PSTAT 126 Final, Part I

1) Obtaining data samples for a multiple regression problem from the dataset “PSTAT 126 Final Dataset 1” (which will be sent in an email with that subject line), utilize any or all of the various regression model-fit assessment tools available in R that we have discussed in the course to help identify and validate a linear regression model that appears to best fit the given data. These will for example likely include, but may not be limited to, visual plots and/or graphs (for example, residual vs. fitted value plots), results of statistical hypothesis tests, numerical regression model accuracy measures, and linear model summary output reports, as well as any other applicable diagnostic tools that we have considered in the course. You should use the solutions to Problems 1 and 5 in Homework #3 as a general guide as to what types of information your answer could or should contain (and how to go about answering this question), but there may have been course material introduced after Homework #3 was assigned that may also be relevant here. You will likely need to try a number of different variable transformations on the response variable and/or predictor variables, using the diagnostic tools to decide which of the resulting regression models appear consistent with the data and which do not. You should **clearly** identify one or perhaps two candidate regression models that you believe are most consistent with the given data. Please try to include screenshots of the graphical plots you use as well as quote any relevant R code output results. You can or even should include the R code itself as well if you feel it helps to support your argument. Please explain your reasoning as to why the model(s) you propose may be the right one(s) in plain, natural language; you do not necessarily have to identify the “right” model to get a great deal of partial credit or even perhaps full credit.

2) In this problem work within the Simple Linear Regression (SLR) context:

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n,\ n = 1, \dots, N,$$
$$\epsilon_n \sim N(0, \sigma^2), n = 1, \dots, N.$$

(a) Show that that $\mathrm{E}[\hat{\beta}_1] = \beta_1$, where E[ ] denotes expectation. Please do not simply quote a theorem statement for this part of this problem or those below, but instead give a mathematical argument. (Also note that showing $\mathrm{E}[\hat{\beta}_m] = \beta_m$ for $m = 0$ is similar to the case $m = 1$, and you only need to include in your answer the case for $m = 1$.)

(b) Show that Var($\hat{\beta}_1$)= $\frac{\sigma^2}{S_{xx}}$, where $S_{xx} = \sum_{n=1}^{N}(x_n - \bar{x})^2$ and Var( ) denotes the variance.

(c) Show that Var($\hat{\beta}_0$)= $\sigma^2\left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)$, where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$.

(d) Are we able to conclude, directly from parts (a) and (b), that $\hat{\beta}_1$ is normally-distributed with mean $\beta_1$and variance $\frac{\sigma^2}{S_{xx}}$? Why or why not?

3) Agents at a call center get a score of 1 if a caller was satisfied with a particular call and a score of 0 if not. The company wants to see if it can accurately predict, including generating a probability estimate for the prediction, whether customers will be satisfied with a call based on relevant predictors involved such as, for example, length of the call, number of months of experience of the agent, time of day that the caller calls, etc. What is a natural regression method to use to build such a predictive model? Please first describe in detail (without any R or other software code) how you would algorithmically/mathematically set up a regression-based model to solve this problem, including how you could generate probability value estimates. You can assume there are M predictor variables. Then describe how you could set up and numerically solve such a problem in practice using R. For this part, do include the R code. You can use the built-in mtcars dataset, which does include 0/1-valued variables, as a stand-in dataset for this part of the problem. In your proposed model, use the “vs” variable -- a 0/1-valued variable -- as the response variable to serve as a stand-in for the score of a call-center call. Taking M=2, you should use “wt” and “disp” as the stand-in predictor variables. What estimates for the intercept and the coefficients of wt and disp do you get? What probability values for 0 and 1 do you get from this model when wt = 2.8, disp = 160?

4) Obtain data samples from the dataset “PSTAT 126 Final Dataset 2” (which will be sent in an email with that subject line) with a single predictor variable. Then follow the same instructions as for Problem 1 above of this final.