

Data Analysis Project 1:
ESTIMATING THE EARNINGS EQUATION

Due: Thursday, November 1, 2001

General instructions: You may do this project in a group of up to three (3) students. Since you will have to do the final project on your own, I strongly encourage each group member to participate and learn how to use the software. Commands from Excel menus are indicated as follows: **File > Open** means you should click on File and then select Open from the menu. Variable names are in all upper-case letters. Something you actually have to hand in is indicated in *italics*.

In this project you will examine data from the March 2000 Current Population Survey (CPS), contained in the following file:

earn2000.xls Random sample of U.S. adults (18+) in the labor force who worked full-time, year-round in 1999 and were not self-employed.

I have attached a brief description of the CPS and a listing of the variables in our data set and their definitions.

What you hand in should be typed, and all tables and figures should be numbered or labeled for easy identification when you discuss them in the text of your answers. Although you will be asked to hand in part of this project as separate homework assignments, please make sure to hand in all the results for all parts when you hand it in on November 1.

Getting started

How to download the data from the web

1. Go to the course web page:

<http://lsb.scu.edu/~wsundstrom/econ150/>

Click on Data sets

2. To download, right-click on the data set you need (earn2000), and you should then be given the option of saving the file somewhere. Save it to your hard disk or a floppy.

Make sure Excel is set up to do data analysis

In Excel, click on **Tools**. Near the bottom of the menu you should see an option for **Data Analysis**. If it is there, you are all set. If it is not there, you need to click on **Tools > Add-Ins**. You will then be able to add the Analysis ToolPak, which has what you need.

Part 1: Descriptive statistics

1. Open the data set (earn2000.xls) in Excel. Variable names are across the top row, and the variables are in columns. If you forget what a variable is, see the definitions attached to this handout (also on the web site).
2. Scroll to the bottom of the data. *How many observations (N) are there in the sample?*
3. Take a look at observation (person) #15. *Describe this person. Report her or his:*
 - C gender
 - C age
 - C race
 - C region of residence
 - C educational attainment
 - C earnings from wages and salary
 - C occupation
 - C marital status
4. Now obtain some basic statistics for the following three variables: **age**, **school**, and **wagesal**. Throughout this assignment, we will use **wagesal** (annual wage and salary income) as our measure of earnings. For each of these three variables, do the following:
 - C Generate means, etc. using **Tools > Data Analysis > Descriptive Statistics**. Note that you must specify the Input Range (cells that contain all the data). Include the variable name in the input range and check Labels in First Row.

C Generate percentiles using **Tools > Data Analysis > Rank and Percentile**.

Note: When given the option for where to put the output, it is convenient to place it on a new worksheet in the same workbook.

C *Print out your results using the **Print** button, and hand in the printout with your assignment.*

C *For each of the three variables, compare the mean and median. If they are different, what is the significance of one being bigger than the other?*

C *For each of the three variables, find the 10th, 25th, 50th, 75th, and 90th percentiles. Examine the extremes of the distribution (top and bottom few values). Are there any obvious “outliers”?*

5. Use the same descriptive statistics procedure to answer the following questions. (Hint: the proportion of Asians in the sample is given by the mean of the **asian** variable.) You do not need to hand in the actual printouts for these questions, just the number for each.

C *What percentage of individuals are white? Black? Asian? Latino?*

C *What percentage of individuals are female?*

C *What percentage of individuals are married with spouse present?*

C *What percentage of individuals graduated from high school but had no further formal education?*

C *What percentage of individuals have more than a high-school education?*

6. Now examine the subsample of women. The best way to do this in Excel is sort all your data by **female**, using either **Data > Sort** or **Data > Filter**, then copy the observations with **female** = 1 into a new sheet.

Using this restricted (female) sample, answer all parts of questions 4 and 5 over again for females only.

7. *Repeat all of #6 for the subsample of males only.*

8. *Using your separate results for men and women (#7 and #8), compare the men and women in the sample. In terms of the variables you have examined, in what ways are men and women similar? In what ways are they different?*

Part 2: Regression basics

This part follows fairly closely what I demonstrated in lecture.

1. Using the earnings data (full data set), create an XY scatter plot of earnings (vertical) against years of school (horizontal). To do this, it helps to have the X variable (**school**) in one column and the Y variable (**wagesal**) in the next column immediately to the right. Then use the chart wizard to create an XY scatter.

Print out the scatter plot and answer the following questions about the scatter plot:

- C Does the relationship between earnings and schooling appear to be positive or negative?*
- C Are there any obvious outliers?*
- C Sketch in a line that you think fits the data and estimate what its equation is.*

2. Estimate a simple regression with earnings (**wagesal**) hours as the dependent (Y) variable and schooling (**school**) as the independent or X variable. To do this, use **Tools > Data Analysis > Regression** and proceed as we did in class. Make sure to use the variable labels. Check off the boxes for residuals and residual plots. Put the output on a new worksheet.

Print out the resulting table of results, and answer the following questions:

- C Write out the formula for the equation you have estimated.*
- C Interpret the R-squared.*
- C Interpret the ANOVA table and the F-statistic.*
- C Calculate a 95% confidence interval for the slope coefficient. Is the slope significantly different from zero? Could you reject the hypothesis that an additional year of schooling increases wage and salary income by \$1000? By \$2000? Explain.*

3. Now examine the residual plot from this regression. *Does the residual have any systematic relationship to the variable **school**? Are there outliers? Would you say the residual appears random?*
4. Extra credit: *Repeat question #2, but this time try to remove the obvious outlier(s) from the data. Do your regression results change much?*

Part 3: Multiple regression

We now predict earnings using **school** and **exper**, where **exper** = **age** - **school** - 6. This variable is often called “years of potential work experience.”

1. Run a regression in which the dependent variable is **wagesal** and the independent variables are **school** and **exper**. (Note that all the X-variables (regressors) must be in adjacent columns.) *Print out the results.*
 - C Check the general validity of the regression by examining and discussing the R-squared, F-statistic, and residuals.*
 - C Interpret the coefficients on **school** and **exper**. Are these coefficients statistically significant?*
2. Now run a similar regression, but this time use as the dependent variable the *natural log* of wage and salary earnings (do NOT take the logs of **school** or **exper**.) To do this you need to generate a new column of numbers with the variable name **lwagesal** (the Excel function LN(..) produces the natural log). *Print out the results.*
 - C Again, check the general validity of the regression by examining and discussing the R-squared, F-statistic, and residuals. Do the residuals appear to be more randomly*

distributed? Do you prefer this regression or the regression from #1? Explain.

- C *Interpret the coefficients on **school** and **exper**. Are these coefficients statistically significant?*

*For the remainder of this Data Analysis Project, continue to use the log of earnings (**lwagesal**) as your dependent variable.*

3. To your regression from #2, now add as a regressor a dummy variable for **female** (include **school** and **exper** as regressors too). *Print out the results. Controlling for schooling and experience, does the evidence suggest that women earn more than, less than, or about the same as men? Explain (consider both economic and statistical significance in your answer).*
4. To your regression from #3, add two interaction terms: the interaction between **female** and **school**, and the interaction between **female** and **exper** (include all the regressors you used in #3). This requires creating two new columns of numbers. For instance, the interaction of **female** and **school** could be **femsch = female*school**. *Print out and carefully interpret the results:*
 - C *Draw a diagram representing the relationship between schooling and log earnings implied by your regression, with separate lines for men and women (assume that **exper** = 10 years in each case). Is there evidence that the returns to schooling are significantly different between men and women?*
 - C *Draw a diagram representing the relationship between work experience and log earnings implied by your regression, with separate lines for men and women (assume that **school** = 14 years in each case). Is there evidence that the returns to experience are significantly different between men and women?*
 - C *Use an F-test to test the hypothesis that the coefficients on **female** and the two **female** interactions are all equal to zero (that is, test the general hypothesis that there are no significant gender differences in the earnings equation estimated here). See the handout on F-tests for details.*
5. Be prepared to discuss your findings in class: No formal presentations, but think about how your findings for questions #3-4 might be explained.

Current Population Survey (CPS): Overview

(From CPS web site: <http://www.bls.census.gov/cps/overmain.htm>)

The Current Population Survey (CPS) is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years.

The CPS is the primary source of information on the labor force characteristics of the U.S. population. The sample is scientifically selected to represent the civilian noninstitutional population. Respondents are interviewed to obtain information about the employment status of each member of the household 15 years of age and older. However, published data focus on those ages 16 and over. The sample provides estimates for the nation as a whole and serves as part of model-based estimates for individual states and other geographic areas.

Estimates obtained from the CPS include employment, unemployment, earnings, hours of work, and other indicators. They are available by a variety of demographic characteristics including age, sex, race, marital status, and educational attainment. They are also available by occupation, industry, and class of worker. Supplemental questions to produce estimates on a variety of topics including school enrollment, income, previous work experience, health, employee benefits, and work schedules are also often added to the regular CPS questionnaire.

CPS data are used by government policymakers and legislators as important indicators of our nation's economic situation and for planning and evaluating many government programs. They are also used by the press, students, academics, and the general public.

Variable definitions for CPS data: earn2000.xls

Note: “=1” indicates a “dummy variable” equal to 1 if the condition holds, 0 otherwise

Variable name	Definition
age	age in years
asian	=1 if Asian or Pacific Islander
assocdeg	=1 if Associate degree
attendhs	=1 if attended but didn't complete high school
bachdeg	=1 if Bachelor's degree
black	=1 if race is black
citstat	citizenship status: 1= Native-born US, 2= Native-born US outlying areas, 3= Native-born US abroad US parent, 4= Foreign born-naturalization, 5= Not a US citizen
divsep	=1 if divorced or separated
doctdeg	=1 if doctorate degree, including MD, DDS, etc.
exper	potential years of work experience = age - school - 6
faminc	total family income in 1997 (\$)
female	=1 if female
forborn	=1 if foreign-born
govtemp	=1 if employed by government
herelt5	=1 if moved to USA permanently in last 5 years
hsgrad	=1 if high-school grad (no further school)
hsplus	=1 if more than high school education
industry	industry (see attached codes)
kidun18	number of children in family under 18
kidun6	number of children in family under 6

latino	=1 if Latino (Latin American ethnicity)
marital	1=married, spouse present, 3=married, spouse absent, 4=widowed, 5=divorced, 6=separated, 7=never married
married	=1 if married (not including separated)
mastdeg	=1 if Master's degree
midwest	=1 if living in Midwest region
nevermar	=1 if never has been married
nogradhs	=1 if did not finish high school
nonwhite	=1 if race is nonwhite
northeast	=1 if living in Northeast region
occup	occupation (see attached codes)
private	=1 if private (nongovernment) employer
profdeg	=1 if professional degree
schlt9	=1 if less than 9 years of school
school	total years of school (derived from ranges of grades completed)
somecoll	=1 if some college, but no degree
south	=1 if living in South region
spspres	=1 if married with spouse present
wagesal	total wage and salary income earned in 1997 (\$)
west	=1 if living in West region
widowed	=1 if widowed

Industry codes

- 01 . Agriculture
- 02 . Mining
- 03 . Construction manufacturing
- 04 . Manufacturing- durable goods
- 05 . Manufacturing- nondurable goods
 . transportation, communications,
 . and other public utilities
- 06 . Transportation
- 07 . Communications
- 08 . Utilities and sanitary services
 . wholesale and retail trade
- 09 . Wholesale trade
- 10 . Retail trade
- 11 . Finance, insurance and real
 . estate services
- 12 . Private household miscellaneous
 . services
- 13 . Business and repair
- 14 . Personal services, except
 . private household
- 15 . Entertainment professional and
 . related services
- 16 . Hospital
- 17 . Medical, except hospital
- 18 . Educational
- 19 . Social services
- 20 . Other professional
- 21 . Forestry and fisheries
- 22 . Public administration
- 23 . Armed Forces

Occupation codes

Managerial & professional

01 .Executive, admin. & managerial

02 .Professional specialty

Technical, sales & admin. support

03 .Technicians & related support

04 .Sales

05 .Administrative support, incl.
.clerical

Service

06 .Private household

07 .Protective service

08 .Other service

09 .Precision production, craft &
.repair

Operators, fabricators & laborers

10 .Machine operators, assemblers &
.inspectors

11 .Transportation & material
.moving

12 .Handlers, equip. cleaners, etc.

13 .Farming, forestry & fishing

14 .Armed Forces