Project Description
Math 342 Fall 2019

Directions: You may work independently or as part of a two-person team. Each member of a team will receive the same grade. Teams are expected to work independently of each other.

Project Description: You will be doing a multiple linear regression analysis on a real (not simulated) dataset. You will write a short paper describing your analysis and any conclusions drawn from the analysis.

Data: Each team will collect data on a quantitative response variable (y) that they believe is related to *three or more* independent variables. Avoid categorical variables.

You may use data from any of the following sources:

- an experiment/survey in a book or journal
- an website or online database (e.g. Carnegie Mellon University's statlib database, www.gapminder.org, Kaggle, or https://catalog.data.gov/dataset).
- an experiment or survey you conduct yourself.

If you did not conduct your own experiment or survey, you *must* cite your source and you must perform your own independent analysis.

Timeline:

- Project proposal and dataset description due: November 12.
- First draft due: November 19.
- Final report due: November 26.

Project proposal and dataset: This should be a brief description of your project, which should include:

1. Each group member's name.
2. The response variable.
3. The predictor variables.
4. The source of your data or method of sample collection.
5. The sample size. I am especially interested in making sure that your sample size is big relative to the number of predictor variables (rule of thumb: $n \geq 10k$).
6. A brief statement of why you believe the response is related to the predictors.
7. A statement of one or more scientific questions you want to answer with the dataset.

Final report: This should include the following sections:

1. Abstract. A one-paragraph summary of your project: The scientific question you are answering, the type of analysis run, and the results/key findings. This is the one-paragraph version of your write-up.

2. Introduction. A short introduction that explains the scientific question(s) being addressed in your study. This should motivate why someone should care about the scientific questions you are asking.

3. Methods. This is where you explain your dataset and what you did with it. It should have the following subsections:

   (a) Data. This explains the variables of interest—what they are and any relevant details about how they were measured.

   (b) Regression model. What model did you finally run? What transforms did you do to your variables and why? If you did model selection, justify why this is your "best" model (reference to stepwise procedures, nested models, and/or best subsets may be appropriate here). You don't need to mention every model you looked at. This subsection should include the regression output and ANOVA table of your final model. Briefly describe the results (e.g. quote R^2, comment on the significance of each predictor variable).

   (c) Model Diagnostics. This should be brief. Check the assumptions underlying your regression model (i.e., residual diagnostics, normality). Also discuss any outliers, influential points, etc. If you decide there is a problem in this part of the analyses, describe the remedy you have taken. In the event that all the measures you have taken fail to fix the problem, briefly suggest some possible alternatives.

4. Conclusions. This should be a paragraph or two about the scientific question of interest. What did you learn using your regression?

5. Supplementary plots. Your report should include the following plots: (1) A matrix plot of the predictor vectors (the final predictors used in your model), (2) the 2×2 grid of diagnostic plots that R prints out for your regression.

Grading: The project will be graded based on the thoroughness of your analysis and the overall presentation of your results.

Other comments: Here are some things I thought of as I was grading the Spring 2010 projects.

- Grammar: Avoid ending sentences with prepositions. Watch your grammar and spelling!

- Avoid slangy or colloquial words and phrases. I shouldn't be able to drag out your report in 20 years and be able to tell exactly when it was written by the style of your language. In particular, the phrase "we had issues..." is *very* early 2000's.

- Avoid using technical words in a non-technical manner. Watch the words "normal," "correlated," and "significant." In particular, "correlated" means that the correlation coefficient was significantly different than zero, so if you use that word, you should also report the p-value. The word "significant" means that a test was performed and the p-value was less than 0.05. So, if you use the word significant, it should be accompanied by a p-value (and a description of the test you performed).

- Dealing with influential points: If you have an influential point, run the regression with and without the point to see *if the coefficients of your regression change*. If the regression coefficients don't change, then it doesn't matter much that the point is influential, because it is telling the same story as the other points. A common mistake is for students to check *if* R^2 *changes*. Checking R^2 doesn't tell you if that single influential point is masking what the other points are telling you (in which case you need to be strongly concerned that your regression is being dictated by a single point). Looking at R^2 just tells you whether your overall fit is as good with and without the influential point.

- In the Spring 2010 class, it turns out I wasn't thorough enough in my treatment of multicollinearity. I showed them that nearly collinear variables would make the regression coefficients hard to estimate. The result would be large p-values for those coefficients. However, if the variables are not *exactly* collinear, a large enough sample will allow you to tease apart the effects of the two variables. With enough sample, you will be able to get small standard errors (and hence can get low p-values). The bottom line: you don't need to worry about multicollinearity if you have a big enough sample (unless, of course, your variables are precisely linearly related).