STA674: Regression Analysis and Design of Experiments

Assignment #3

Submission:

You must format your assignments as a pdf. Handwritten assignments will not be accepted.

When are ready to submit your assignment, copy your R (or RStudio work) or SAS code and paste it at the end of your document. *Don't forget to add comments to help the grader follow your work.* Keep in mind this is the first assignment where "***inititiative taken with statistical software***" is considered separately for 1 point—please re-check the syllabus for clarification on earning that point. Collaboration during the process of solving the problems is not only allowed but encouraged; that said, the submissions are each expected to be an individual effort reflecting the individual's work. Identical submissions or even submissions found to be **"too close to be coincidental" will be flagged and given no credit**.

Each problem is worth 4 points—longer problem parts are worth more. Please submit it **at or before the due date** via electronic submission on Canvas.

Homework Questions

1. Use multiple regression analysis to study the variation in mercury concentration in largemouth bass living in Florida lakes. The data (bass.csv on Canvas) come from a study of 53 lakes in Florida sampled from the summer of 1990 to the spring of 1991 (Lange, Royals, and Connor 1993). During this time, samples of water were taken from the lakes and the follows factors were measured: pH, alkalinity, amount of chlorophyll from suspended plant matter, and the concentration of calcium. At the same time, fish were caught and their flesh was tested for mercury levels. The response variable is the average mercury level in the flesh of bass in each of the 53 lakes (avg_mercury). You will use this data to determine if the level of mercury in the fish can be predicted based on the water chemistry. Import the data and fit the multiple regression model, fitting all possible models.

   a. Give the R^2 and AdjustedR^2 for the best models with one, two, three, and four predictors. Comment on these results (include the variables involved.)

   b. Suppose that you want to predict the average mercury level of fish in a new lake with alkalinity 3.0, calcium 2.5, chlorophyll 2.5, and pH 6.0. The predicted value for the model including all four predictors is .545 (.0164, 1.073) [*mean* (*PI*).] The predicted value for the model including only alkalinity, calcium, and chlorophyll is .532 (.0133, 1.051). Have the predicted values and the prediction intervals changed considerably between the two models? Explain why or why not (based on the inspection of these results.)

   c. Explain how your results of a) and b) agree.

2. The data set (Canvas: body.csv) contains records of CHEST_DIAM, , CHEST_DEPTH, ANKLE_DIAM, WAIST_GIRTH, WRIST_GIRTH, WRIST_DIAM (all in cm.), AGE (years), WEIGHT (kg.), HEIGHT (cm.), and GENDER (1=male) for 108 individuals. We will be looking for the best set of variables to (parsimoniously?) model WEIGHT. Even though 6 explanatory variables only gives 2^9=512 possibilities for "all possible" regressions, we'll try to be more methodical about it.

   a. First, use forward selection to find the best model for WEIGHT. Give the model.

   b. Next, use backwards elimination to find the best model for WEIGHT. Give the model. (It *may* be the same model—noteworthy, either way.)

   c. Finally, as you did in problem 1, fit all possible models, and find the highest adjustedR^2 (no need to report the model)—check models with 5, 6, and 7 variables with the highest adjustedR^2 compared to yours (hint, hint)—is there much difference?

Test (preparation) Questions

These questions, are general to the topics of the first half of the course—they will be graded as homework questions, but preparing their answers will provide you with a ***tremendous*** level of preparedness for the first exam.

3. A temperature data set similar to that we have seen in class is analyzed in output included at the end of the test. Two (of several) variables in the data are:

   Average high January temperature (High_F).

   Latitude of the city (latitude).

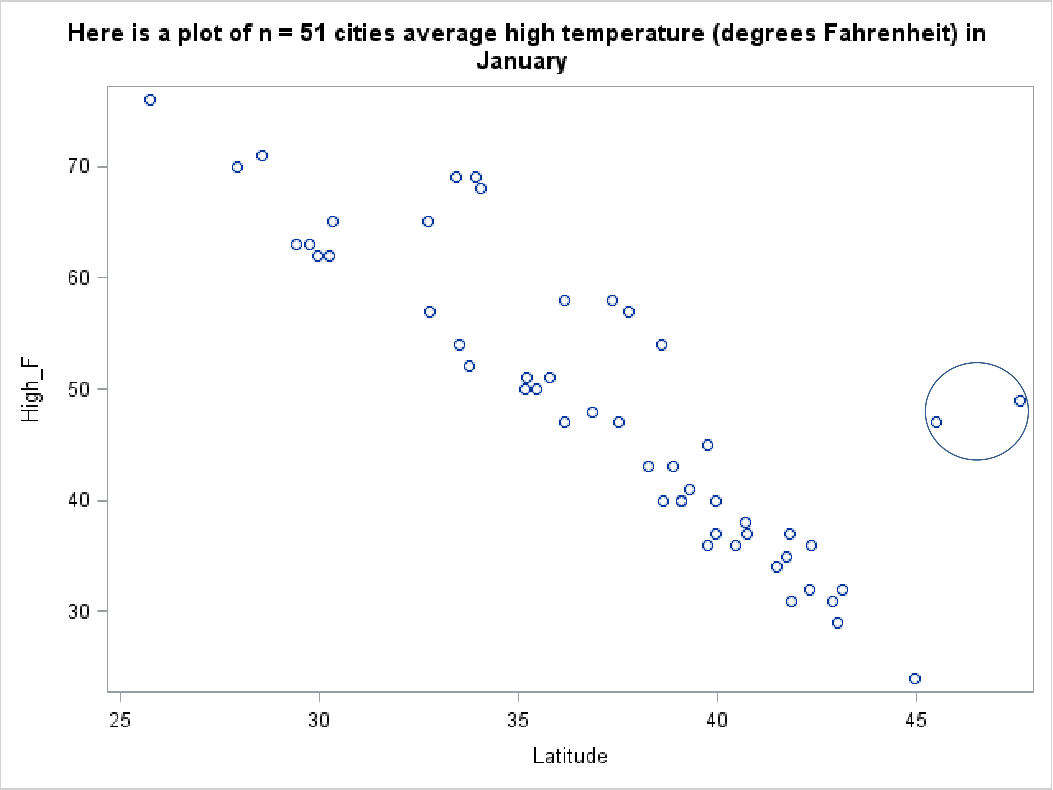
   Let x_{1i} represent the latitude of the city and y_i the average high January temperature for the i^{th} observation. The average high temperature is fitted as the response variable to the latitude of the city as the explanatory variable.

   a. Write out the complete linear model for this problem, explaining the components of the model.

   b. Fit this linear model, use software to find the least squares estimates for the coefficients, and interpret those estimates in context.

   c. Are the assumptions of the least squares analysis satisfied? Check all four assumptions, commenting (briefly) on each—whether each is satisfied or not, give justification for your conclusion. (Note: this is an instance where you may reasonably address the independence assumption.)

4. Regardless to your answer to 3.c., we are going to treat the observations of Portland, Oregon, and Seattle, Washington (circled on the plot below) as unusual (not a statistical term).



a. Using statistical terms (and statistical support from your software of choice), would we classify these two observations as influential, outliers, or both?

b. Make a new data set without those two observations and re-fit the model from problem 1; compare/contrast the results (some suggestions: Analysis of Variance table F-value, Root MSE, R^2, and, from the Parameter Estimates table, the t Value for Variable Latitude.) Some of these you would expect to increase, others to decrease—is it consistent with your expectations from a.?

5. Using the same data set, we now add an additional variable; the variables in the data now are:

Average high January temperature (High_F).

An indicator variable indicating whether the city's name starts with a letter early in the alphabet (A-M), or late in the alphabet (N-W): 0= late letters and 1= early letters (earlyind).

Latitude of the city (latitude).

Let x_{1i} represent the latitude of the city, x_{2i}, earlyind, represent whether the city's first letter comes in the first half of the alphabet, , and y_i the average high January temperature for the i^{th} observation. The model fit to the data is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

Answer the following questions:

a. Is the overall linear model useful in predicting the average January temperature? If so, how much of the variance in the average high January temperature is explained by the model?

b. For each significant variable, interpret the parameter estimate in the context of the problem.

c. Comment on the reasonability of predicting average high January temperature on a city's first letter, and to what degree this agrees or conflicts with your intuition (definitely want to look at the cities in the data set and think about the nations involved in settling our country.) Here, I am asking this along the lines of, "How reasonable is it to predict someone's height based on their middle name?"—don't look for a trick question. But the data "throw you a curve ball" (have an unexpected result.)