Final Project Analysis and Report
STAT 334, Winter 2021

Due Date: Tuesday, March 16 by 5:00pm (note this is one day later than what is in the syllabus)

Grading: The project is worth 15% of the final course grade (equivalent to one midterm exam)

Everyone's project grade begins at a set value (typically around 80%). From that base value, your grade will go up or down based on things that you do well or poorly in your project. Doing just an average job on your project will get you the base grade. To get an above average grade, you need to do a thorough job of model fitting and you need to have good explanations and interpretations in your report. When you have finished the project, you must fill out a survey evaluating the contributions of your fellow group members. Your project grade may be adjusted based on these peer evaluations.

If there are any problems with your group working together, you should let me know immediately.

This project must be the result of only the work of you and your project partners. No outside help is allowed. You are allowed to ask questions to me by e-mail or during office hours. You cannot get help from other students, other professors, or search the internet to look for previous analyses of your data. What you turn-in must be 100% your own work.

Your project analysis is an *exploratory observational study*. The goal of the project is to produce a **final model** that "best" **explains** the relationship between your response variables and the predictor variables in your regression. Since *explanation* (and not just prediction) is the goal of the project, we will have to worry about the effects of multicollinearity. **Since we are doing an exploratory analysis, you do not need to use the Bonferroni adjustment.**

I will provide the project data in **csv** format.

The major steps of your project analysis will include:

a. Your initial model (containing all of the original predictor variables)
b. A check of assumptions, multicollinearity, and influential observations for the initial model
c. Changes to your data or model to fix any problems in the initial model
d. Variable reduction (or "How I got from my initial model to my final model")
   This step may include variable selection methods, removing variables, adding interactions, rechecking assumptions, etc., iteratively until you find a good final model.
e. A check of assumptions, multicollinearity, and influential observations for the final model
f. Discussion of any remaining (unfixable) problems with your final model (such as violated regression assumptions, collinearity, etc.).
g. Interpretation of the effects of all significant predictors in the final model.

Be sure to document all of your work in the steps above as part of your project report. If some of these steps require little or no work for your project, you must still mention them and explain why.

Variable reduction means eliminating insignificant variables from your regression model. Ideally, your final model should contain only predictor variables that are important (significant) predictors of the response. However, there are a few circumstances in exploratory studies when we should leave insignificant variables in the regression. (These are described on the next page.)

You may use the automated variable selection methods discussed in class to help you choose your final models; however, remember that these automated methods should not be used until problems with regression assumptions, multicollinearity, and influential observations have been fixed. You may need to make transformations, add interactions or indicator variables, or alter the model in other ways in order to make your model better. The model selected by an automated method is not necessarily the final model. It may need adjustments before you get to a final model.

Circumstances in which insignificant variables may appear in the final model:

- When the only reason the variable is insignificant is because of multicollinearity and the multicollinearity problem cannot be fixed or is not strong enough to justify simply throwing out one of the variables.
- When using polynomial regression if x^2 is significant, x stays in the model even if it is insignificant.
- When using interactions in regression if x_1:x_2 is significant, both x_1 and x_2 stay in no matter what.
- When using indicator variables or interactions with them, all indicator variables or interactions that are produced from the same categorical variable stay in as long as the partial F-test for all of those indicators (or interactions) that go together says that they are significant as a group.

Defining New Variables

You may define new X-variables as part of your analysis. This can be done by centering variables, using power transformations (to fix assumptions), dividing one variable by another or subtracting variables (to fix multicollinearity), or creating polynomials. Outside knowledge about the data can be used to create new variables. **Any new variables that are created should be described in your project report.** (Some projects will not require this step.)

Choosing the Final Model:

The final model should be a model that...

1. Satisfies all of the assumptions of regression
2. Has as low multicollinearity
3. Has a high R^2-adjusted (or low C_p, AIC, SBC) among models that satisfy 1 and 2.
4. If two models are similar on conditions 1, 2, and 3 then the best model should be the one that is the easiest to interpret (fewest variables and transformations).

Before deciding that a model is THE final model, you should always calculate a partial F-test that compares the final (reduced) model to a full model containing all of the starting variables—including the appropriate transformations and any new interactions or polynomials that you have added. Remember that the reduced model must be a subset of the variables in the full model. To make the full model, start with your final model (including any interactions, polynomials, and transformations), then add back in any predictors that were removed during variable reduction. If this partial F-test shows that the reduced (final) model is better than the full model, and the other conditions above are met, then you have your final model.

Influential Observations:

You should check for influential observations at least twice during your analysis: once in your preliminary model and again in your final model.

If you detect influential observation(s) in the preliminary model, you should identify them by name (or ID number if there is no name) and report two sets of results for your preliminary model: one with the influential observation(s) and one without. Your report should describe which coefficients were affected by the observation(s) and whether the influence affects the significance of any variables. The `compareCoefs` command in the `car` package is good for this. After doing this, you will have to proceed with the rest of your project using one model or the other. You may choose to proceed with the model that omits the influential observation(s) if you restrict the scope of your model and discuss this in your report. In discussing your final model, you must also mention that the scope of your model is restricted to a smaller set of possible X-values and describe the limits of model.

If you have influential observation(s) in your final model, you should identify them by name (or row number if there is no name) and report two sets of results: one with the influential observation(s) and one without. In presenting the model without the influential observation(s), you should tell how the scope of the model is limited because you have removed those observations.

Whenever possible in the project, demonstrate things that we have learned in class. Your project grade will be based on what you try to do not just on what you get right or wrong. You can do everything right and still get a lower project grade than another group that does a more in-depth analysis with better interpretations even if they make a few mistakes.

Project Report Deliverables

When submitting your project, upload the following **3** files to the project submission link in Canvas:

1. A **PDF file** of your **project report** of up to **3** pages (see Guidelines below)
2. A **separate PDF file** of your report **appendix** that contains all plots and tables used in writing your report. (The appendix could be a PDF from a knitted R Markdown file, or just a PDF of copied tables and plots from R. If you use R Markdown, you must knit the file before submitting it. It would be helpful to use Echo=TRUE when knitting the file. Delete any unnecessary output such as long listings of the data in variables before submitting the appendix.)
3. A **separate R script or R Markdown file** containing the code that used to produce the output for your report.

Do not combine these files together. They should be submitted as 3 separate files.
Your project will be penalized if you fail to provide these files.

Guidelines for Writing the Project Report

The report to be handed in is a typed, single-spaced report of the regression analysis of your project data. The body of this report should be no longer than **3** pages of text (**single-spaced in a 12-point font**). Please stay within the pages limits. (If you must go a few lines over this limit, but please do not go far over the limit. At some point, I will just stop reading and give you credit only for what is within the page limit.)

A summary of your analysis and any tables and plots should be placed in an appendix after the body of the report. There is no limit to the size of the appendix, only include output from models that you use in your report. (For example, you may try many transformations to fix the regression assumptions, but only include the one that does the best job in the appendix. **Your grade on the project depends entirely on your written report. Any analysis that you did that is not mentioned in your report does not exist.**

The **project report (maximum 3 pages)** should include the following:

1. A brief description of the your analysis including:
   - all of the problems in your initial model
   - things you did to fix these problems
   - how you did your variable reduction (including any tests that you ran to help with this)
   - what interactions you tried to add to your model and whether any were statistically significant

2. Interpretations of your final model in the context of the data. This should include:
   - interpretations of the coefficients of all significant predictors
     (If your response has been transformed using a logarithm, make your interpretations in terms of the original scale rather than the log scale. If your response has been transformed using something other than a logarithm, you will have to make the interpretation using the units of your transformed variable. You also have to correctly refer to any transformed predictor variables in the interpretations of their slope coefficients.)
   - interpretation of the intercept of the mode (if you can)
   - interpretation of the adjusted R^2 of the regression to describe the model fit
   - a discussion of any remaining problems (violated assumptions, multicollinearity, etc.) in the final model
   - a list of any observations that were removed (for example, for being influential) and what effect they had on the results

Your report should reference specific plots and tables as they are labeled in your appendix (for example: Model 1, Table 1A, Plot 1B, etc.)

Part of your grade will be based on the quality of your writing, so please make sure to check your report for spelling and grammar.

The project **appendix** should be:

- An **organized** list of the plots and tables from your analysis in the order that you did them. **The plots and tables in the appendix must be numbered and these numbers should be used to refer to them in your report.**
- The models should be numbered consecutively starting from your initial model (Model 1). **Along with the number of each model, you should include the code for the `lm` command that created the model so that it is easy to see which variables are in the model.**
- All of the plots and tables should be labeled with the model number and a letter for each plot or table. For example 1A could be the regression summary for the initial model. 1B could be the residual plots from the initial model, etc. Similar output can appear under the same heading. For example, even though confidence intervals and VIFs come in separate tables from the regression summary, you could group all of them together and call it, Table 1A. Several residual plots together could be Plot 1B. Any residual tests that you use could be grouped together as Table 1C. A partial F-test for Model 3 could be Table 3B, etc.

Your project grade will depend in large part on the organization and presentation of your results, so plan on finishing most of your analysis a few days before the deadline so that you can spend time on the last few days to write and organize your report. Any major deviations from the format described here will make it more difficult to see what you have done with your project and may hurt your grade.