Final BUEC 333. Version D.

August 11, 2014, 12:00-15:00

Questions

1. Consider the example used in the chapter on panel data. We have a panel data set on $n = 48$ U.S. states during $T = 7$ periods, from 1982 up to and including 1988. The total number of observations is 336.

   (a) Is this a balanced panel? Explain. **Yes: 7*48=336, so there are not gaps.**

   (b) For each state, in each time period, let Y_{it} denote the number of annual traffic deaths per 10000 in the population. Let X_{it} denote the beer tax in 1988 U.S. dollars. Temporarily ignore the data after 1982, so that we have a cross-section of 48 states. The estimated regression line gives

   $$\hat{Y}_{i,1982} = 2.01 + 0.13X_{i,1982}.$$

   Alternatively, we can use fixed effects regression to estimate the effect fixed effects regression line is

   $$\hat{Y}_{i,t} = \underset{(0.29)}{\hat{\alpha}_i - 0.66X_{i,t}}.$$

   Do you think that the Least Square assumptions hold, i.e. do you believe that the 0.13 in the first result comes from an unbiased estimator? If YES: explain what causes the difference between 0.13 and -0.66. If NO: explain why the Least Square assumptions are unlikely to hold. Include in your answer: "tax on beer". **They do not hold: there are omitted variables that the FE accounts for.**

2. [Based on SW, 14.7] Suppose that Y_t follows the stationary AR(1) model

   $$Y_t = 3.9 + 0.5Y_{t-1} + u_t$$

   where u_t is i.i.d. with $E(u_i) = 0$ and $\text{var}(u_i) = 9$.

   (a) Compute the mean and variance of Y_t. **E(Yt) = 3.9/(1-0.5)=7.8. Var(Yt) = 9 / (3/4) = 12**

   (b) Compute the first autocovariance of Y_t. **Cov(Yt,Yt-1)=0.5*12=6**

   (c) Compute the second autocovariance of Y_t. **Cov(Yt,Yt-1)=0.5^2*12 = 3**

3. [Stock and Watson, 2.5] "In September, Seattle's daily high temperature has a mean of 59 degrees Fahrenheit. The standard deviation is 9 degrees Fahrenheit." Remember that, to convert from degrees Fahrenheit to degrees Celsius, we need to subtract 32 and then multiply by 5/9, so

   $$T_C = \frac{5}{9} \times (T_F - 32).$$

   For the questions that follow, indicate which formulas you are using.

   (a) What is the mean of Seattle's daily high temperature in September in degrees Celsius? **5/9*(59-32)=15**

   (b) What is the standard deviation of Seattle's daily high temperature in September in degrees Celsius? **5/9*9=5**

   (c) What is the variance of Seattle's daily high temperature in September in degrees Celsius? **5*5=25**

4. Suppose you want to estimate the population mean of Y, $\text{E}(Y) = \mu_Y$. You have a random sample of size n, $\{Y_1, \cdots, Y_n\}$. For simplicity, assume that $n = 2$. Then, the sample average $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, simplifies to $\bar{Y} = (Y_1 + Y_2)/2$.

   (a) Is the sample average $\bar{Y}$ unbiased for μ_Y? Explain. **E(1/2(Y1+Y2))=1/2 E(Y1+Y2) = 1/2 * (E(Y1) + E(Y2)) = 1/2 * 2 mu = mu**

(b) Now, consider $\tilde{Y} = \frac{1}{4}Y_1 + \frac{1}{4}Y_2$. What is the variance of $\tilde{Y}$? On the basis of the variances, do you prefer $\bar{Y}$ or $\tilde{Y}$? **1/8 σ_Y^2, which is smaller than $var\left(\bar{Y}\right)$, so you prefer $\tilde{Y}$**

(c) What is wrong with $\tilde{Y}$? **It is biased**

(d) Now, consider the estimator $\check{Y} = \left(\mu_Y + \bar{Y}\right)/2$. Why is this not a good estimator? (Hint: the answer has nothing to do with efficiency, unbiasedness, or consistency.) **The estimator is based on the unknown quantity that yo u are trying to estimate!**

5. Consider the following estimated regression equation that describes the relationship between a student's weight and height:

$$\widehat{WEIGHT} = 100 + 6.0\,HEIGHT$$

(a) A student has height 5. What is the regression's prediction for that student's weight? **100+6*5=130**

(b) In the sample, the sample average of HEIGHT is 4. What can you say about the sample average for WEIGHT? **124**

Now, an additional variable is included, is ID, a student's SFU ID. Obviously, this is a nonsensical variable to include: it is not in any way related to a student's weight. The new estimated regression equation is

$$\widehat{WEIGHT} = 101.5 + 5.98\,HEIGHT + 0.02\,ID$$

(c) Someone's weight has nothing to do with their SFU ID. Still, the R^2 went up from 0.74 to 0.75. How is this possible? **The R^2 never decreases**

(d) If the post office box number is not related to a student's weight, should the estimated coefficient not be equal to 0? How could it be that it is 0.02? **Sampling variability. Even if $\beta_{ID} = 0$, $\hat{\beta}_{ID} \sim \mathcal{N}\left(0, \sigma_{\hat{\beta}}^2\right)$**

6. Let D_i be a dummy variable. Consider the model that consists of the equation

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 D_i X_i + u_i$$

and the standard OLS assumptions

(a) Draw a graph to visualize this model. **One graph, two lines. One is labelled $E\left(Y_i|X_i, D_{,i} = 1\right)$, the other $E\left(Y_i|X_i, D_i = 0\right)$. The horizontal axis is labelled X_i. The two lines have different slopes, and different intercepts. The intercepts (β_0, $\beta_0 + \beta_2$) and slopes ($\beta_1, \beta_1 + \beta_3$) are clearly marked in the graph.**

(b) What is the interpretation of β_3? **Difference in slopes $\partial E\left(Y_i|X_i, D_{,i}\right)/\partial X_i$ between $E\left(Y_i|X_i, D_{,i} = 1\right)$, $E\left(Y_i|X_i, D_i = 0\right)$.**

7. [Based on SW, Exercise 8.7] This problem is inspired by the study of the gender gap in top corporate jobs in Bertrand and Hallock (2001). The study compares total compensation among top executives in a large set of U.S. public corporations in the 1990s.

(a) Let $Female$ be an indicator variable that is equal to 1 for females and to 0 for males. A regression of the logarithm of earnings onto $Female$ yields

$$\log\left(\widehat{Earnings}\right) = 6.48 - 0.44 Female$$

where the estimated regression coefficient -0.44 has a standard error of 0.05. Explain what the -0.44 means. **In expectation, c.p., women earn 44% less.**

(b) Does this regression suggest that there is gender discrimination? Explain. **No: omitted variables.**

(c) Two new variables are added to the regression: $\log\left(MarketValue\right)$, where $MarketValue$ is a measure of firm size, in millions; and $Return$, the stock return, in percentage points. The resulting estimated regression line is

$$\log\left(\widehat{Earnings}\right) = 3.68 - 0.28 Female + 0.37 \log\left(MarketValue\right) + 0.004 Return$$

where the standard errors for the three regressors are 0.04, 0.004, and 0.003, respectively. The coefficient estimate for $Female$ has changed from -0.44 to -0.28. Why has it changed? **Omitted variables in the first regression**

8. A mixed bag of questions:

   (a) Describe the difference between "internal validity" and "external validity". **Book**

   (b) List two threats to internal validity. (The book lists five). **Book**

   (c) What are the two conditions an instrumental variable must satisfy? **(1) Relevance and validity, OR (2)** $\mathbf{E}(u_i|Z_i) = 0$ **and** $Cov\left(Z_i, X_i\right) \neq 0$**.**

9. In the context of your second hand-in assignment, consider the following code and output. Why are there "NA"s in the row for "occ9"? **Multicollinearity / dummy variable trap.**