables are used in the exercises suggested in the next section. You should feel free to experiment by trying alternative regression specifications with the extra variables. Further details of some of the variables are given in Section 6.

2. EXERCISES

Note: In all of the regressions you should include a constant. Most regression packages do this automatically for you. Some, like EViews and its TSP predecessors, require you to specify a constant if you wish to include one.

Exercise 1 Simple regression analysis

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 Section 6). Perform a regression of S on $ASVABC$ and interpret the regression results. Comment on the value of R^2 .

Exercise 2 Simple regression analysis

Do earnings depend on education?

$EARNINGS$ is the hourly earnings of the respondent, in dollars, in 2002. Perform a regression of $EARNINGS$ on S and interpret the regression results. Comment on the value of R^2 .

Exercise 3 t tests of regression coefficients and confidence intervals

- Perform a t test on the coefficients of the regression in Exercise 1. Explain the implications of the result of the test. Calculate a 95% confidence interval for the slope coefficient.
- Perform a t test on the coefficients of the regression in Exercise 2. Calculate a 95% confidence interval for the slope coefficient.

Exercise 4 F test of the goodness of fit of a regression equation

Calculate the F statistic corresponding to the value of R^2 in Exercise 1 and verify that it is equal to that printed in the regression results. Perform an F test for the goodness of fit.

Exercise 5 Multiple regression analysis

Does educational attainment depend on parents' education?

Regress S on $ASVABC$ and SM , the years of schooling of the respondent's mother. Interpret the regression results and perform t tests. Repeat the regression using SF , the years of schooling of the father, instead of SM , and again including both 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 education of her children. Do your regression results support this view?

Exercise 6 Multiple regression analysis

Do earnings depend on work experience as well as education?

Regress $EARNINGS$ on S and EXP , interpret the regression results, and perform t tests.

Exercise 7 Multiple regression analysis

Use of the Frisch–Waugh method of representing the relationship between the dependent variable and one explanatory variable.

Using your *EAEF* data set, make a graphical representation of the relationship between *S* and *SM* using the technique described above, assuming that the true model is as in Exercise 5. To do this, regress *S* on *ASVABC* and *SF* and save the residuals. Do the same with *SM*. Plot the *S* and *SM* residuals. Also regress the former on the latter, and verify that the slope coefficient is the same as that obtained in Exercise 5.

Exercise 8 Explanation of the size of a standard error

If the regression model

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

is fitted using OLS, and if the specification is correct and if the Gauss–Markov conditions are satisfied, the standard error of b_2 is given by

$$s.e.(b_2) = \frac{s_u}{\sqrt{n \text{Var}(S)}} \times \frac{1}{\sqrt{1 - r_{S,EXP}^2}}$$

Using this expression, explain why the standard errors of the coefficients of *S* and *EXP* are greater for the male subsample than for the female subsample (standard errors in parentheses; the regressions use Data Set 21).

males

$$EARNINGS = -33.1526 + 3.7042S + 0.3106EXP$$

(9.2803) (0.4395) (0.2870)

females

$$EARNINGS = -11.2484 + 1.7125S + 0.2150EXP$$

(3.6135) (0.2199) (0.1130)

Further data:

	<i>males</i>	<i>females</i>
s_u	18.2360	8.2122
n	270	270
$r_{S,EXP}$	-0.3483	-0.0261
$\text{Var}(S)$	7.2555	5.1665
$\text{Var}(EXP)$	17.0172	19.5683

Exercise 9 Correlated explanatory variables

What kind of ability is important for educational attainment?

Regress S on SM , SF and $ASVAB2$, $ASVAB3$, and $ASVAB4$, the three components of the $ASVABC$ composite score. Compare the coefficients and their standard errors with those of $ASVABC$ in a regression of S on SM , SF and $ASVABC$. Calculate correlation coefficients for the three $ASVAB$ components.

Exercise 10 Correlated explanatory variables: use of a restriction

Investigate the determinants of family size by regressing $SIBLINGS$ on SM and SF using your $EAEF$ data set. SM and SF are likely to be highly correlated (find the correlation in your data set) and the regression may be subject to multicollinearity. Introduce the restriction that the theoretical coefficients of SM and SF are equal and run the regression a second time replacing SM and SF by their sum, SP . Evaluate the regression results.

Exercise 11 F tests of the goodness of fit in a multiple regression model

Perform an F test, stating clearly your null and alternative hypotheses,

- (a) for the regression in Exercise 5
- (b) for the regression in Exercise 6

Exercise 12 Nonlinear regression analysis

The dependence of earnings on education – an alternative model specification

Define a new variable $LGEARN$ as the (natural) logarithm of $EARNINGS$, regress it on S and EXP , and interpret the regression results (for the interpretation of a model of this type, see Section 3 of this manual). Perform a t test on the slope coefficient.

Note: Save the data set after defining $LGEARN$. If you do not, you will have to define $LGEARN$ again the next time you wish to use it.

Exercise 13 Nonlinear regression analysis

What is the relationship between weight and height?

Define $LGWT85$ and $LGHEIGHT$ as the (natural) logarithms of $WEIGHT85$ and $HEIGHT$, respectively, and regress $LGWT85$ on $LGHEIGHT$. Give an interpretation of the regression results, perform a t test on the coefficient of $LGHEIGHT$, and explain the implications of the regression results.

Note: The interpretation of the coefficients is different from that in Exercise 12 because both the dependent and the explanatory variables are in logarithmic form. For help with the interpretation, see Section 4 of this manual.

Exercise 14 Box–Cox test

Discriminating between functional forms using a Box–Cox test

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. Perform a Box–Cox test.

Exercise 15 Regression analysis with a dummy variable

Does the sex of an individual affect educational attainment?

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?

Exercise 16 Regression analysis with a dummy variable

Is there sex discrimination in earnings?

Regress *LGEARN* on *S*, *EXP*, and *MALE*. Interpret the coefficients and perform *t* tests. See Section 3 for guidance concerning the interpretation of dummy variable coefficients in a semi-logarithmic model.

Exercise 17 Regression analysis with multiple categories of dummy variable

Does ethnicity affect educational attainment?

In the 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.

Exercise 18 Regression analysis with multiple categories of dummy variable

Are earnings subject to ethnic discrimination?

Regress *LGEARN* on *S*, *EXP*, *MALE*, *ETHHISP*, and *ETHBLACK*. Interpret the regression results and perform *t* tests on the coefficients.

Exercise 19 F test of the explanatory power of a group of dummy variables

- 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 15 and 17.
- Do the same for earnings by comparing the residual sums of squares in the regressions in Exercises 16 and 18.

Exercise 20 Evaluation of the effect of changing the omitted category in a regression with dummy variables

Repeat Exercise 17 making *ETHBLACK* the reference (omitted) category. Evaluate the impact on the interpretation of the coefficients and the statistical tests.

Exercise 21 Evaluation of the effect of changing the omitted category in a regression with dummy variables

Repeat Exercise 18 making *ETHBLACK* the reference (omitted) category. Evaluate the impact on the interpretation of the coefficients and the statistical tests.

Exercise 22 Dummy variable trap

Repeat Exercise 16 including *FEMALE* as well as *MALE*. Regress *LGEARN* on *S*, *EXP*, *MALE*, and *FEMALE*. Discuss the regression results.

Exercise 23 Slope dummy variable

Is the effect of the ASVABC score on educational attainment different for males and females?

Define a slope dummy variable *MALEASVC* as the product of *MALE* and *ASVABC*:

$$MALEASVC = MALE * ASVABC$$

Regress *S* on *ASVABC*, *SM*, *SF*, *ETHBLACK*, *ETHHISP*, *MALE*, and *MALEASVC*, interpret the equation and perform appropriate statistical tests.

Exercise 24 Slope dummy variable

Is the effect of education on earnings different for males and females?

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.

Exercise 25 Interactive dummy variable

Are there ethnic variations in the effect of the sex of a respondent on educational attainment?

A special case of a slope dummy variable is the interactive dummy variable defined as the product of two dummy variables. Define interactive dummy variables *MALEBLAC* and *MALEHISP* as the product of *MALE* and *ETHBLACK*, and of *MALE* and *ETHHISP*, respectively:

$$MALEBLAC = MALE * ETHBLACK$$

$$MALEHISP = MALE * ETHHISP$$

Regress *S* on *ASVABC*, *SM*, *SF*, *MALE*, *ETHBLACK*, *ETHHISP*, *MALEBLAC*, and *MALEHISP*. Interpret the regression results and perform tests.

Exercise 26 Chow test

Are educational attainment functions are different for males and females?

Regress S on $ASVABC$, $ETHBLACK$, $ETHHISP$, SM , and SF (do not include $MALE$). Repeat the regression using only the observations for the male respondents. Repeat it again using only the observations for the female respondents. Perform a Chow test.

Exercise 27 Chow test

Are earnings functions different for males and females?

Regress $LGEARN$ on S , EXP , $ETHBLACK$, and $ETHHISP$ (do not include $MALE$). Repeat the regression using the observations for the male respondents. Repeat it again using the observations for the female respondents. Perform a Chow test.

Exercise 28 Comparison of a Chow test with an F test using a full set of dummy variables

Are there differences in male and female educational attainment functions?

This question has been answered by Exercise 26 but nevertheless it is instructive to investigate the issue using the dummy variable approach. Define the following slope dummies combining $MALE$ with the parental education variables:

$$MALESM = MALE * SM$$

$$MALESF = MALE * SF$$

and regress S on $ASVABC$, $ETHBLACK$, $ETHHISP$, SM , SF , $MALE$, $MALEASVC$, $MALEBLAC$, $MALEHISP$, $MALESM$, and $MALESF$. Next regress S on $ASVABC$, $ETHBLACK$, $ETHHISP$, SM , and SF only. Calculate the correlation matrix for $MALE$ and the slope dummies. 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 26) and then perform t tests on the coefficients of the slope dummy variables in the first regression. Comment on the test results.

Exercise 29 Comparison of a Chow test with an F test using a full set of dummy variables

Where are the differences in male and female earnings functions?

Define a slope dummy variable $MALEEXP$ as the product of $MALE$ and EXP :

$$MALEEXP = MALE * EXP$$

Regress $LGEARN$ on S , EXP , $ETHBLACK$, $ETHHISP$, $MALE$, $MALES$, $MALEEXP$, $MALEBLAC$, and $MALEHISP$. Next regress $LGEARN$ on S , EXP , $ETHBLACK$, and $ETHHISP$ only. Calculate the correlation matrix for $MALE$ and the slope dummies. 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 27) and then perform t tests on the coefficients of the slope dummy variables in the first regression. Comment on the test results.

Exercise 30 Oaxaca decomposition of the difference in male and female earnings

In Exercise 27, using a Chow test, you will probably have discovered that the earnings functions are different for males and females. In Exercise 29 you will have gone further and have found out which characteristics have significantly different coefficients. Now we will quantify the factors responsible for the difference in male and female earnings. Differences in the coefficients are only part of the story. Differences in the sizes of the variables can also be important. In this exercise you will quantify the effects of both.

Using the male subsample, regress $LGEARN$ on S , EXP , $ETHBLACK$, and $ETHHISP$. Still with the male subsample, find the mean value of $LGEARN$ and the explanatory variables. Do the same for the female subsample.

- (a) Provide an interpretation of the difference in the mean values of $LGEARN$ for males and females.
- (b) Decompose the difference using the Oaxaca decomposition described in Section 5.

Exercise 31 Omitted variable bias

Regress S (1) on $ASVABC$ and SM , (2) on $ASVABC$ only, and (3) on SM only. Compare the coefficients of $ASVABC$ in regressions (1) and (2) and explain the direction of the change. Also compare the coefficients of SM in regressions (1) and (3) and explain the direction of the change.

Exercise 32 Omitted variable bias

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.

Exercise 33 Omitted variable bias

Regress $LGEARN$ (1) on S , EXP , $MALE$, $ETHHISP$ and $ETHBLACK$, and (2) on S , EXP , $MALE$, $ETHHISP$, $ETHBLACK$, 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.)

Exercise 34 Redundant explanatory variable

Regress $LGEARN$ on S , EXP , $ASVABC$, $MALE$, $ETHHISP$ and $ETHBLACK$. Repeat the regression adding $SIBLINGS$. Calculate the correlations between $SIBLINGS$ and the other explanatory variables. Compare the results of the two regressions.

Exercise 35 Proxy variables

Do earnings depend on length of 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):

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

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).

Exercise 36 F test of a restriction

Is previous work experience as valuable as experience with the current employer?

The variable *EXP* measures total work experience apart from jobs taken while still in school or college. (The reason for eliminating these is that they are likely to be makeshift jobs that are irrelevant to subsequent employment.) *TENURE* measures work experience with the current employer. Define another variable *EXPBEF*, experience before working for the current employer, as

$$EXPBEF = EXP - TENURE$$

and regress *LGEARN* (1) on *EXP*, *S*, *ASVABC*, *MALE*, *ETHBLACK*, and *ETHHISP*, and (2) on *EXPBEF*, *TENURE*, *S*, *ASVABC*, *MALE*, *ETHBLACK*, and *ETHHISP*. Set up the null hypothesis $H_0: \delta_1 = \delta_2$, where δ_1 is the coefficient of *EXPBEF* and δ_2 is the coefficient of *TENURE* in the second regression. Explain why the regression with *EXP* is the correct specification if H_0 is true, while the regression with *EXPBEF* and *TENURE* should be used if H_0 is false. Compare the coefficients of *EXPBEF* and *TENURE* in the second regression. Do they appear to be similar? 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.

Exercise 37 t test of a restriction

Regress *LGEARN* on *S*, *ASVABC*, *EXP*, *MALE*, *ETHBLACK*, *ETHHISP*, and *TENURE*. Demonstrate that a *t* test on the coefficient of *TENURE* is a test of the restriction described in the previous exercise. Verify that the same result is obtained. Do this for the combined sample and also for males and females separately.

Exercise 38 Heteroscedasticity

Choose a specification for an earnings function with *EARNINGS* as the dependent variable. (You could use that in Exercise 6, but you may add additional variables if you wish). Sort the observations by size of *S* (your instructor will have to tell you how to do this) and perform a Goldfeld–Quandt test to test for heteroscedasticity in the *S* dimension. Repeat using *LGEARN* as the dependent variable.

Exercise 39 Measurement error, instrumental variables estimation and Hausman test

It is possible that the *ASVABC* test score is a poor measure of the kind of ability relevant for earnings. Accordingly, perform an OLS regression of *LGEARN* on *S*, *EXP*, *ASVABC*, *MALE*, *ETHBLACK*, and *ETHHISP* using your *EAEF* data set and an IV regression using *SM* as an instrument for *ASVABC*. Perform a Hausman test to evaluate whether *ASVABC* appears to be subject to measurement error. Consider repeating the exercise using *SM*, *SF*, *SIBLINGS*, and *LIBRARY* jointly as instruments for *ASVABC*.

Exercise 40 Simultaneous equations estimation

In principle *ASVABC* might be a positive function of *S*, in which case the educational attainment model should have two equations:

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

$$ASVABC = \alpha_1 + \alpha_2 S + v$$

Fit the second equation, (1) using OLS, (2) using instrumental variables estimation with *SM* as an instrument. Investigate analytically the likely direction of the bias in the slope coefficient in the OLS regression, and check whether a comparison of the OLS and IV estimates confirms your analysis.

Exercise 41 Binary choice models

What are the factors influencing going to college?

Define a binary variable *COLLEGE* that is equal to 1 if $S > 12$, 0 otherwise. Regress *COLLEGE* on *ASVABC*, *MALE*, *SM*, and *SF*

- (a) using ordinary least squares
- (b) using logit analysis
- (c) using probit analysis

Calculate the marginal effects using logit and probit analysis and compare them with those obtained using OLS.

Exercise 42 Sample selection bias

Does sample selection bias affect the OLS estimate of the return to college education?

Using your *EAEF* data set, investigate whether there is evidence that selection bias affects the least squares estimate of the returns to college education. Define $COLLYEAR = S - 12$ if $S > 12$, 0 otherwise, and $LGEARNCL = LGEARN$ if $COLLYEAR > 0$, missing otherwise. Use the heckman procedure to regress *LGEARNCL* on *COLLYEAR*, *EXP*, *ASVABC*, *MALE*, *ETHBLACK*, and *ETHHISP*, with *ASVABC*, *MALE*, *ETHBLACK*, *ETHHISP*, *SM*, *SF*, and *SIBLINGS* being used to determine whether the respondent attended college. Run the equivalent regression using least squares. Comment on your findings.

Exercise 43 Mini-project

Summarize what you have discovered concerning the determinants of educational attainment and earnings, running additional regressions if desired. Include a table giving descriptive statistics for

the variables in the introduction to your report. It is a good idea to go further and provide descriptive statistics of the more important variables separately for major subgroups, for example, for males and females. You can then compare the absolute proportion difference in, say, mean earnings with the estimate from the corresponding dummy variable (in this case, *MALE*). The dummy variable coefficient will often be smaller than the absolute difference because other factors included in the regression specification will have accounted for part of the difference.

3. INTERPRETATION OF A SEMI-LOGARITHMIC EARNINGS FUNCTION

For various reasons, some of which may become apparent in the exercises, the semi-logarithmic specification is considered to be the most satisfactory form of earnings function. A simple example is

$$LGEARN = \beta_1 + \beta_2 S + u$$

β_2 can be interpreted as the proportional increase in earnings attributable to an additional year of schooling. To see this, note that

$$EARNINGS = e^{\beta_1 + \beta_2 S + u}$$

Hence if S is increased by one unit (one year), $EARNINGS$ is multiplied by e^{β_2} , which is approximately $(1 + \beta_2)$ if β_2 is small. If β_2 is not small, the approximation breaks down and one has to calculate the proportional increase as $(e^{\beta_2} - 1)$.

Example

```
. reg LGEARN S;
```

Source	SS	df	MS			
Model	47.0744894	1	47.0744894	Number of obs =	540	
Residual	153.387029	538	.285106002	F(1, 538) =	165.11	
Total	200.461518	539	.371913763	Prob > F =	0.0000	
				R-squared =	0.2348	
				Adj R-squared =	0.2334	
				Root MSE =	.53395	

LGEARN	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	.1182303	.0092011	12.85	0.000	.1001558	.1363047
_cons	1.153066	.1293726	8.91	0.000	.8989281	1.407203

The table shows the output from a regression of $LGEARN$ on S using the DAT21 data set and Stata. The coefficient of S is 0.118, implying that each additional year of schooling increases earnings by a proportion 0.118, that is, 11.8 percent. Calculating the exact effect using $(e^{\beta_2} - 1)$, one obtains a slightly greater proportional estimate 0.125, or equivalently 12.5 percent.

Interpretation of dummy variable coefficients in a semi-logarithmic regression

(Do not bother with this until you have become familiar with the use of dummy variables in the ordinary linear specification)

Suppose that the dummy variable $MALE$ is included in a semi-logarithmic regression model with coefficient δ :

$$LGEARN = \beta_1 + \beta_2 S + \delta MALE + u$$

Rewriting the model as

$$EARNINGS = e^{\beta_1 + \beta_2 S + \delta MALE + u} = e^{\beta_1} e^{\beta_2 S} e^{\delta MALE} e^u$$

it can be seen that the new term multiplies *EARNINGS* by e^0 when *MALE* = 0 and e^δ when *MALE* = 1. e^0 is of course 1, so the new term has no effect for females. For males, the new term multiplies *EARNINGS* by e^δ which, if δ is small, approximates as $(1 + \delta)$. Hence, if δ is small, the model implies that males earn δ proportionally more than females. If δ is not small, the proportional difference is $(e^\delta - 1)$.

Example

```
. reg LGFEARN S MALE;
```

Source	SS	df	MS			
Model	61.2786779	2	30.639339	Number of obs =	540	
Residual	139.18284	537	.259185923	F(2, 537) =	118.21	
Total	200.461518	539	.371913763	Prob > F =	0.0000	
				R-squared =	0.3057	
				Adj R-squared =	0.3031	
				Root MSE =	.5091	

LGFEARN	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	.1140801	.0087908	12.98	0.000	.0968116	.1313486
MALE	.3250321	.043906	7.40	0.000	.2387836	.4112807
_cons	1.047976	.1241658	8.44	0.000	.8040654	1.291886

The table shows the effect of including *MALE* as an explanatory variable. Its coefficient is 0.229, implying, as a first approximation, that males earn a proportion 0.3250, that is 32.5 percent, more than females. In this case the coefficient is so large that one should definitely calculate $(e^\delta - 1)$, which is 0.384, implying that males earn 38.4 percent more than females.

4. INTERPRETATION OF A LOGARITHMIC REGRESSION

Nonlinear relationships of the type

$$Y = \beta_1 X^{\beta_2}$$

are very common in economic theory, for example in demand functions and Cobb–Douglas production functions. If Y is related to X in this way, its elasticity with respect to X is constant and equal to β_2 . Differentiating Y with respect to X , we have

$$\frac{dY}{dX} = \beta_2 \beta_1 X^{\beta_2-1} = \beta_2 \frac{Y}{X}$$

and hence the elasticity of Y with respect to X , given by the left side of the next equation, is equal to β_2 .

$$\frac{\frac{dY}{dX}}{\frac{Y}{X}} = \beta_2$$

The elasticity is the proportional change in Y per proportional change in X . One way of putting this into concrete terms is to say that, if X changed by 1 percent, Y would change by β_2 per cent.

To fit equations of this type, you linearize the equation by taking the logarithms of both sides:

$$\begin{aligned} \log Y &= \log \beta_1 X^{\beta_2} \\ &= \log \beta_1 + \log X^{\beta_2} \\ &= \log \beta_1 + \beta_2 \log X \end{aligned}$$

So far we have not specified whether we are taking logarithms to base e or to base 10. We shall always use e as the base, so we shall be using what are known as “natural” logarithms. This is standard in econometrics. Purists sometimes write \ln instead of \log , but this is unnecessary. Nobody uses logarithms to base 10 any more. They were tabulated in the dreaded log tables that were universally employed for multiplying or dividing large numbers until the early 1970s. With the invention of the pocket calculator, they have become redundant, along with the slide rule. They are not missed.

If we write $Y' = \log Y$, $X' = \log X$, and $\beta_1' = \log \beta_1$, the equation may be re-written

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

and you can fit it using ordinary regression analysis. All regression packages have built-in facilities for generating new variables like Y' and X' from existing ones. The coefficient of X' will be a direct estimate of the elasticity β_2 and the intercept will be an estimate of $\log \beta_1$. To obtain an estimate of β_1 , you calculate $e^{\beta_1'}$.

5. THE OAXACA DECOMPOSITION OF EARNINGS DIFFERENTIALS

The Oaxaca decomposition is used to quantify the difference in earnings of two categories of individual in a sample. A very common example, one of the exercises above, is the decomposition of the difference in male and female earnings. Another, which will be used as an example here, is the decomposition of the difference in the earnings of union members and non-members.

First of all the theory. Suppose we fit an earnings function for non-members. To keep the analysis simple, we will start with a model with only one explanatory variable, S :

$$LGEARN^{NM} = b_1^{NM} + b_2^{NM} S^{NM}$$

The intercept is determined by the following equation:

$$b_1^{NM} = \overline{LGEARN}^{NM} - b_2^{NM} \overline{S}^{NM}$$

Thus the mean value of the dependent variable is equal to the estimated intercept plus the mean value(s) of the explanatory variable(s) multiplied by its (their) regression coefficient(s):

$$\overline{LGEARN}^{NM} = b_1^{NM} + b_2^{NM} \overline{S}^{NM}$$

This is always true with ordinary least squares regression and of course is exactly what you would expect. When we fit the corresponding equation for union members, we will get the parallel relationship

$$\overline{LGEARN}^M = b_1^M + b_2^M \overline{S}^M$$

Subtracting the union member equation from the non-member equation, we have

$$\overline{LGEARN}^{NM} - \overline{LGEARN}^M = b_1^{NM} - b_1^M + b_2^{NM} \overline{S}^{NM} - b_2^M \overline{S}^M$$

Now if we subtract and add the term $b_1^{NM} \overline{S}^M$, we obtain

$$\overline{LGEARN}^{NM} - \overline{LGEARN}^M = b_1^{NM} - b_1^M + b_2^{NM} \overline{S}^{NM} - b_2^{NM} \overline{S}^M + b_2^{NM} \overline{S}^M - b_2^M \overline{S}^M$$

which can be re-written

$$\overline{LGEARN}^{NM} - \overline{LGEARN}^M = \left[b_2^{NM} (\overline{S}^{NM} - \overline{S}^M) \right] + \left[b_1^{NM} - b_1^M + (b_2^{NM} - b_2^M) \overline{S}^M \right]$$

We have thus decomposed the difference in the mean logarithm of earnings into the two components in square brackets. The first component gives that part of the difference attributable to the difference in quantities, in this case, mean years of schooling, for non-members and members, weighted by the estimated coefficient for non-members. The second term gives that part of the difference attributable to differences in the coefficients, weighted by the mean for members.

If we had had more explanatory variables, the first component would have been the sum of terms like $b_2^{NM} (\bar{S}^{NM} - \bar{S}^M)$, and the second component would have been the difference in the intercepts plus the sum of terms like $(b_2^{NM} - b_2^M) \bar{S}^M$. The first component is sometimes referred to as the part of the difference attributable to differences in *endowments*, and the second as the part of the difference attributable to the differences in *prices*.

You may have spotted one arbitrary step in the analysis. We could equally well have subtracted and added the term $b_2^M \bar{S}^{NM}$. If we had done that, we would have ended up with a decomposition in which the differences in endowments were weighted by the prices for union members, and the differences in price—s were weighted by the endowments of non-members. Usually the result is fairly similar.

Example

Oaxaca decomposition, earnings of non-members and members of unions								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean quantities (endowments)			Regression coefficients (prices)				
	members	non-members	difference	members	non-members	difference	(3)x(4)	(2)x(6)
<i>LGEARN</i>	2.8680	2.7702	0.0978					
<i>S</i>	14.0673	13.7821	0.2852	0.0660	0.1240	-0.0580	0.0188	-0.7994
<i>EXP</i>	17.1091	17.0017	0.1074	0.0206	0.0193	0.0013	0.0022	0.0221
<i>ASVABC</i>	49.9767	51.1062	-1.1295	0.004250	0.004771	-0.0005	-0.0048	-0.0266
<i>MALE</i>	0.6250	0.4702	0.1548	0.3900	0.2559	0.1341	0.0604	0.0631
<i>ETHBLACK</i>	0.0962	0.1147	-0.0185	-0.1431	-0.1352	-0.0079	0.0026	-0.0009
<i>ETHHISP</i>	0.0577	0.0619	-0.0042	0.0921	0.0714	0.0207	-0.0004	0.0013
Constant	1.0000	1.0000	0.0000	1.1393	0.3800	0.7593	0.0000	0.7593
Total							0.0789	0.0188

The first two columns give the means of *LGEARN* and the explanatory variables used in the analysis. Column (3) is the difference of (1) and (2). Columns (4) and (5) give the coefficients obtained in separate regressions for non-members and members, respectively (*COLLBARG* equal to 0 and 1, respectively; the union sample included those respondents whose earnings were covered by collective bargaining, even if they were not actually union members themselves). Column (6) is the difference between (4) and (5). Column (7) is (3)x(4) and gives the component attributable to differences in endowments. Column (8) is (2)x(6) and gives the component attributable to differences in prices.

The mean log of earnings is 2.8680 for union members and 2.7702 for non-members, implying that the hourly earnings of union members are approximately 10 percent higher than those of non-members (the difference is actually in the geometric means, not arithmetic means, but it is not worth worrying about the distinction). Looking at the totals in columns (7) and (8), we see that differences in endowments rather than differences in prices are responsible for the gap. The most important difference is in the proportion of males. This is much higher for the union subsample, and since males are paid more than females, especially in the union subsample, this accounts for most of the difference in average earnings in the two subsamples. It should be noted that the sample is rather small for such a specialized application. There are only 540 observations in all, and only 104 of them relate to union members.

6. FURTHER DETAILS OF THE VARIABLES

The meanings of most of the variables are obvious from the definitions given in Section 1. This section provides information on those that may need some further explanation.

The FAITH variables

The dummy variables are defined according to the response to the question "In what religion were you raised?", asked during the 1979 interview.

The ASVAB variables

The Armed Services Vocational Aptitude Battery is a series of ten tests taken by potential recruits to the military. Nearly all the NLSY79 respondents took the test as part of a project sponsored by the Department of Defense to obtain updated information on the distribution of scores that could be expected and hence to allow the raw score on a test item (number of correct responses) to be mapped to a distribution with mean 50 and standard deviation 10.

Eight of the tests are power tests, that is, tests where the questions start by being very easy and then progressively become more difficult, with enough time allowed for time not to be a factor. Three of these are cognitive tests (relating to basic intelligence) and five are knowledge tests. The variables *ASVAB2* – *ASVAB4* are the scores on the cognitive tests. *ASVAB2* is arithmetic reasoning, *ASVAB3* is word knowledge, and *ASVAB4* is paragraph comprehension. Even the most difficult test items are fairly easy, the general purpose of the *ASVAB* being to discriminate among those whose education is limited to high school, the most important source of recruitment to the armed forces.

The *ASVABC* score is a composite of *ASVAB2* – *ASVAB4* constructed specifically for the present data sets. It combines *ASVAB2* with weight 0.5, with *ASVAB3* and *ASVAB4*, each with weight 0.25. An adjustment has been made to preserve the standard deviation at approximately 10 without changing the mean of approximately 50. (A similar composite, known as the Armed Forces Qualification Test score, is constructed by the military but, because it is scaled in the form of percentiles, it cannot be compared directly with the scores from which it is constructed.)

The other two tests are speed tests, consisting of very easy items with no difficulty in gradient but with so little time allowed that only a small minority of respondents can complete. One is *ASVAB5*, numerical operations, where a typical test item is multiplying 3 by 3. The other is *ASVAB6*, coding speed, in which four-digit numbers are translated to words using a simple key, the key being changed periodically. Again, the difficulty of each item is very low; the score depends on concentration and short-term memory. *ASVAB5* and *ASVAB6* are not specified in any of the exercises but you should try experimenting with them.

The LIBRARY variable

This dummy variable is defined according to the response to the 1979 interview question "When you were about 14 years old, did you or anyone else living with you have a library card?". Try using it in the educational attainment function.

The COLBARG variable

This is defined to be 1 if the respondent said that her or his earnings in 2002 were determined by a collective bargaining agreement. Try using it in the earnings function.