Professor Joyce Jacobsen

Economics 300
Summer 2010

Assignment #9

Due Monday 6/28/10 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. Use of earnings data in a text file and a statistics program to study earnings-experience profiles.

The dataset baseball.txt contains 818 observations on major league baseball players from 2009. The set has three columns of data, separated by tabs:

1) 2009 earnings (in thousands) for the player (ranges from 400 to 33000)
2) year that the player first entered the major leagues (ranges from 1988 to 2009)
3) 1 if the player is a pitcher and 0 if the player is not a pitcher ("batter")

1) Download this dataset from the course webpage, read it into your statistics program, and calculate the years of experience in the majors for each player as of 2009.

2) Construct a confidence interval to test whether pitchers and batters earn the same on average. Do you reject or fail to reject this hypothesis?

3) Plot earnings vs. years of experience for all players. Looking at the graph, does there appear to be a relationship? Now do this for pitchers and batters separately. Does there appear to be a stronger relationship for one group than for the other?

4) Calculate the regression line: earnings = a + b*(years of experience). Looking at the regression equation, does there appear to be a relationship between earnings and experience? Now do this for pitchers and batters separately. Does there appear to be a stronger relationship for one group than for the other?

5) Discuss how to go about "improving" on the calculations in **4** of the relationship between earnings and experience for baseball players. [Hint: think about what other factors you might want to control for to get a "clean" estimate of the effect of experience on earnings for players]

B. A random sample of 122 Americans showed the following relationship between annual earnings Y (in dollars) and education X (in years):

$$Y = \$14{,}000 + \$900X$$

Average earnings $\overline{Y}$ = \$24,800 and average education $\overline{X}$ = 12 years, with $\Sigma x^2 = 800$. The residual standard deviation about the fitted line is s = \$8400.

1) Calculate a 95% confidence interval for the population slope. Is the relation of earnings to education statistically discernible at the 5% level?

2) Calculate the p-value for the null hypothesis that earnings do not increase with education.

3) Predict the earnings of a man who has completed 3 years of high school (X = 11). Include an interval wide enough that you would bet on it at odds of 95 to 5.

4) Can you conclude from this that each year of education is worth \$900 in increased annual earnings? Why or why not?

C. [Note. You can use the computer to help you do this problem, or do it all by hand.]
Suppose a random sample of 5 families had the following annual after-tax incomes and savings:

Family	Income X (in $000s)	Savings S (in $000s)
A	45	4
B	50	6
C	35	3
D	25	1
E	60	9

1) Estimate the population regression line $S = \alpha + \beta X$

2) Graph the five points and the fitted line.

3) Construct a 95% confidence interval for the slope β. Can you reject the hypothesis that $\beta = 0$?

4) Theory suggests that the population marginal propensity to save (β) is nonnegative. Accordingly, you use instead a one-sided test. State the null and alternative hypotheses in symbols, calculate the p-value for H_0, and construct the relevant one-sided 95% confidence interval. Based on these results, can you reject the hypothesis that β is negative?

D. Assume (heaven forbid) that after I have graded all the finals, I lose your test before recording your grade. I then need to predict what your grade was. There are several methods I could use:

a. Assume that your grade was the same on the final as it was for the rest of the course, so:

$Y = X$, where Y = % correct on the final and X = % correct on earlier coursework

b. Assume that your grade was the same on the final as the average for the class, so:

$Y = \overline{Y}$

c. Fit a regression line to the data and substitute in your X to get the estimate of Y:

$Y = a + bX$

Which method will yield the best prediction (where "best" means "on average, closest to the true value")? Explain carefully why you have rejected the other two methods as inferior.

E. The average smog level in Middletown (Y), measured on an index, and average number of cars driven down Main Street (X), measured in thousands, was recorded monthly for 5 years and the following data resulted:

$\overline{X} = 27.38$ $\quad \sum x^2 = 14350.18$ $\quad \sum xy = 6236.75$

$\overline{Y} = 26.75$ $\quad \sum y^2 = 20327.25$ $\quad n = 60$

$Y = 14.85 + 0.43X$ $\quad s = 17.43$

1) Verify the regression equation.

2) Guess the long-run average of the Middletown smog index if the number of cars were held steady at 25 (thousand). Along with your guess, include a 95% confidence interval.

3) If another month is sampled, guess the smog index value if the number of cars is 25 (thousand). Along with your guess, include an interval that you are 95 percent sure contains the true value.