Fall 2002 - STAT 515 - Project

The Analysis of Two Related Variables

The goal of this assignment is to analyze two quantitative variables (that may or may not be related to each other) to see if you can predict one from the other. The data can either be from a census, a survey, or an experiment. It may be real data you found in a published source (not the text book, and not one that has already analyzed the data!) or it may be data you gathered yourself. Both variables need to be continuous (or at least have a large number of different levels if discrete).

The goal is to analyze the data and to present the results so that someone who had 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 accept or 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 estimated 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 by 4:30 pm on Friday, December 6th. Various homework assignments between now and then will be related to the project, but you are free to change your chosen topic at any point. You should be sure to get approval of the data you are using before doing too much work however. In the past students have chosen inappropriate data (not continuous for example) and received failing grades on the project.

Specifics for the Fall 2002 STAT 515 Project

3) If the data comes 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 150) from the population, discuss how the sampling could be improved if you were allowed more money and time.

If the data comes 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.

If the data is census data or fixed measurement (e.g. area of states or reported death rates): Describe how the measurements were gathered, how these measurements have changed over time, what some alternate census levels are (e.g. State vs. PMSA), whether it seems to be reasonable that these results will be the same in the future, and whether the variables need to be adjusted to remove the effects of other variables such as total population or cost of living.

4) 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 the reader help in interpreting the graphs and statistics by telling them what they should be seeing.

If the data is from an experiment or sample, construct confidence intervals for the means and standard deviations of the variables. Say if we can trust these intervals or not (that is, are the assumptions met).

5) 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 that the slope is zero, the coefficient of determination (r^2) , and the standard error about the line (square root of MSE). 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!)

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