Data Analysis Project 2
Due 5:00 pm Nov 21

1 Instructions

1. This exam is week-long take-home data analysis exam.
2. You are allowed to use your textbook as well as other reference books you feel you might need. You should use the statistical software R to perform your analysis.
3. **You are not allowed to consult with any person other than your professor and your teaching assistants.**
4. You are expected to comply with the CMU policy on academic integrity. Unauthorized help will result in failing the exam, and possibly more severe disciplinary action.
5. Submit two files to Canvas: one is the PDF of your report and the other is your code (either a text file or R Markdown file).
6. Clearly comment your code so that it is clear which parts of your code go with which parts of your report.
7. Do not submit Word files; they will not be graded.
8. **There are many correct approaches to data analysis. You should not think that there is a single, correct approach to analyzing these data. The important thing is to explain clearly what you are doing and why.**
9. Your report is limited to a maximum of 7 pages (including graphs and tables). Make sure your report is clear and easy to read.

2 Data and Research Problem

The data can be found at:

`http://stat.cmu.edu/~larry/=stat401/PlantData.txt`

The first row has the variable names. The variables are:

Variable Name	Description
NR	Native plant species richness
Area	area in hectares
Latitude	latitude in degrees North Lat
Elev	elevation in meters above sea level
Dist	distance from mainland in km
Soil	number of soil types
Years	years since isolation
Deglac	years since deglaciation
Human.pop	human population

The main outcome is native plant species richness, the count of the number of different plant species.

There are several research questions and goals:

(1) The investigator hypothesizes that native species richness (NR) can be predicted from Area, Latitude, Elev, Dist, Soil, Years, Deglac, Human.pop.

(2) The investigator hypothesizes that the most important predictors are Area, Elevation and Soil types.

(3) The investigator hypothesizes that better models will be obtained if a log transformation is applied to each covariate.

The study was inspired by the following article:

```
McMaster, Robert T. "Factors influencing vascular plant diversity on
22 islands off the coast of eastern North America." Journal of
Biogeography 32.3 (2005): 475-492.
```

Your are welcome to track down and read this article if you wish. However, I don't recommend it. We are using a completely different dataset. His analysis of his data might lead you towards an analysis that does not work well for our data.

3 The Report

Your report should have the following sections:

1. **Introduction**. Briefly describe the data and the research problem.
2. **EDA.** Provide graphical displays or numerical summaries for all variables. Describe your results.
3. **Modeling.** Start by building any multiple linear regression models you think are appropriate.
4. **Diagnostics and model selection.** Use diagnostics to evaluate your models. Take any actions you think are justified such as: transformations, removing outliers, removing variables etc. Explain what decisions you make and explain why you made them.
5. **Final Models.** Summarize your final models: report the parameter estimates, standard errors, confidence intervals, and p-values. Interpret the fitted models in the context of the problem. If you have several models, then compare them. Note that there are several ways to compare different regression models: (i) partial F-tests, (ii) residuals and diagnostics and (iii) cross-validation (or other measures of prediction error).
6. **Discussion.** What are your final conclusions? Mention any limitations of your analysis.

4 Suggestions for Writing a Clear Report

1. Your writing should well-organized, free of grammatical errors, and written in complete sentences.
2. All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.
3. Figures and tables should be easy to read, with informative captions, axis labels and legends.
4. Make sure your organized, commented and that it is easy for others to read and understand.