Bachelor of Science in Business Administration & Bachelor of Science in Finance and Accounting BOEK 2024

12/13/2024

Problem 1

The following sums were obtained from 10 sets of observations on Y, X_1, and X_2:

$$\sum Y = 20 \quad \sum X_1 = 30 \quad \sum X_2 = 40$$
$$\sum Y^2 = 88.2 \quad \sum X_1^2 = 92 \quad \sum X_2^2 = 163$$
$$\sum YX_1 = 59 \quad \sum YX_2 = 88 \quad \sum X_1X_2 = 119$$

Estimate the regression of Y on X_1 and X_2, including an intercept term.

Problem 2

Consider the two regressions

$$y = \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + u,$$

and

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + u,$$

where $z_1 = x_1 - 2x_2$, $z_2 = x_2 + 4x_3$, and $z_3 = 2x_1 - 3x_2 + 5x_3$. Let $\mathbf{X} = [x_1\, x_2\, x_3]$ and $\mathbf{Z} = [z_1\, z_2\, z_3]$. Show that the columns of $\mathbf{Z}$ can be expressed as linear combinations of the columns of $\mathbf{X}$, that is, that $\mathbf{Z} = \mathbf{XA}$, for some 3×3 matrix $\mathbf{A}$. Find the elements of this matrix $\mathbf{A}$.

Show that the matrix $\mathbf{A}$ is invertible, by showing that the columns of $\mathbf{X}$ are linear combinations of the columns of $\mathbf{Z}$. Give the elements of $\mathbf{A}^{-1}$. Show that the two regressions give the same fitted values and residuals.

Precisely how is the OLS estimate $\hat{\beta}_i$ related to the OLS estimates $\hat{\alpha}_i$, for $i = 1, \dots, 3$? Precisely how is α_1 related to the β_i, for $i = 1, \dots, 3$?

Consider the following linear regression:

$$y = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + u,$$

where y is $n \times 1$, $\mathbf{X}_1$ is $n \times k_1$, and $\mathbf{X}_2$ is $n \times k_2$. Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the OLS parameter estimates from running this regression.

Problem 3

Given the following least-squares estimates,

$$\begin{aligned} C_t &= \text{constant} + 0.92Y_t + e_{1t}, \\ C_t &= \text{constant} + 0.84C_{t-1} + e_{2t}, \\ C_{t-1} &= \text{constant} + 0.78Y_t + e_{3t}, \\ Y_t &= \text{constant} + 0.55C_{t-1} + e_{4t}, \end{aligned}$$

calculate the least-squares estimates of β_2 and β_3 in

$$C_t = \beta_1 + \beta_2 Y_t + \beta_3 C_{t-1} + u_t.$$

Problem 4

Consider the following regression model in deviation form:

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$$

with sample data

$$n = 100 \quad \sum y^2 = \frac{493}{3} \quad \sum x_1^2 = 30 \quad \sum x_2^2 = 3$$

$$\sum x_1 y = 30 \quad \sum x_2 y = 20 \quad \sum x_1 x_2 = 0$$

Compute the LS estimates of β_1 and β_2, and also calculate R^2.

Problem 5

The data listed in Table 1 are extracted from Koop and Tobias's (2004) study of the relationship between wages and education, ability, and family characteristics. Shown in the table are the first year and the time-invariant variables for the first 15 individuals in the sample.

Let $\mathbf{X}_1$ equal a constant, education, experience, and ability (the individual's own characteristics). Let $\mathbf{X}_2$ contain the mother's education, the father's education, and the number of siblings (the household characteristics). Let y be the log of the hourly wage.

Table 1: Subsample from Koop and Tobias Data

Person	**Education**	**ln Wage**	**Experience**	**Ability**	**Education**		**Siblings**
					Mother's	**Father's**	
1	13	1.82	1	1.00	12	12	1
2	15	2.14	4	1.50	12	12	1
3	10	1.56	1	-0.36	12	12	1
4	12	1.85	1	0.26	12	10	4
5	15	2.41	2	0.30	12	12	1
6	15	1.83	2	0.44	12	16	2
7	15	1.78	3	0.91	12	12	1
8	13	2.12	4	0.51	12	15	2
9	13	1.95	2	0.86	12	12	2
10	11	2.19	5	0.26	12	12	1
11	12	2.44	1	1.82	16	17	2
12	13	2.41	4	-1.30	13	12	5
13	12	2.07	3	-0.63	12	12	2
14	12	2.20	6	-0.36	10	12	2
15	12	2.12	3	0.28	12	12	3

a.

Compute the least squares regression coefficients in the regression of y on $\mathbf{X}_1$. Report the coefficients.

b.

Compute the least squares regression coefficients in the regression of y on $\mathbf{X}_1$ and $\mathbf{X}_2$. Report the coefficients.

c.

Regress each of the three variables in $\mathbf{X}_2$ on all the variables in $\mathbf{X}_1$ and compute the residuals from each regression. Arrange these new variables in the 15 × 3 matrix $\tilde{\mathbf{X}}_\mathbf{2}$. What are the sample means of these three variables? Explain the finding.

d.

Compute the R^2 for the regression of y on $\mathbf{X}_1$ and $\mathbf{X}_2$. Repeat the computation for the case in which the constant term is omitted from $\mathbf{X}_1$. What happens to R^2?

e.

Compute the adjusted R^2 for the full regression including the constant term. Interpret your result.

f.

Referring to the result in part c, regress y on $\mathbf{X}_1$ and $\tilde{\mathbf{X}}_\mathbf{2}$. How do your results compare to the results of the regression of y on $\mathbf{X}_1$ and $\mathbf{X}_2$? The comparison you are making is between the least squares coefficients when y is regressed on $\mathbf{X}_1$ and $\mathbf{M}_1\mathbf{X}_2$, and when y is regressed on $\mathbf{X}_1$ and $\tilde{\mathbf{X}}_\mathbf{2}$. Derive the result theoretically. (Your numerical results should match the theory, of course.)

C4

Use the data in ATTEND for this exercise.

(i) Obtain the minimum, maximum, and average values for the variables *atndrte*, *priGPA*, and *ACT*.

(ii) Estimate the model

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u,$$

and write the results in equation form. Interpret the intercept. Does it have a useful meaning?

(iii) Discuss the estimated slope coefficients. Are there any surprises?

(iv) What is the predicted *atndrte* if $priGPA = 3.65$ and $ACT = 20$? What do you make of this result? Are there any students in the sample with these values of the explanatory variables?

(v) If Student A has $priGPA = 3.1$ and $ACT = 21$ and Student B has $priGPA = 2.1$ and $ACT = 26$, what is the predicted difference in their attendance rates?

C8

Use the data in DISCRIM to answer this question. These are ZIP code–level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

(i) Find the average values of *prpblck* and *income* in the sample, along with their standard deviations. What are the units of measurement of *prpblck* and *income*?

(ii) Consider a model to explain the price of soda, *psoda*, in terms of the proportion of the population that is black and median income:

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + u.$$

Estimate this model by OLS and report the results in equation form, including the sample size and R^2. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on *prpblck*. Do you think it is economically large?

(iii) Compare the estimate from part (ii) with the simple regression estimate from *psoda* on *prpblck*. Is the discrimination effect larger or smaller when you control for *income*?

(iv) A model with a constant price elasticity with respect to *income* may be more appropriate. Report estimates of the model

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + u.$$

If *prpblck* increases by 0.20 (20 percentage points), what is the estimated percentage change in *psoda*? (Hint: The answer is 2.xx, where you fill in the "xx.")

(v) Now add the variable *prppov* to the regression in part (iv). What happens to $\hat{\beta}_1$?

(vi) Find the correlation between log(*income*) and *prppov*. Is it roughly what you expected?

(vii) Evaluate the following statement: "Because log(*income*) and *prppov* are so highly correlated, they have no business being in the same regression."

C11

Use the data in MEAPSINGLE to study the effects of single-parent households on student math performance. These data are for a subset of schools in southeast Michigan for the year 2000. The socioeconomic variables are obtained at the ZIP code level (where ZIP code is assigned to schools based on their mailing addresses).

(i) Run the simple regression of *math4* on *pctsgle* and report the results in the usual format. Interpret the slope coefficient. Does the effect of single parenthood seem large or small?

(ii) Add the variables *lmedinc* and *free* to the equation. What happens to the coefficient on *pctsgle*? Explain what is happening.

(iii) Find the sample correlation between *lmedinc* and *free*. Does it have the sign you expect?

(iv) Does the substantial correlation between *lmedinc* and *free* mean that you should drop one from the regression to better estimate the causal effect of single parenthood on student performance? Explain.

(v) Find the variance inflation factors (VIFs) for each of the explanatory variables appearing in the regression in part (ii). Which variable has the largest VIF? Does this knowledge affect the model you would use to study the causal effect of single parenthood on math performance?

C13

Use the data in GPA1 to answer this question. We can compare multiple regression estimates, where we control for student achievement and background variables, and compare our findings with the difference-in-means estimate in Computer Exercise C11 in Chapter 2.

(i) In the simple regression equation

$$colGPA = \beta_0 + \beta_1 PC + u,$$

obtain $\hat{\beta}_0$ and $\hat{\beta}_1$. Interpret these estimates.

(ii) Now add the controls *hsGPA* and ACT—that is, run the regression *colGPA* on *PC*, *hsGPA*, and ACT. Does the coefficient on PC change much from part (ii)? Does $\hat{\beta}_{hsGPA}$ make sense?

(iii) In the estimation from part (ii), what is worth more: Owning a PC or having 10 more points on the ACT score?

(iv) Now to the regression in part (ii) add the two binary indicators for the parents being college graduates. Does the estimate of β_1 change much from part (ii)? How much variation are you explaining in *colGPA*?

(v) Suppose someone looking at your regression from part (iv) says to you, "The variables *hsGPA* and ACT are probably pretty highly correlated, so you should drop one of them from the regression." How would you respond?