STAT 540, Fall 2015

Due on **12:00 A.M., Friday, September 18th**

Homework 3

Homework format for all STAT 540 homework this term: Please label all problems clearly and turn in an organized homework assignment. You don't need to spend hours producing beautifully typeset homework, but you won't get credit if we can't find or read your answer. Unless noted otherwise, turn in the following (as appropriate for the problem).

- Theoretical derivation (when asked for).
- Numerical results **with an explanation of your solution**, written in complete sentences. If computer code is absolutely necessary to provide context here, then include it–nicely formatted–within the solution (otherwise, see below).
- Appropriate graphics. Use informative labels, including titles and axis labels. Try to put multiple plots on the page by using, for example, the R command `par(mfrow=c(2,2))`.
- **Only as necessary**: Final clean computer code used to answer the problem **attached to the end of your homework**. Only include the rare code excerpts without which we wouldn't be able to figure out what you did. Annotate your code. Number and order the code in order of the problems. When in doubt, leave it out; consider that we will probably never read it.
- Some problems will be relatively open-ended, such as "Here are some data. Analyze them and write a report." I will provide further instructions about reports later. They should be self-contained, with suitable EDA, graphs, numerical results, and **scientific interpretation**. No computer code should be included. The report should be concise: "no longer than necessary".

(1) Show the decomposition of total variations on page 36 in lecture 3, that is show

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

(2) In a study of fuel efficiency, the fuel efficiency (measured as miles per gallon) for 13 compact car brands was measured in both city and highway conditions. The data are in the file `mpg.dat` on the class web site. You are welcome to use a computer, but please know how to do the computations (once you have r) by hand. You will have to do the computation for part (e) by hand.

(a) You are asked to describe the strength of the association between city and highway mpg. What is the most appropriate measure of association for this problem? Justify your choice.

(b) Compute the correlation coefficient of city and highway MPG for the 13 compact autos and test $H_0 : \rho = 0$.

(c) Use Fisher's z-transformation to give a 99% confidence interval for the correlation coefficient in the population of compact autos. (You can assume that 13 autos is large enough for this approach).

(d) Test the hypothesis that the population correlation is 0.90.

(e) The test in part (d) was motivated by the fact that an earlier study of 13 luxury automobiles yielded a sample correlation of 0.90. Explain why the test in part (d) is not an appropriate way to compare the two sample correlations. Propose and carry out a suitable test.

(3) Derive the equation $\sum_j h_{ij}^2 = h_{ii}$ on page 10 in lecture 47.

(4) The data in `anscombe.dat` on the class web site are classic data sets constructed by Anscombe to illustrate the need for graphical diagnostics. There are four data sets in this single data file. There are three columns: set, x, and y. The observations in data set 1 have set = 1, and so on. The issue is to decide, for each data set, whether a linear regression is a good description of the relationship between y and x.

(a) Fit a simple linear regression, predicting y using x, separately to each data set. Using only the numerical results (estimates, standard errors, tests, r^2, whatever else you might think of), is the linear regression a good description for data set 1? for data set 2? for data set 3? for data set 4?

(b) Plot y versus x, and also the residual versus predicted value for each data set. Make sure you understand how the pattern in the residual plot relates to the pattern in the observations (no answer required). After looking at the graphs, is the linear regression a good description for data set 1? for data set 2? for data set 3? for data set 4? Explain why or why not for each data set.

(5) The file `snow1.dat` on the class web site contains data from a snow gauge calibration study. A snow gauge is an instrument that measures the wetness of snow, which is crucial in western states for predicting water availability. Wet snow is denser than dry snow. Density is time consuming to measure directly; the snow gauge instrument measures a quantity called gain that depends on the density. The data at hand were collected to calibrate the instrument, that is describe how gain is related to density. When the snow gauge is used, the gain is measured and used to predict the snow density. Polyethylene blocks were used as substitute for snow. These can be manufactured in different densities. The density is set by the process used to manufacture the blocks. The data set (in `snow1.dat`) includes 9 densities. Ten blocks of each density were measured.

(a) The investigators plan to use regression to describe the relationship between gain and density. Which variable (gain or density) should be used as the X variable? Which is the Y variable? If it doesn't matter, say so. Briefly explain.

(b) Rightly or wrongly, the investigators decide to use density as the X variable and gain as the Y variable. For all this and subsequent parts of this question, please assume that the usual simple linear regression model is appropriate. Estimate the slope and intercept of the regression of gain on density.

(c) If you assume a linear relationship, is density related to gain? Report your p-value and a short conclusion.

(d) Predict the average gain when the density = 0.2 and calculate a 95% confidence interval for the average gain at density = 0.2.

(e) Calculate a 95% prediction interval for gain measurements when the density = 0.2. The 95% prediction interval includes 95% of all observations at density = 0.2.

(f) Plot the residuals vs the predicted values. Do you have any concerns?

(g) These data include multiple observations at each density. Hence it is possible to compute the mean and standard deviation for each density, then use the Box-Cox method to choose a transformation that equalizes the variances. Calculate the mean and standard deviation for each of the 9 densities, then regress log sd (Y) on the log mean (Y). What transformation (if any) does this suggest?

(h) Rightly or wrongly, the investigators decide to use a log transformation. Fit the regression of Y = log gain on X= density. (You don't have to report anything about this regression). Plot the residuals vs. predicted values. Are there any concerns?

(i) We have talked about how the standard error of the slope depends on the number of observations and spread of the X's. This problem explores some of those issues, using the snow density and gain study. For all parts here, use the regression of $Y = \log(\text{gain})$ on X = density.

  (i) The investigators are concerned about the large standard error for the slope, β_1. They plan to repeat the study using the same 9 densities. Based on the current data and previous studies, they believe 0.25 is a good estimate of σ, the standard deviation of observations among blocks of the same density. It may help to know that $\sum_i (X_i - \bar{X})^2 = 0.4557$ when that sum is calculated for 1 block of each of the 9 densities. The investigators want to reduce the standard error of the slope to 0.075. They will use n blocks at each density, i.e. the same number of blocks at each density. What is the appropriate n?

  (ii) It is expensive to buy these blocks, especially in all the different densities. And, it is hard to get some of the densities. It is easy and cheap to get blocks with densities of 0.001 and 0.686. If the investigators only used these two densities and used n of each, what is the appropriate n so that the s.e. of the slope is 0.075?

(iii) Do you have any concerns about the design in part (ii)? Would you ever recommend the design with two densities in part (ii)? Would you ever recommend the design with 9 densities in part (i)?

(6) Textbook problems:

(a) Problem 3.4: Copier maintenance. Parts (a)-(f) only. In (a), replace a dot plot with a histogram. In (e), prepare a normal probability plot only.