Problem Set 2

Francisco Brady

6 Feb 2024

### Part 1: Data Analysis

The dataset Earnings\_and\_Height.dta contains data from the 1994 National Health Interview Survey. This is a subset of the data used in the Anne Case and Christina Paxson paper “Stature and Status: Height, Ability, and Labor Market Outcomes”, Journal of Political Economy, 2008, 116(3): 499-532. The dataset contains information on 17,870 workers. A more detailed codebook is provided at the end of the problem set.

#### 1. Explore and describe the data.

───────────────────────────────────────────────────────────────────  
 (1)   
   
 min p50 mean p75 max  
───────────────────────────────────────────────────────────────────  
earnings 4726.391 38925.34 46875.32 84054.75 84054.75  
educ 0 13 13.53733 16 19  
───────────────────────────────────────────────────────────────────  
N 17870   
───────────────────────────────────────────────────────────────────

##### a. Describe the variation in earnings. What is the mean, median, min, max, 25th percentile, 75th percentile? (Note: In addition to pasting Stata or R output, please interpret what you find using 1-2 sentences).

The mean earnings is $46,875.32, with a median value of $38,925.34. The lowest earnings amount is $4736.39, while both the 75ht percentile and the max are $84,054.75. The mean is greater than the median, which can indicate a positive skew.

##### b. Describe the variation in years of schooling. What is the mean, median, min, max, 25th percentile, 75th percentile? (Note: In addition to pasting Stata or R output, please interpret what you find using 1-2 sentences).

The mean years of education is 13.53, with a median value of 13. The years of education is 0. The 75th percentile is 16 years, and the max years of education is 19. The data has a positive skew.

##### c. What is the correlation between earnings and years of schooling? Interpret the correlation using a few sentences (e.g. is it positive or negative? Weak or strongly correlated?)

(obs=17,870)  
  
 │ earnings educ  
─────────────┼──────────────────  
 earnings │ 1.0000  
 educ │ 0.3880 1.0000

Education and earnings are weakly positively correlated. This can be seen by visualizing the data, as years of education increase, earnings also increase, although the correlation is not as strong below the mean of earnings.

#### 2. Regression Analysis I

##### a. What is the unconditional relationship between earnings and years of education? As a first step, use a bivariate regression to regress earnings on years of education and paste the output below.

. regress earnings educ, robust  
  
Linear regression Number of obs = 17,870  
 F(1, 17868) = 3414.09  
 Prob > F = 0.0000  
 R-squared = 0.1505  
 Root MSE = 24815  
  
─────────────┬────────────────────────────────────────────────────────────────  
 │ Robust  
 earnings │ Coefficient std. err. t P>|t| [95% conf. interval]  
─────────────┼────────────────────────────────────────────────────────────────  
 educ │ 3953.761 67.66637 58.43 0.000 3821.129 4086.394  
 \_cons │ -6648.031 918.8802 -7.23 0.000 -8449.125 -4846.937  
─────────────┴────────────────────────────────────────────────────────────────

The relationship between earnings and education is positive and significant. On average, one year increase in schooling is associated with an additional $3,953.76 dollars in earnings. The t statistic is 56.27 and the estimates are significant at greater than 95% confidence levels.

##### b. Write out the equation for the sample regression line that corresponds to this regression.

##### c. Find and label the following on your Stata or R output and interpret these statistics with words.

###### i. :

The coefficient on Education is 3953.761, which is the average increase in earnings associated with an additional year of education.

###### ii. Standard error of :

The standard error is 67.666 on Education, which represents the average distance of the observed values from the regression estimate.

###### iii. t-test for null hypothesis that :

The t statistic for the test that the coefficient on education is 0 is 58.43. The absolute value of the t statistic is larger than 1.96, which means that we can reject the null hypothesis that the coefficient is 0 with a 95% level of confidence.

###### iv. :

The intercept/constant is -6648.013, which is the value estimated by the equation for someone with 0 years of education.

###### v. :

The for the model is .1505, which means that around 15% of the variance in earnings can be predicted from years of education. This is the overall strength of the association.

##### d. Based on this regression, what is the predicted earnings for someone with 12 years of schooling? What about 16 years of schooling? Show your work.

#### 3. Regression Analysis II

##### a. Regress earnings on the variable indicating the sex of the worker (check the codebook below to verify the values of this variable).

. regress earnings sex, robust  
  
Linear regression Number of obs = 17,870  
 F(1, 17868) = 49.08  
 Prob > F = 0.0000  
 R-squared = 0.0027  
 Root MSE = 26887  
  
─────────────┬────────────────────────────────────────────────────────────────  
 │ Robust  
 earnings │ Coefficient std. err. t P>|t| [95% conf. interval]  
─────────────┼────────────────────────────────────────────────────────────────  
 sex │ 2838.752 405.2162 7.01 0.000 2044.489 3633.015  
 \_cons │ 45621 268.7055 169.78 0.000 45094.31 46147.69  
─────────────┴────────────────────────────────────────────────────────────────

##### b. Write out the equation for the sample regression line that corresponds to this regression.

##### c. What is the “omitted” category in this regression?

In this regression, the sex variable is coded as a binary variable (0/1), with 0 indicating female and 1 indicating male. The omitted group in the regression is male identification.

##### d. Find and label the following on your Stata or R output and interpret these statistics with words.

###### i. :

The coefficient on is 2838.752, which is the average contribution to earnings of being male.

###### ii. Standard error of :

The standard error on is 405.2162, which is the average distance of the observed values from the values predicted by the regression equation.

###### iii. t-test for null hypothesis that :

The t statistic for the test that the coefficient on sex is 0 is 7.01. The absolute value of the t statistic is larger than 1.96, which means that we can reject the null hypothesis that the coefficient on sex is 0 with a 95% level of confidence.

###### iv. :

The constant is 45,621, which is the value produced by the estimate for someone who identifies as a woman.

###### v. :

The for the model is .0027, which means that around .027% of the variance in earnings can be predicted by the sex of the individual. This is the overall strength of the association.

##### d. Based on this regression, what is the predicted earnings for a worker who identifies as male? As female? Show your work.

#### 4. Model Fit

Based on the two models above, which characteristic – education or sex – explains more of the variation in earnings in this data?

The model using education explains almost 15% more of the variance in earnings. The is higher than the model using sex identity. In addition, the standard errors for the regression using education are much smaller, indicating that the observed values are closer to the values predicted by the model.

#### 5. Introduction to multiple regression

In assessing the relationship between earnings and education, you might want to control for several possible confounders. Create a table (in Excel or Word or some other spreadsheet tool) that looks like Table 7.1 from S&W (though you can simply report the R-squared and not the adjusted R-squared or SER). Each column should correspond to a separate regression and each row should correspond to a different variable, with education in the first row. Each cell of the table should report the coefficient and the standard error for the given variable and specification. The table should contain the follow specifications:

(Hint: to control for years of experience, generate a new variable, exp, where exp=age-educ-6; this is a fairly standard way of measuring potential years of experience in labor economics—we take a person’s age, subtract the number of years they were in school, and subtract 6 to account for school starting age, the balance is the expected number of years of work experience).

Column 1: Regress earnings on education (same model estimated above)  
Column 2: Regress earnings on education and years of experience  
Column 3: Regress earnings on education and years of experience and sex  
Column 4: Regress earnings on education and years of experience and sex and height and whether the individual is married

────────────────────────────────────────────────────────────────────────────────────  
 Model 1 Model 2 Model 3 Model 4   
   
────────────────────────────────────────────────────────────────────────────────────  
Education of Indiv~l 3953.761\*\*\* 4347.498\*\*\* 4345.374\*\*\* 4292.414\*\*\*  
 (67.67) (69.36) (69.13) (67.15)   
exp 326.536\*\*\* 327.744\*\*\* 292.110\*\*\*  
 (18.27) (18.21) (16.86)   
Sex 2751.913\*\*\* -355.257   
 (368.96) (479.20)   
height 438.043\*\*\*  
 (60.71)   
married 20920.802\*\*\*  
 (337.50)   
\_cons -6648.031\*\*\* -18960.229\*\*\* -20173.272\*\*\* -60282.661\*\*\*  
 (918.88) (1104.10) (1113.83) (3931.18)   
────────────────────────────────────────────────────────────────────────────────────  
Adj. R-Squared 0.150 0.165 0.168 0.306   
R-Squared 0.151 0.165 0.168 0.306   
Observations 17870.000 17870.000 17870.000 17870.000   
────────────────────────────────────────────────────────────────────────────────────  
\* p<0.10, \*\* p<0.05, \*\*\* p<0.01

##### a. Discuss the change in magnitude and statistical significance of the coefficient on educ from column 1 to 2.

In regression 1, the coefficient on education is 3953.761. In regression 2, the estimate for education rose to 4347.498. In both regressions, the coefficient of education is significant at greater than the 95% level. After adding exp, the of the model increased.

##### b. Interpret the coefficient on exp in column 2. Please discuss the magnitude and statistical significance.

In regression 2, the coefficient on experience is 326.536. The coefficient is significant at greater than the 95% level. This means that on average an additional year of experience, holding education equal, is associated with an increase of around $326 dollars in earnings.

##### c. From the results in column 2, would you argue that education or experience matters more for earnings? Why?

Based on the magnitude of the coefficients in column 2, an additional year of education is associated with a larger increase in earnings (holding experience equal), than an additional year of experience. This supports the argument that education matters more for experience in explaining earnings for this sample.

##### d. Discuss the difference between the R-squared across columns 1 through 4.

In the first 3 regressions, the is between .15 and .168, meaning that those regressions were able to explain around 15-16.8% of the variance in the sample. As more variables are added to the regressions, the increases, indicating that the additional variables increase the ability of the overall model to explain the variance between the observed values and the predicted values. It is important to note that the increase in can happen because of how it is calculated. The additional variables not only increased the , but decreased the standard errors on the coefficients for most of the variables. This indicates that adding those variables improved the explanatory power of the model. In the last regression, the doubles to .304, which supports the inclusion of the height and married variables. Interestingly, the coefficient on the sex variable actually loses significance when holding height and marital status equal.

##### e. In about 2 sentences, describe the relationship between earnings and education and whether including these controls changes this relationship much.

The relationship between earnings and education is positive and statistically significant, with an additional year of education increasing earnings around $4300, on average. The addition of controls for experience, sex, height, and marital status increase the precision of the education estimate and are all statistically significant to the model.