

Problem set 9

1. Access the data in the **burglary.xls** file, which contains information about burglary arrests and employment levels for 90 counties in the United States. Conduct a regression of Burglary Arrests vs. Employed (which contains the number of employed people in the civilian workforce in that county.)
 - (a) What do these results suggest?
 - (b) Are these results surprising to you?
 - (c) Identify any counties that are outliers or highly leveraged or influential observations.
 - (d) What is the probability that a normal random variable will be over 5.6065 standard deviations from the mean (as the LA County residual is)?
2. Access the **beerdata.xls** dataset, which contains data on beer consumption and income levels per capita for 19 European counties. Conduct a regression of beer consumption vs. income levels per capita.
 - (a) On average, as income increases by \$1,000 per capita, how much does beer consumption increase?
 - (b) Does this relationship make sense?
 - (c) How would you answer to part a) if the outliers were removed from the data? (This is generally not a good idea, but we are using the removal of outliers to see how strongly they impact some of our results.)
3. A Midwestern hotel chain has noticed much variation in its electricity costs and would like to be able to explain these changes for planning and budgeting reasons. It has collected samples from random hotels during random months during the past years. The variables include the hotels' electricity costs per room and the average temperature that month. These data are available in the **electricitycosts.xls** file. Use R to conduct a regression of electricity costs per room vs. average temperature.
 - (a) Does the relationship seem significant?
 - (b) Plot residuals versus predicted values for this regression. Does this graph give you any thoughts on improving the model?
 - (c) Use the tools discussed in class to build an improved model.

4. Headhunter Inc.

Headhunter Inc, a firm specialized in worker recruiting for other firms, wants to perform an analysis of the productivity of their typewriting clerks. They suspect that some of them are using computer skills to waste time in Facebook and Twitter instead of their typewriting jobs. For this they collect some data, which you can find in the file **words.xls**, on number of words typed per minute (Words), achieving score on a computer test (Computer) and experience measured in years of seniority in the job (Experience).

- (a) They first run a regression to analyze the effect of computer skills on words typed, and they concluded that there is no evidence that computer skills are being used in social networks, because the effect is positive and statistically significant. Do you agree?
- (b) Afterwards they included also Experience as an explanatory factor in the above regression. Is the regression in (a) biased? What is the size of the bias?
- (c) Use an auxiliary regression of Experience as dependent variable and Computer as explanatory variable to explain the possible bias in (a) and (b), and to give a precise explanation on the effects observed (here you have to say what is the direct effect of Computer on Words, and what is the indirect effect).

5. The effects of different variables on education

A consultancy firm wants to analyze the effects of different variables on wages in the telecommunication sector in the UK. For this they buy a data set from the UK Office of National Statistics, consisting on survey with 935 family heads.

The data set **education.xls** consists in observations of wages (logarithm of monthly earnings), education (in years), IQ score and tenure (years with the actual employer). The label of the variables are lwages, educ, IQ and tenure, respectively.

- (a) Suppose that the true model is

$$\log(wages)_i = \beta_0 + \beta_1 educ_i + \beta_2 tenure_i + \epsilon_i$$

where the variables in the equation are the ones defined before and ϵ_i is the error term for individual i . Estimate the return to education coefficient. What is the percentage increase in wages if

education increases in one year? Now imagine that the data set does not include *tenure*, so you have to estimate

$$\log(wages)_i = \beta_0 + \beta_1 educ_i + \epsilon_i$$

Estimate the return to education coefficient. How are these estimates compared? Are they similar? Why? Verify your findings by running a regression of *tenure* on education.

- (b) Using the model that we call “true” in the first question, test if a year of education is as worth as a year of *tenure*. Indicate clearly the null and alternative hypothesis.
- (c) Suppose now that in the true model *tenure* is replaced by *IQ*:

$$\log(wages)_i = \beta_0 + \beta_1 educ_i + \beta_2 IQ_i + \epsilon_i$$

Estimate the return to education coefficient. Now estimate a “short” version of the true model,

$$\log(wages)_i = \beta_0 + \beta_1 educ_i + \epsilon_i$$

Estimate the return to education coefficient by OLS. How do these estimators compare? Verify the omitted variable bias and compare to the estimates that we had when we omitted *tenure* instead of *IQ*.