Project Module 1 – Due Monday, Sep. 29

STAT-UB.0103 – Statistics for Business Control and Regression Models

Working with real data is the best way to learn statistics. This project gives you an opportunity to do just that.

First, choose a group of 2–4 people. **One person in your group must email the TA with your group members' names before 11:59PM on Friday, Sep. 12**. Alternatively, if you would like the TA to find a group for you, you can email him and let him know.

Once you have a dataset, **you must email me the data and a scatterplot of your data (as described in question 2 below) before 11:59PM on Sunday, Sep. 21**. Although it is not required, I recommend that you come by my office for about 5 minutes to discuss your plans for the project. You can come during office hours, or if you would like to meet with me at a different time, please email me to set up an appointment.

Your writeup is due by the end of office hours on Monday, Sep. 29.

Dataset Selection

First, you must find or create a dataset with at least 30 observations and at least 3 variables. That is, **the dataset must have at least 30 rows, storing the observations, and at least 4 columns, storing the names of the observations and variables;** in total, there must be at least 30 names and $30 \times 3 = 90$ values for the variables.

To find a dataset, first pick a topic or type of observation that you find interesting. Here are some example topics:

- Albums
- Celebrities
- Cities, States, or Countries
- Companies
- Movies
- New York City restaurants
- Professional athletes
- Smartphone Apps
- Sports teams

There are many other possibilities.

You must choose a dataset that you are actually interested in. All of your data must be from the same year, and your data must be no more than 5 years old. In particular, you are not allowed to use any of the following topics: Oysters, Norweigian roads, or SAT scores.

Once you pick a topic, choose at least 30 individuals and choose identifying names for the individuals. Next **pick at least 3 variables**, and record the values of these variables for each individual in the dataset. At least 2 of the variables must be quantitative.

For example, if you pick “iPhone Apps” as your topic, you should select 30 applications, perhaps by randomly sampling from the top 100 listed in the iTunes App Store. The “observations” will be “iPhone Apps,” and the “names” will be the names of the samples (e.g., “2048,” “Angry Birds,” etc.) Next, choose 3 variables to measure. Here are some possible variables: average star rating, number of reviews, number of downloads, price, release year, type of application, number of Google hits, etc.

For some good sources on the web, visit Professor Jeffrey Simonoff’s website `http://www.stern.nyu.edu/~jsimonof/classes/datalink.html` . For financial data you can visit the website of the Federal Reserve Bank of St. Louis, `http://research.stlouisfed.org/fred/`.

For the project, you will designate one of your variables as the “response” variable; the other two will be “predictor variables.” **Your response variable must be quantitative.** Also, **at least one of your predictor variables must be quantitative.** For example, if your observations are iPhone Apps, then your response variable might be “downloads,” and your predictor variables might be “star rating” and “type of application.” In Module 3 of the project, you will try to predict the response variable using the predictor variables.

Writeup

Once you have chosen and collected your data, please answer, briefly, each of the following questions. **The writeup for this module should be approximately three pages, including the Minitab output.** Please organize your output in an easy-to-read format. The graphs should be reasonably small, but still readable.

1. Describe your data set (including the source). Where did you get the data? Why are you interested in it? What do you hope to learn? What are the observations? What are the response and predictor variables? Before exploring your data set, state some hypotheses (guesses) about how the variables should be related, perhaps based on your knowledge and experience.

2. Make a scatterplot of your response variable (on the Y-axis) versus one of the predictor variables (on the X-axis). Describe the pattern you see. Is this pattern consistent with what you expected? Note any apparent outliers in the plot, and find out which observations (and names) correspond to the outliers. Can you propose a “cause” for these outliers? Repeat the entire procedure for the other predictor variables.

3. Can you think of any other variables (not in your data set) that might be useful in predicting Y? Try to list a few possibilities.

4. For each variable, obtain Minitab’s Graphical Summary. To do this, use **Stat** ⇒ **Basic Statistics** ⇒ **Graphical Summary**. Enter the variables of interest in the dialog box. For each variable, the graph gives, first, a histogram with a ”normal curve” superimposed. (You don’t need to worry about this normal curve for now. The curve can be removed by clicking on it and then hitting the Delete key.) The graph also gives a boxplot (on its side, corresponding to the X-axis of the histogram) as well as other numerical and graphical information.

   Note any points which are outliers (or at least the two or three most extreme ones), according to the boxplots. Do these correspond to outliers you found in the scatterplots?

5. Often, the variability of a quantity depends on its size. For example, the variation in the incomes of the top 10% of earners is much greater than in the bottom 10% of earners.

   If one of your variables suffers from this size-dependent variability:

   (a) The histogram will show a right-skewed distribution.

   (b) The mean will be larger than the median.

   (c) The boxplots will show that the median line is towards the low side (in this case, left side) of the box.

   (d) The boxplot will show more outliers on the high side than on the low side.

   For each variable, based on the descriptive statistics output, decide if your response variable has the problem described above. If so, and if all of the data values for this variable are positive, try taking logs of the variable. To do this, use **Calc** ⇒ **Calculator**. If, for example, you want to create a variable, Log10Price, from the existing variable Price, type Log10Price in the box marked "Store result in variable:", and type "LOGTEN('Price')" in the box marked "Expression:" Then create the descriptive statistics graph again for the log of the variable, and decide whether the problem is reduced, according to the criteria (a)-(d) above.

   Please note that if a variable, say X, has negative values, then taking logs is NOT appropriate, so there is no point in trying it in this case. (Minitab will simply generate an error message). If a variable is non-negative, but takes a few zero values, try using LOGTEN(X + c) for a small value of c (e.g, one half of the smallest nonzero value that X takes).

   The reason we worry so much about taking logs is that it often helps the subsequent statistical analysis. In particular, taking logs tends to bring the high outliers more in line with the rest of the data, while at the same time "blowing up" the picture at the low end, so that these points can now be seen more clearly.

6. Rerun the scatterplots (and answer the rest of question 2) using the logged variables wherever this was found appropriate in question 5). Here are some examples of what I mean: If you decided to take logs of predictor variable X2 only, then you should run a scatterplot of your response variable (let's call it Y) against $\log_{10}$(X2). If you decided to take logs of X2 and X3, then you should run scatterplots of Y versus $\log_{10}$(X2) and Y versus $\log_{10}$(X3). If you decided to take logs of Y only, then you should run scatterplots of $\log_{10}$(Y) versus all of the (non-logged) predictor variables. If you decided not to take the log of any of the variables, you do not need to do anything. For each scatterplot you create here), compare it with the corresponding one from question 2). Did taking logs help you to uncover a relationship between the variables?

Points to Remember

Pay particular attention to the following points:

- One person in your group must email the TA with your group members' names before 11:59PM on Friday, Sep. 12.
- You must email me the data and a scatterplot of your data (as described in question 2 below) before 11:59PM on Sunday, Sep. 21.
- Your writeup is due by the end of office hours on Monday, Sep. 29.
- The dataset must have at least 30 rows, storing the observations, and at least 4 columns, storing the names of the observations and the variables.
- You must choose a dataset that you are actually interested in.
- All of your data must be from the same year, and your data must be no more than 5 years old.
- Pick at least 3 variables.
- Your response variable must be quantitative.
- At least one of your predictor variables must be quantitative.
- The writeup for this module should be approximately three pages, including the Minitab output.

You will lose points or be forced to re-do your assignment if you fail to meet these guidelines.