Statistics Citadel
Jeffrey R. Russell
Homework 4

1. Use the dataset censuswage.xls. This dataset contains wages per hour and years of education for a sample of workers from the 1990 US census.

   Run a regression of wages on education (ie put wages on the left side and education on the right side).

a. Find a 95% confidence interval for the mean wage of all high school graduates. Verify the interval using R.
b. Use the formulas in class to find an interval in which there is a 95% chance that a high school graduate's wage would fall in. Verify the interval using R.

2. Use the ceosalary dataset for this question. This data set contains the salary and other information for a sample of ceo's. Regress CEO salary (salary) (left hand side variable) on years with the company (comten), years as CEO (ceoten), and sales (sales). Report all coefficient point estimates, their standard errors, and 95% confidence intervals for the true coefficients. What evidence does this regression provide about the following implications of two hypothetical corporate finance theories?

a. Theory A implies that CEOs' overall time with their firm is an important determinant of their compensation and controlling for the firm's sales and how long the person has been CEO, their pay should be higher the longer they have been with the company overall.
b. Theory B implies that overall tenure with the firm and sales are irrelevant, the only variable that is important for CEO compensation is how long the person has been CEO.

3. The data for this question are in murder.xls. The file contains data on the murder rates for different states in each of 3 different years: 1987, 1990, and 1993. Here are the variable descriptions.

Variable	Description
state	State identifier
year	Year: 1987, 1990, or 1993
mrdrate	Murder rate in that state during the given year; defined as murders per 100,000 population
exec	number of executions in that state during the year
unemp	unemployment rate in that state during the year; a percentage

Our goal in this problem is to investigate how the murder rate is related to the unemployment rate and the presence of capital punishment. We are interested in estimating the model:

$$Mrdrate_i = \alpha + \beta_1 Unemp_i + \beta_2 CapPun_i + \varepsilon_t$$

Where $Mrdrate_i$ is the murder rate, $Unemp_i$ is the unemployment rate, and $CapPun_i$ is a dummy variable that equals one if the state employed capital punishment in a given year and zero otherwise.

a. First off, the variable **exec** in this data tells us *how many* executions were performed. We are interested in *whether* or not executions were performed.
Define a dummy variable $CapPun_i$ that takes the value 1 if the state performs capital punishment and zero if not.

Now run the regression of **mrdrate** on **unemp** and the dummy variable you just created.

b. What is the interpretation of the coefficient β_2? Explain in your own words.

c. Test the null hypothesis $H_o: \beta_2 = 0$ at the 5% level. Do you reject? Consider the following statement: "Controlling for economic conditions, the murder rate is positively correlated with the presence of capital punishment". Does this dataset provide evidence for or against this statement? How strong is this evidence?

d. Suppose a given state has an unemployment rate of 6%. What is the 95% predictive interval for the murder rate if the state DOES NOT employ capital punishment? What is the 95% predictive interval for the murder rate if the state DOES employ capital punishment?
Are these intervals "big"?

e. Now make a scatterplot of mrdrate versus unemp. Do you notice anything?
It turns out that the state code "9" corresponds to the District of Columbia (Washington DC). Print out your scatterplot, circle the three DC observations, and turn it in.
Delete the three DC observations from the data set and re-run the regression. How do your answers to parts (c) and (d) change?
Discussion: Give an intuitive explanation as to WHY your answers to parts (c) and (d) change when the DC observations are dropped.
Also, be VERY CAREFUL in interpreting results of regressions like these! Remember there's no hard and fast rule about whether we "should" drop outliers, the point here is that we need to be AWARE of them. More importantly, remember that *correlation does not imply causation.* It could be really misleading to say that capital punishment "causes" murders!!

4. Business week and U.S. News and World Report publish rankings of the top 20 Business Schools. The Business Week overall ranking is based on rankings obtained from students and from firms that recruit MBAs. Along with the rankings, the magazine reports information about the cost of getting an MBA

degree and the graduates' average starting salaries. The file rank.xls contains data on the 1992 Business Week and U.S.NWR rankings. It also contains data on the overall cost of getting an MBA and the average entry salaries for graduates of the top 20 schools in that year.

a. What is the correlation between overall cost and entry salary?

b. Plot a scatter plot of salary versus cost. What do you see?

c. We are interested in answering the question "is there a payoff from spending more?" and run a regression. What value is the slope estimate? What do you conclude based on the value of the slope estimate alone?

d. Do you think you need to be careful when interpreting the result from part (c)? (Hint: what does the scatter plot from point (b) suggest to you?)

Basic Statistics Citadel

Prediction Challenge

NOT DUE UNTILTUESDAY

Use the rent data in the files **rent596.txt** and **rent100.txt** to answer question 1. This data is for 596 apartments in Chicago in 1999 and contains the following variables: monthly rent in dollars, number of bathrooms, the year the apartment's building was built, square feet (in 100s), an indicator equal to one if the apartment has air conditioning and zero otherwise, an indicator equal to one if the apartment comes with a parking space and zero otherwise, and the number of rooms in the apartment. The data has already been divided into two parts for an out-of-sample prediction performance comparison. The 596 data points in the file **rent596.txt** are the estimation sample and the remaining 100 observations in the file labeled **rent100.txt** are the testing sample set aside for prediction comparisons.

Suppose that you have decided to enter the "Try and Out-predict the professor (and all the other study groups)" contest. The goal of this contest is to build a regression model for rents to use for predicting rents given apartment characteristics that outperforms the professor's favorite regression model and those of the other study groups. Your grade for this problem will NOT depend on your actually being able to build a model that does better than the professor's or those of the other study groups, only upon a good faith effort to do so. However, no contest is complete without a prize, so there will be a (small) prize awarded to the study group that out-predicts the professor's model by the largest margin according to the out-of-sample prediction exercise that will be held in class on Wednesday. This data is real. I have deleted any obvious errors in the final 100 test observations.

a. First, using the 596 observations in your estimation sample, replicate Jeff's favorite rent regression. This is a regression of rent upon a constant, number of bathrooms, square feet, number of rooms, number of rooms squared, and an indicator for whether it has air conditioning.

b. Use your insight to construct two different regression specifications to use to predict rents (hopefully better than Jeff's) as a function of the apartments' characteristics. Define variables, write out regression equations, and then estimate both using the 596 observations in the estimation sample. Present point estimates, standard errors, and R-squared for these two regressions. Note that you may consider any transformations of the raw data you believe would be relevant in helping you to predict rents. (E.g. Jeff's regression includes a new variable, number of rooms squared, that is meant to allow for the possibility that the expected change in rent when the number of rooms increases may not be constant.)

c. Now use both Jeff's regression estimates and your two sets of regression estimates from part b. to construct predictions for rents for each of the 100 observations in the testing sample. Compute and report sums of the squared prediction errors for all three models: Jeff's and the two you have created. Which model does the best according to this criterion? Summarize the relative performance of your models to Jeff's by computing the ratios of each of your two models sum of squared prediction errors divided by the sum of squared prediction errors for Jeff's model.

d. Choose the best model from the competition you conducted in part (c) to be your contest entry specification. In order to get the most precise estimates of the parameters in your entry specification, re-estimate this model using BOTH the estimation sample and the prediction testing sample: all 696 observations. Report your estimated parameters and R-squared on your homework write-up. Be sure to remember your regression specification. Designate at least one group member to bring a laptop to class on Wednesday with a copy of the full dataset (696 observations) and be prepared to reproduce your regression and form fitted values using additional data. I will make available a new data set consisting of 100 new observations. The winner of the prediction contest will be determined in class when each study group conducts an out-of-sample prediction exercise (using squared forecast errors as the measure of prediction quality) with additional data that I will make available at the start of class.