1. (Lecture 2) (12 points) Use the data in CHARITY [obtained from Franses and Paap (2001)] to answer the following questions:
2. (2 points) What is the average gift in the sample of 4,268 people (in Dutch guilders)? What percentage of people gave no gift?
3. (2 points) What is the average mailings per year? What are the minimum and maximum values?
4. (2 points) Estimate the model

by OLS and report the results in the usual way, including the sample size and R-squared.

1. (4 points) Interpret the slope coefficient. If each mailing costs one guilder, is the charity expected to make a net gain on each mailing? Does this mean the charity makes a net gain on every mailing? Explain.
2. (2 points) What is the smallest predicted charitable contribution in the sample? Using this simple regression analysis, can you ever predict zero for gift?

Answer:

1. The average gift is about 7.44 Dutch guilders. Out of 4,268 respondents, 2,561 did not give a gift, or about 60 percent.
2. The average mailings per year is about 2.05. The minimum value is .25 (which presumably means that someone has been on the mailing list for at least four years), and the maximum value is 3.5.
3. The estimated equation is
4. The slope coefficient from part (iii) means that each mailing per year is associated with – perhaps even “causes” – an estimated 2.65 additional guilders, on average. Therefore, if each mailing costs one guilder, the expected profit from each mailing is estimated to be 1.65 guilders. This is only the average, however. Some mailings generate no contributions, or a contribution less than the mailing cost; other mailings generated much more than the mailing cost.
5. Because the smallest *mailsyear* in the sample is .25, the smallest predicted value of *gifts* is . Even if we look at the overall population, where some people have received no mailings, the smallest predicted value is about two. So, with this estimated equation, we never predict zero charitable gifts.
6. (Lecture 3) (14 points) The file CEOSAL2 contains data on 177 chief executive officers and can be used to examine the effects of firm performance on CEO salary.
7. (3 points) Estimate a model relating annual salary to firm sales and market value. Make the model of the constant elasticity variety for both independent variables. Report the results in the usual way.
8. (4 points) Add *profits* to the model from part (i), re-estimate the model and report the results in the usual way. Why can this variable not be included in logarithmic form? Would you say that these firm performance variables explain most of the variation in CEO salaries?
9. (3 points) Add the variable *ceoten* to the model in part (ii), re-estimate the model and report the results in the usual way. What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?
10. (4 points) Find the sample correlation coefficient between the variables log(*mktval*) and *profits*. Are these variables highly correlated? What does this say about the OLS estimators? [Hint: You can use the stata command correlate.]

Answer:

1. The constant elasticity equation is
2. We cannot include *profits* in logarithmic form because *profits* are negative for nine of the companies in the sample. When we add it in levels form, we get

The coefficient on *profits* is very small. Here, *profits* are measured in millions, so if *profits* increase by $1 billion, which means  – a huge change – predicted salary increases by about only 3.6%. Together, these variables (and we could drop *profits* without losing anything) explain almost 30% of the sample variation in log(*salary*). This is certainly not “most” of the variation.

1. Adding *ceoten* to the equation gives

This means that one more year as *CEO* increases predicted salary by about 1.2%.

1. The sample correlation between log(*mktval*) and *profits* is about .78, which is fairly high. As we know, this causes no bias in the OLS estimators, although it can cause their variances to be large. Given the fairly substantial correlation between market value and firm profits, it is not too surprising that the latter adds nothing to explaining CEO salaries. Also, *profits* is a short term measure of how the firm is doing, while *mktval* is based on past, current, and expected future profitability.
2. (Lecture 4) (17 points) Refer to the example used in Lecture 4 to compare the returns to education at junior colleges and four-year colleges. The model after rearrangement is

where *totcoll* is total years of college. Use the data set TWOYEAR, which comes from Kane and Rouse (1995).

1. (5 points) Run the regression above and report the OLS estimates in the usual form, including the standard errors, sample size and R-squared. How do you interpret ? Is it statistically significant?
2. (2 points) The variable *phsrank* is the person’s high school percentile. (A higher number is better. For example, 90 means you are ranked better than 90 percent of your graduating class.) Find the smallest, largest, and average *phsrank* in the sample.
3. (4 points) Add *phsrank* to the model and report the OLS estimates in the usual form. Is *phsrank* statistically significant? How much is 10 percentage points of high school rank worth in terms of wage?
4. (3 points) Compare regression results in (i) and (iii), does adding *phsrank* to the model substantively change the conclusions on the returns to two- and four-year colleges? Explain.
5. (3 points) The data set contains a variable called *id*. Explain why if you add *id* to the model you expect it to be statistically insignificant. What is the two-sided p-value?

Answer:

1. The estimated OLS equation is as follows:

represents the difference of the return to a year at a junior college and the return to a year at a university. Its estimated value indicates that the return to a year at a junior college is about one percentage point less than a year at a university. The absolute value of *t* statistic is about 1.47, which is insignificant at even 10% level.

1. The minimum value is 0, the maximum is 99, and the average is about 56.16.
2. When *phsrank* is added to the model, we get the following:

So *phsrank* has a *t* statistic equal to only 1.27; it is not statistically significant. If we increase *phsrank* by 10, log(*wage*) is predicted to increase by (.0003)10 = .003. This implies a .3% increase in *wage*, which seems a small increase given a 10 percentage points increase in *phsrank*.

1. Adding *phsrank* makes the *t* statistic on *jc* even smaller in absolute value, about 1.34, but the coefficient magnitude is similar (from -.0102 to -.0093). Therefore, the basic point remains unchanged; the return to a junior college is estimated to be somewhat smaller compared to the return to a university, but the difference is not significant at standard significant levels.
2. The variable *id* is just a worker identification number, which should be randomly assigned (at least roughly). Therefore, *id* should not be correlated with any variable in the regression equation. It should be insignificant when added to the model. In fact, its *t* statistic is very low, about .54, and the two-sided p-value is 0.587.
3. (Lecture 4) (9 points) Use the data set GPA1 to answer this question.
4. (3 points) Run the regression *colGPA* on *PC*, *hsGPA*, and *ACT* and obtain a 95% confidence interval for βPC. Is the estimated coefficient statistically significant at the 5% level against a two-sided alternative?
5. (3 points) Discuss the statistical significance of the estimates and in part (i). Is *hsGPA* or *ACT* the more important predictor of *colGPA*? Explain.
6. (3 points) Add the two indicators *fathcoll* and *mothcoll* to the regression in part (i). Is either individually significant? Are they jointly statistically significant at the 5% level?

Answer:

1. The 95% confidence interval around is (.044, .271). The point estimate of .157 is statistically different from zero at the 5% level against a two-sided alternative (*p*-value = .007).
2. The coefficient on *hsGPA* is statistically significant even at the 1% level (*p*-value = .000). While the coefficient on *ACT* is not statistically different from 0 even at the 10% level (*p*-value = .413). Besides, the point estimate of *hsGPA* is 0.447 while that of *ACT* is only .009. This suggests that the marginal effect of a 1 unit change in *hsGPA* on *colGPA* is larger in magnitude than the marginal effect of a 1 unit change in ACT score.
3. Neither *fathcoll* nor *mothcoll* are individually significant at the 5% level (*t* statistics of 0.68 and -0.06 respectively). We can test joint significance using an F-test. The *F* statistic is .24 following an F2,135 distribution under H0. The *p*-value is .7834, so *fathcoll* and *mothcoll* are not jointly statistically significant at the 5% level.
4. (Lecture 5) (10 points) Use the data in WAGE1 for this exercise.
5. (4 points) Estimate the equation

and report the OLS estimates in the usual form. Save the residuals and plot a histogram.

[Hint: 1) You can obtain the residuals of each prediction by using the **residuals** command and storing these values in a variable named whatever you’d like, e.g., predict resid\_wage, residuals. 2) You can use the **histogram** command to plot a histogram, e.g., histogram resid\_wage.]

1. (4 points) Repeat part (i), but with log(*wage*) as the dependent variable.
2. (2 points) Would you say that Assumption MLR.6 is closer to being satisfied for the level-level model or the log-level model? Explain.

Answer:

1. The estimated equation is

Seen below is a histogram of the 526 residuals, . The histogram uses 22 bins.



1. With log(*wage*) as the dependent variable, the estimated equation is

Again, the histogram for the residuals from this equation is given below. The histogram uses 22 bins.



1. The residuals from the log(*wage*) regression appear to be more normally distributed. The histogram for the *wage* residuals is notably skewed to the left. In the *wage* regression, there are some very large residuals (roughly equal to 15) that lie almost five estimated standard deviations () from the mean of the residuals, which is identically zero, of course. Residuals far from zero do not appear to be nearly as much of a problem in the log(*wage*) regression.
2. (Lecture 5) (13 points) The model we used in class to explain the standardized outcome on a final exam (*stndfnl*) in terms of percentage of classes attended, prior college grade point average, and ACT score is

.

1. (2 points) Argue that
2. (3 points) Use the equation above to estimate the partial effect of *priGPA* on *stndfnl* when *priGPA* is at its mean value 2.59, and *atndrte* is also at it mean value 82. Interpret your estimate. [Hint: The estimated OLS equation can be found in Lecture 5.]
3. (4 points) Show that the equation can be re-written as

where .

How do you interpret ?

1. (4 points) Following (iii), suppose that, in place of , you put . Now how do you interpret the coefficients on *atndrte* and *priGPA*?

Answer:

1. The result can be easily shown when you take the derivative of *stndfnl* with respect to *priGPA*.
2. The partial effect . It means that at the mean value of percentage of classes attended, increasing prior college grade point average by one point is expected to increase *stndfnl* by .362 standard deviations from the mean final exam score.
3. (Mathematical derivation omitted) represents the partial effect of *priGPA* on *stndfnl*, at the mean value of *atndrte*.
4. Now the coefficient on *atndrte* represents the partial effect of *atndrte* on *stndfnl*, at the mean value of *priGPA*; and the coefficient on *priGPA* represents the partial effect of *priGPA* on *stndfnl*, at the mean value of *atndrte*.