Professor Joyce Jacobsen

Economics 300
Summer 2010

Assignment #12

Due Friday 7/2/10 by 6 p.m. in the Econ 300/QAC201 slot in the Economics Alcove

Please show the calculations used to arrive at your answers. Round final answers to the second decimal place if necessary.

A. A random sample of 500 father-son pairs found a correlation of .60 between the heights of father and son. The heights of the fathers had an average of 68 inches and a standard deviation of 2 inches; the heights of the sons had an average of 69 inches and a standard deviation of 3 inches.

1) What height would you predict for a son drawn at random if the father's height is unknown?

2) What height would you predict for a son drawn at random if the father's height is 72 inches? Include a 95% confidence interval.

B. Suppose you know that for a particular sample of people, wages are a linear function of person i's age, sex, race, and union status:

$$wage_i = \beta_0 + \beta_1(age_i) + \beta_2(union_i) + \beta_3(race_i) + \beta_4(sex_i) + e_i$$

Where age is measured in years,
sex is a dummy variable (0 = woman, 1 = man),
race is a dummy variable (0 = nonwhite, 1 = white),
union status is a dummy variable (0 = nonunion, 1 = union),
and $e_i$ is a random error term with zero mean and no serial correlation

But your data set only contains information on each person's age in 2007, their wage in 2007 and 2008, and their union status in each of 2007 and 2008.

1) What are the problems, if any, with running the following regression, using 2007 data?

$$wage = b_0 + b_1(age) + b_2(union)$$

In particular, how would you expect the estimated effects of age and union status to differ from the true effects of age and union status?

2) Show how to transform the true regression equation from levels to first differences.

3) Explain how estimating the regression equation in first differences solves at least some of the problems you found for the regression in **1**.

C. To forecast monthly sales of its milk, a dairy uses an exponential smoothing formula:

Forecast = 0.25*(last month's sales) + 0.75*(forecast of last month's sales)

1) The actual sales in March were 640 units higher than the forecast sales of 27,710 units. What is the forecast for April?

2) If April sales turned out to be 27,760, what is the forecast for May?

D. The following are U.S. 2004-2007 quarterly data on the number of unemployed persons (in thousands):

| | Quarter | | | |
|---|---|---|---|---|
| Year | I | II | III | IV |
| 2004 | 8916 | 8082 | 8001 | 7598 |
| 2005 | 8326 | 7498 | 7475 | 7064 |
| 2006 | 7518 | 6933 | 7104 | 6446 |
| 2007 | 7320 | 6771 | 7199 | 7020 |

The data above were fitted with a multiple regression (for T = 1, 2, …, 16):

$$Y = 8690 - 96T - 603Q_2 - 384Q_3 - 676Q_4$$

1) Calculate the residual $e_t$.

2) Calculate r, using a calculator, computer program (spreadsheet or statistics program), or graph, for the simple autoregression:

$$e_t = re_{t-1}$$

3) Calculate the predicted values using the multiple regression above for the four quarters of 2008. Compare the predicted to the actual values of 8067, 8099, 9370, and 10161: Is there an apparent systematic prediction bias upwards or downwards?

4) Given your answers to **2** and **3**, is it possible to improve on this regression by utilizing the information in the estimated residuals? Explain.

E. Assume you are a funding officer at the National Institutes for Health. You have sent out a call for research proposals concerning the effects of a new drug treatment for AIDS. Consider the scientific merit and the ethical and political aspects of the following four proposals. Assuming you only have money to fund one of the four projects, which one would you pick, and why? Can you think of a better proposal of your own? Be explicit as to how it is better.

a. Draw a random sample of n persons with AIDS. Randomly assign dosage levels D to these people. For each person, record the drug dosage level and how long the person lives (measured in months M from the time the study begins). The study continues until all the persons have died. Then calculate the regression line of M against D, interpreting the coefficient of D as the effect of the drug.

b. Draw a random sample of n persons with AIDS. Randomly assign and record dosage levels D for these people and collect background information on them, including age, sex, nutritional habits, etc. After 5 years, find out if they are still alive or not (so M = 0 if dead, 1 if alive). Then calculate the multiple regression of M on D and all of the background variables, interpreting the coefficient of D as the effect of the drug.

c. Draw a random sample of n persons with AIDS. Give half of them a standard dosage of the drug and half of them a placebo. After five years, find out how many in each group are still alive. Then construct a 95% confidence interval for the difference in mortality rates between users and nonusers.

d. Ask for volunteers until you have gathered a group of n persons with AIDS. For each person, flip a coin to decide whether or not the person will receive the standard dose of the drug or the placebo. After five years, find out how many in each group are still alive. Then construct a 95% confidence interval for the difference in mortality rates between users and nonusers.