Problem Set 4

Antonio Pacifico Federica Romei

April 3, 2012

1 Empirical Exercises

1. For this exercise we are going to use the dataset Affairs.dta that contains infidelity data from a survey conducted by Psychology today in 1969.

   This dataset contains observations for the following variables

   - $affairs$: How often engaged in extramarital sexual intercourse during the past year.
   - *education*: numeric variable coding level of education:
     - 9 = grade school,
     - 12 = high school graduate,
     - 14 = some college,
     - 16 = college graduate,
     - 17 = some graduate work,
     - 18 = master's degree,
     - 20 = Ph.D., M.D., or other advanced degree;
   - *gender*: dummy variable
     - =1 if female,
     - =2 if male;

   You will not use other variables hence you don't need the description.

   (a) Construct a statistical procedure to test whether male and female are not different when it comes to cheating.

   (b) Test at 5% whether people with higher level of education (above the mean) cheat as people with lower level of education (below the mean).

2. For this exercise we use the dataset CigarettesB.dta that contains cross-section data on cigarette consumption on 46 U.S. States, for the year 1992. The dataset contains the following two variables:

   - *price*: avarage price of cigarette packs per state;
   - *packs*: cigarette consumption (in packs) per person of smoking age (>16 years old).

   (a) You are working in the Health Department. You want to know whether cigarette price has an effect on smoking habits, in prticular, whether price hikes lead to a reduction in smoking. Do a regression and comment the output (*Hint:* you have to comment carefully significance of coefficients, R^2, etc.)

   (b) Based on the result of the regression, do you think you can reach a scientifically solid conclusion? Explain.

   (c) If you express prices in Euro instead of dollar, how will β_1 and β_0 change?[2]

   (d) Your boss want the regression expressed in cigarettes smoked. How will β_0 and β_1 change?

2 Theory Exercises

1. Let $X \sim Bernoulli(p)$. Define $Z = 5$.
   (a) Is Z a random variable? (Explain)
   (b) Find $E(ZX)$
   (c) Find $Var(ZX)$
   (d) Propose an estimator of $E(ZX)$

2. Let $X_1, \ldots, X_n$ denote an i.i.d. sample of size n from a random variable X. Suppose $Var(X) < 1$ and that $Z = a + bX$, for some constants a and b.
   (a) Give an expression for the sample mean of Z, $\bar{Z}$ in terms of the sample average of X, $\bar{X}$.
   (b) Show that $\bar{Z}$ is unbiased for $E(Z)$;
   (c) Show that $\bar{Z}$ is a consistent estimator of $E(Z)$ (Hint: Is the function $g(t) = a + bt$ continuous)

3. Let X denote the annual salary of professors at LUISS, measured in thousands of Euro. Suppose that $\sigma_X = 11.27$. Suppose you have at your disposal the following observations on X:

$$(X_1, X_2, \ldots, X_{10}) = \{54.74, 63.85, 70.78, 63.69, 70.16, 64.52, 55.54, 92.31, 72.96, 64.58\}.$$

[2](Assume that all the packs contain 20 cigarettes and that 1euro=1.39$.)

(a) Provide an estimate of μ_X;

(b) Construct a 90% confidence interval for μ_x;

(c) Test that $H_0 : \mu_x = 60$ against $H_1 : \mu_x > 60$ at 10% significance level.

(d) Test that $H_0 : \mu_x = 60$ against $H_1 : \mu_x \neq 60$ at 5% significance level.

4. A random sample of 500 owners of single-family homes is drawn from the population of a city. Let the random variable X denote annual household income, in thousands of Euro, and the random variable Y denote the value of the house, also in thousands of euro. The following information is available

$$n = 500$$

$$\sum_{i=1}^{n} X_i = 24.838$$
$$\sum_{i=1}^{n} Y_i = 107.226$$
$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = 66.398$$
$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 1,398,308$$
$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 194.293$$

(a) Compute the mean and standard deviation of the value of the houses in this sample. Do the same for household income;

(b) Compute the correlation coefficient between income and house value;

(c) Construct a 95% confidence interval for the mean value of houses;

(d) Construct a 90% confidence interval for the mean value of household income;

(e) Using a two-tailed test at 5% significance level, test the hypothesis that household income is equal to 110,000 against a dual-sided alternative.

5. Consider the following minimization problem[3]

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

   (a) Write down the first order conditions for β_0 and β_1;

   (b) Solve the first order conditions for β_0 to find that the optimal β_0 is given by

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X};$$

   (c) Solve the first order conditions for β_1 to find that the optimal β_1 is given by

$$\hat{\beta}_1 = \frac{\widehat{cov(X,Y)}}{\hat{\sigma}_X^2},$$

   where $\widehat{cov(X,Y)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})$ and $\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

6. Suppose that a random sample of 200 twenty-year-old men is selected from a population and that these men's height and weight are recorded. A regression of weight and height yields:

$$\widehat{Weight} = -99.41 + 3.94 \times Height, R^2 = 0.81, SER = 10.2$$

   where weight is measured in pounds and Height is measured in inches.

   (a) What is the regression weight prediction for someone who is 80 inches tall?

   (b) A man has a late growth spurt and grows 1.5 inches over the course of a year. What is the regression prediction for the increase in this man weight?

   (c) Suppose that instead of measuring weight and height in pounds and inches, these variables are measured in centimeters and kilograms. What are the regression estimates from this new centimeter-kilogram regression (*Hint: 1 in = 2.54 cm, 1 pound = 0.454 kilograms).*

7. Females, on average, are shorter and weigh less than males. One of your friends, who is a pre-med student, tells you that in addition, females will weigh less for a given height. To test this hypothesis, you collect height and weight of 29 female and 81 male students at your university. A regression of the weight on a constant, height, and a binary variable, which takes a value of one for females and is zero otherwise, yields the following result:

$$\widehat{Studentw} = \underset{(43.39)}{-229.21} - \underset{(5.74)}{6.36} \times Female + \underset{(0.62)}{5.28} \times Height,\ R^2 = 0.50,\ SER = 20.99 \quad (1)$$

[3]Notice that $\frac{1}{n}\sum_{i=1}^{n} Y_i X_i - \bar{X}\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})$, and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

where `Studentw` is weight measured in pounds and `Height` is measured in inches (standard errors in parentheses).

(a) Interpret the results. Does it make sense to have a negative intercept?

(b) You decide that in order to give an interpretation to the intercept you should rescale the height variable. One possibility is to subtract 5 ft. or 60 inches from your `Height`, because the minimum height in your data set is 62 inches. The resulting new intercept is now 105.58. Can you interpret this number now? What effect do you think the rescaling had on the two slope coefficients and their t-statistic? Do you thing that the regression R^2 has changed? What about the standard error of the regression?

(c) Calculate t-statistics and carry out the hypothesis test that females weigh the same as males, on average, for a given height, using a 10% significance level. What is the alternative hypothesis? What critical value did you use?

(d) Suppose you ran the following regression

$$Studentw = \beta_0 + \beta_1 Female + u$$

That is, you voluntarily omitted `Height` from the regression in (1). Do you expect $\hat{\beta}_1$, the OLS estimate from from the above model, to be larger or smaller than -6.36? Explain using your intuition.