Problem Set 2.

1. Results of Regressions of Average Hourly Earnings on Gender, and Education Binary variables and other characteristics using data for 1998 from the Current Population Survey. The data set consists of information on 4,000 full-time full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The worker's age ranged from $25 - 34$ years.

Regressor	(1)	(2)	(3)
College	5.46	5.48	5.44
	(0.21)	(0.21)	(0.21)
Female	−2.64	−2.62	−2.62
	(0.20)	(0.20)	(0.20)
Age		0.29	0.29
		(0.04)	(0.04)
Northeast			0.69
			(0.30)
Midwest			0.60
			(0.28)
South			−0.27
			(0.26)

Standard errors are in parenthesis. The dependent variable is Average Hourly Earnings (AHE).

College: binary variable (1 if college, 0 if high school)

Female: binary variable (1 if female, 0 if male)

Age: age (in years)

Northeast: binary variable (1 if Region is Northeast, 0 if otherwise)

Midwest: binary variable (1 if Region is Midwest, 0 if otherwise)

South: binary variable (1 if Region is South, 0 if otherwise)

Using the regression results in column (1):

a. Do workers with college degrees earn more, on average, than workers with only high school degrees? How much more? Is the earnings difference estimated from this regression statistically significant at the 5% level?

b. Do men earn more than women on average? How much more? Is the earnings difference estimated from this regression statistically significant at the 5% level?

Using the regression results in column (2):

a. Is age an important determinant of earnings? Explain.

b. Sally is 29-year-old female college graduate. Betsy is a 34-year-old female college graduate. Predict Sally's and Betsy's earnings. Is the expected difference between their earnings significant?

Using the regression results in column (3):

a. Do there appear to be important regional differences?

b. Juanita is a 28−year-old female college graduate from South.

Jennifer is a 28-year-old female college graduate from the Midwest.

bi. Calculate the expected difference in earnings between Juanita and Jennifer. Is this difference statistically significant?

2. From a sample of 200 observations the following quantities were calculated:

$$\begin{aligned} \sum X &= 11.34, \ \sum Y = 20.72, \ \sum X^2 = 12.16 \\ \sum Y^2 &= 84.96, \ \sum XY = 22.13. \end{aligned}$$

Estimate both regression equations: the regression of Y on X, and the regressions of X on Y.

In order to estimate the slope coefficients:

$$\beta_{YX} = \frac{\sum yx}{\sum x^2},\ \beta_{XY} = \frac{\sum yx}{\sum y^2}.$$

we use the following expressions

$$\begin{aligned}
\sum yx &= \sum YX - N\overline{YX} = \\
\sum YX - N\frac{\sum Y}{N}\frac{\sum X}{N} &= \sum YX - \frac{\sum Y \sum X}{N}, \\
\sum x^2 &= \sum X^2 - N(\overline{X})^2 = \\
\sum X^2 - N\frac{(\sum X)^2}{N^2} &= \sum X^2 - \frac{(\sum X)^2}{N}, \\
\sum y^2 &= \sum Y^2 - N(\overline{Y})^2 = \\
\sum Y^2 - N\frac{(\sum Y)^2}{N^2} &= \sum Y^2 - \frac{(\sum Y)^2}{N},
\end{aligned}$$

We obtain

$$\begin{aligned}
\sum yx &= 22.13 - \frac{20.72 \times 11.34}{200} = 20.95, \\
\sum x^2 &= 12.16 - \frac{(11.34)^2}{200} = 11.52, \\
\sum y^2 &= 84.96 - \frac{(20.72)^2}{200} = 82.81.
\end{aligned}$$

It follows that

$$\begin{aligned} \beta_{YX} &= \frac{20.95}{11.52} = 1.82, \\ \beta_{XY} &= \frac{20.95}{82.81} = 0.25, \end{aligned}$$

and

$$\begin{aligned} \alpha_{YX} &= \overline{Y} - \beta_{YX}\overline{X} = 0.10 - 1.82 \times 0.06 = -0.01, \\ \alpha_{XY} &= \overline{X} - \beta_{XY}\overline{Y} = 0.06 - 0.25 \times 0.10 = 0.035. \end{aligned}$$

3. Prove that r^2 (the correlation coefficient squared) in the regression of X on Y can be expressed as

$$r^2 = \beta_{yx}\beta_{xy},$$

where the β's are the LS slopes in the respective regressions.

Answer:

We have:

$$\begin{aligned} r &= \frac{\sum yx}{\sqrt{\sum x^2}\sqrt{\sum y^2}} \Rightarrow r^2 = \frac{\sum yx \sum yx}{\sum x^2 \sum y^2} \\ &= \frac{\sum yx}{\sum x^2} \times \frac{\sum yx}{\sum y^2} = \beta_{YX}\beta_{XY}. \end{aligned}$$

4. Consider the numerical example 1.4.5 in Johnston and Dinardo:

a) Obtain the regression coefficients α and β.

b) Calculate the explained and residual sum of squares as well as the correlation coefficient between Y and X (r).

c) Obtain the estimated standard errors of the regression coefficients.

Answer:

5. a)

$$\begin{aligned} \beta &= \frac{\sum xy}{\sum x^2} = \frac{70}{40} = 1.75, \\ \alpha &= \overline{Y} - \beta \overline{X} = 8 - 1.75 \times 4 = 1. \end{aligned}$$

b)

$$r^2 = \frac{(\sum yx)^2}{\sum x^2 \sum y^2} = \frac{70^2}{40 \times 124} = \frac{4900}{4960} = 0.99.$$

Moreover,

$$\begin{aligned} \text{ESS} &= r^2 \sum y^2 = 123, \\ \text{RSS} &= \sum y^2 - r^2 \sum y^2 = 124 - 123 = 1. \end{aligned}$$

c)

$$s^2 = \frac{\sum \widehat{e}^2}{N-2} = \frac{1}{3} = 0.33.$$

6. A quadratic population regression model relating test scores and income is written mathematically as

$$\text{TestScore}_i = b_1 + b_2 \text{Income}_i + b_3 \text{Income}_i^2 + e_i,$$

where income is the income in the ith district.

Estimating the coefficients of the above equations using OLS for 420 observations yields:

$$\widehat{\text{TestScore}} = \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85}\,\text{Income} - \underset{(0.0048)}{0.0423}\,\text{Income}^2.$$

Test the hypothesis that the relationship between testscore and income is linear.

ONLY FOR THE EC5501

1E. Show that if r is the correlation coefficient between n pairs of variables (X_i, Y_i), then the squared correlation between the n pairs $(aX_i + b, cY_i + d)$, where a, b, c and d are constants, is also r^2.

Answer:

Define

$$\begin{aligned} X_i' &= aX_i + b, \\ Y_i' &= cY_i + d. \end{aligned} \tag{1}$$

Averaging over all the observations gives

$$\begin{aligned} \overline{X}' &= a\overline{X} + b, \\ \overline{Y}' &= c\overline{Y} + d. \end{aligned}$$

Subtracting the above expressions from equation (1) gives

$$\begin{aligned} x_i' &= ax_i, \\ y_i' &= cy_i. \end{aligned}$$

Next, the correlation coefficient between X_i' and Y_i' is given by

$$\begin{aligned} r' &= \frac{\sum x_i' y_i'}{\sqrt{\sum (x_i')^2}\sqrt{\sum (y_i')^2}} = \\ &= \frac{\sum a x_i c y_i}{\sqrt{\sum a^2 {x_i}^2}\sqrt{\sum c^2 {y_i}^2}} = \\ &= \frac{ac \sum x_i y_i}{ac\sqrt{\sum {x_i}^2}\sqrt{\sum {y_i}^2}} = \\ &= \frac{\sum x_i y_i}{\sqrt{\sum {x_i}^2}\sqrt{\sum {y_i}^2}} = r. \end{aligned}$$

2E. Show that if r is the correlation coefficient between n pairs of variables (X_i,Y_i), then the squared correlation coefficient maybe interpreted as the proportion of Y variation attributable to the linear regression on X.

Answer:

We have to prove that

$$\sum y^2 = \sum \widehat{e}^2 + r^2 \sum y^2.$$

We start from the bivariate regression:

$$\begin{aligned} Y &= \alpha + \beta X + \widehat{e} \Rightarrow \\ \overline{Y} &= \alpha + \beta \overline{X}. \end{aligned}$$

Subtracting the second equation from the first gives

$$\begin{aligned} y &= \beta x + \widehat{e} \Rightarrow \widehat{e} = y - \beta x \\ &\Rightarrow \widehat{e}^2 = y^2 + \beta^2 x^2 - 2\beta yx \Rightarrow \\ \sum \widehat{e}^2 &= \sum y^2 + \beta^2 \sum x^2 - 2\beta \sum xy. \end{aligned} \qquad (2)$$

Next, recall that

$$\begin{aligned}\beta &= \frac{\sum xy}{\sum x^2} \Rightarrow \beta \sum x^2 = \sum xy \Rightarrow \\ &\Rightarrow \beta^2 \sum x^2 = \beta \sum xy.\end{aligned}$$

Thus substituting the above expression into equation (2), yields

$$\begin{aligned}\sum \widehat{e}^2 &= \sum y^2 + \beta \sum xy - 2\beta \sum xy = \\ &= \sum y^2 - \beta \sum xy \Rightarrow \\ \sum y^2 &= \beta \sum xy + \sum \widehat{e}^2.\end{aligned}$$

Using

$$\begin{aligned}\beta &= \frac{\sum xy}{\sum x^2} \Rightarrow \beta \sum xy = \frac{\sum xy}{\sum x^2} \times \sum xy = \frac{(\sum xy)^2}{\sum x^2} = \\ &= \frac{(\sum xy)^2}{\sum x^2} \times \frac{\sum y^2}{\sum y^2} = \frac{(\sum xy)^2}{\sum x^2 \sum y^2} \times \sum y^2 = r^2 \sum y^2,\end{aligned}$$

we obtain

$$\sum y^2 = \sum \widehat{e}^2 + r^2 \sum y^2,$$

or

$$\mathsf{TSS} = \mathsf{RSS} + \mathsf{ESS}.$$

Rearranging the above expression gives

$$r^2 = 1 - \frac{\sum \widehat{e}^2}{\sum y^2},$$

or

$$r^2 = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}}.$$