Spring 2011 - STAT 515 – Project Part II

The Analysis of Two Related Variables

The goal of this assignment is to analyze two quantitative variables (that may be related to each other) to see if you can predict one from the other. The data set should consist of a set of individual people or things (say, 25 or more) *on which two variables have been measured*. Both variables need to be continuous (or at least have a *large number* of different levels if discrete). You should have already described your data set to the instructor, who has approved it.

The goal is to analyze the data and to present the results so that someone who has not had a statistics course could understand them. For example, if you use the median as a descriptive statistic you would need to briefly explain to the reader what the median is and why you chose it. When you report the p-value of a hypothesis test, you need to explain what it means and why you would probably reject (or fail to reject) the null hypothesis. You don't need to explain how the tests of hypotheses work, but you do need to explain what the assumptions are.

The project will have to address five main questions:

1) What question are you trying to answer? (e.g., *Can the height of students be used to predict how far someone can jump*?)

2) Why is this question of interest? (e.g., *In grade school one of the tests in gym class is to see how far you can jump. Is this fair to people who are short?*)

3) How was the data gathered, and what limitations does this imply? How would you overcome these limitations? (e.g., *Only students in the fifth period gym class were used, this is bad because…*)

4) Describe the two variables individually. (e.g., *The average height was… Jumping distance was skewed right….*)

5) Describe the relationship between the variables. (e.g., *The jumping distance is predicted to increase by … for each additional inch of height…*)

The paper should be typed, using complete sentences, good grammar, and transition between the various sections. If you are using data collected by someone else, reference the source appropriately. The paper should be between 3 and 5 pages long, excluding any graphs. Some additional specifics of what must be included can be found on the back of this sheet.

The project is due **on or before 3 p.m. Monday, May 2.**

In the past, students have chosen inappropriate data (not continuous, for example) or done the analysis in reverse (predicted *x* from *y* instead of *y* from *x*). Both of these are grounds for receiving a poor grade. Forgetting to answer several of the questions also results in low grades.

Specifics for the Spring 2011 STAT 515 Project

1) If the data come from a sample: Define the desired target population and describe how the sample was collected. If you were not able to sample from the desired population, state what differences you might expect between the population that was actually sampled from and the desired target population. If you were not able to take a simple random sample (page 154) from the population, discuss how the sampling could be improved if you were allowed more money and time.

If the data come from an experiment: Describe how the experiment was carried out, describe any sources of extra variation (e.g. changing temperature, different people making the measurements, etc...). Did you try to control these? Discuss how the experiment could be improved if you were allowed (more) money and time.

2) When describing the variables individually, give the appropriate plots and descriptive statistics to succinctly but thoroughly describe the data. That is, decide which of the graphs **and** statistics best describe the data to the reader. Give readers help in interpreting the graphs and statistics by telling them what they should be seeing.

Construct confidence intervals for the means of each of the variables, separately. Interpret them and say if we can trust these intervals or not (that is, are the assumptions met?).

3) The Model: Fit a linear regression model to your data. Be sure to state what model you are attempting to fit to the data in terms of the variables you are using.

Statistics: The report of the regression you performed should include the following statistics: the estimated regression line, a confidence interval for the slope, the p-value for testing whether the slope is zero, and the coefficient of determination (r^2). Make sure and tell the reader why these statistics should be useful to them, and interpret them in the context of your data set.

Graphics: Give the scatter plot of the data with the regression line.

Assumptions: Check the assumptions needed for the regression. If the assumptions are not met, then don't forget to point out to the reader that they can't entirely trust the confidence intervals and hypothesis tests you found when performing the regression. If you find any outliers, see if they have a significant effect on your results by running the regression again without them and seeing if your regression line changes much. (You don't need to write up all the details on this new regression though!)

4) Finally, don't forget to include a short summary at the end of your paper to tie everything together!

Grading:
The project is mandatory for graduate students. It will be graded out of 30 points, of which the preliminary part is worth 5 points. As an encouragement for working in groups, you will get 2 bonus points if you work in a group of two or three people. When working in groups, each member should contribute significantly to the project. The project is *extra credit* for undergraduate students. Undergraduates choosing to do the project will have 18% of their project grade (i.e., the grade out of 30) added to their final exam score.