Final Report

Due 05/10/2019 by beginning of class

Instructions:
You may work with others on these problems but you must turn in your own work. You may type your solutions in any way you like (LaTeX, Markdown, Office, etc...) as long as you present your work clearly and in an organized way.

Unless otherwise specified, you must hand in a printed copy of your work at the beginning of class.

The task:
The final assignment is to revise the work of your midterm using the more advanced perspective and tools that we have developed since spring break.

You will be expected to perform the following:

1. Perform diagnostic analysis of your model. You should check for our usual assumptions:
   (a) evaluate the structure of the model (e.g., nonlinearity in the relationship, residual versus fitted and partial residual plots);
   (b) evaluate correlation in the errors (this can be particularly difficult in practice, and goes beyond the scope of our class to estimate the covariance matrix, so a discussion of possible sources is sufficient);
   (c) constant variance of the error (residual versus fitted plots and hypothesis tests);
   (d) Gaussianity of the error (Q-Q plots and hypothesis tests);
   (e) identify outliers and evaluate for influential observations changing the fit of the model (studentized residuals, Cook's distance).

   Not all of the above may be appropriate, so you should only include in the write-up your relevant and interesting figures. Other non-interesting results can be summarized in words or tables with the corresponding plots and code in the appendix. If a test is not performed, this choice should be justified in the context of the model with quantitative and qualitative reasoning.

2. Revise the model and perform remediation of the issues:
   (a) if reasonable to do so, handle missing data systematically with methods from class;
   (b) if reasonable to do so, select the model instead based upon an information criterion method;
   (c) if reasonable to do so, consider transformation of scale of the response with e.g., Box-Cox, shifted log, logit or Fisher's z-transformation;
   (d) if reasonable to do so, exclude outliers and re-fit the model;
   (e) if reasonable to do so, transform the explanatory variables by PCA and perform principal component regression;

(f) if reasonable to do so, perform generalized or weighted least squares (generalized least squares is particularly difficult in most cases, so unless you know the covariance structure of the errors already, don't worry about estimating it or performing GLS);

(g) if reasonable to do so, perform polynomial regression with the degree selected systematically;

(h) finally, determine with the usual diagnostics if the remediation has had an effect on the issues previously noted with using the usual assumptions.

Not all of the above will be reasonable to perform with your data. If you do not perform any one of the above, provide a short justification of why it was not reasonable or necessary. **NOTE:** some of the above remediation (like outlier exclusion) depend on the other remedial steps and should be considered simultaneously. You are encouraged to iterate on this several times, but only to summarize your process on how you arrived at your final version of the model, and how (or if) it has addressed the issues noted in the diagnostics.

3. Compare the goodness of fit, uncertainty, explanatory and predictive power of the new model with the model selected in the midterm. The above model evaluations should follow the same guidelines and expectations of the midterm.

4. Make conclusions based on the above comparison. Does it appear that there is a reasonable linear signal in the data? Do these methods perform adequately for actionable prediction or explanation purposes? Does it suggest that other (e.g. nonlinear) methods should be used?

What to turn in:

The report should be **no more than 10 pages, including figures**, but not including any references or appendices in this page count. **Concisely** address the points above. Describe only the most important and interesting parts of your analysis — **code snippets will not be necessary in the main text** but you should clearly describe your methodology for a general scientific audience that is familiar with these techniques. Contextualize your results with what sorts of tests you have made (e.g. for significance), what results you have left out and why these aren't included in your final analysis. Include figures and tables for the most important components of your analysis, and for explanation purposes.

Your corresponding code and work should be included in the final appendix, section 8; I reserve the right to request a copy of the original analysis. If there isn't sufficient documentation in the appendix and this cannot be provided by the student at request, the midterm will not receive any credit. Cases of plagiarism will be handled furthermore with respect to the policy on academic dishonesty.

Your proposal should be written clearly and structured as follows:

Section 1 Introduction. Discuss the data set, your opening research question and why this question is meaningful.

Section 2 Summarize the work performed in the midterm, the issues encountered and your plan to address these in this final work. Discuss your (possibly) revised research question and why this is meaningful.

Section 3 Discuss the diagnostics of the model constructed for the midterm, what assumptions may not be satisfied, and which seem to be "OK".

Section 4 Discuss your remediation steps and your process of re-selecting/ fitting/ revising the model and its variables.

Section 5 Compare the goodness of fit, uncertainty, predictive and explanatory power of the newly formed model with that in the midterm.

Section 6 Discuss what conclusions can be made based on this analysis.

Section 7 References to data sets, papers, books or other works consulted.

Section 8 An appendix including relevant code and work.

Whenever plotting:

- Your plot is clearly labeled in all axes, legends, and the plot includes a clear title.
- The plot must be sensible and easy to read.

Grading — You will be graded for each of the following items:

1. Introduction and motivation – 5 points;
2. Diagnostics – 5 points;
3. Remediation – 5 points;
4. Comparing the goodness of fit and sources uncertainty – 5 points;
5. Comparing the predictive and explanatory power of the models – 5 points;
6. Final discussion and conclusions – 5 points.

In addition, reports that fail to follow the instructions of this assignment, the structure for the proposal, or to meet standards of scientific writing will be subject to a loss of points.