Homework 4B

STAT 242: Intermediate Statistics

The code below just loads some packages and makes it so that enough digits are printed that you won't get confused by rounding errors.

```
library(dplyr) # functions like summarize
library(ggplot2) # for making plots
library(gridExtra)
library(GGally)
library(readr)
library(car)
library(Sleuth3)
options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Problem 1: Galapagos (Adapted from Sleuth3 12.20)

Quote from book:

> The data [read in below] come from a 1973 study. (Data from M. P. Johnson and P. H. Raven, "Species Number and Endemism: The Galapagos Archipelago Revisited," *Science* 179 (1973): 893-5.) The number of species on an island is known to be related to the island's area. Of interest is what other variables are also related to the number of species, after island area is accounted for. The data set includes the following variables:

- Island: a character vector indicating the island
- Total: total number of observed species
- Native: number of native species
- Area: area (km^2)
- Elev: elevation (m)
- DistNear: distance from nearest island (km)
- DistSc: distance from Santa Cruz (km)
- AreaNear: area of nearest island (km^2)

In this analysis, our response variable is `Native`, the number of native species. You will use `Area`, `Elev`, `DistNear`, `DistSc`, and `AreaNear` as possible explanatory variables.

```
galapagos <- ex1220
head(galapagos)
```

```
##          Island Total Native  Area Elev DistNear DistSc AreaNear
## 1        Baltra    58     23 25.09  332      0.6    0.6     1.84
## 2     Bartolome    31     21  1.24  109      0.6   26.3   572.33
## 3      Caldwell     3      3  0.21  114      2.8   58.7     0.78
## 4      Champion    25      9  0.10   46      1.9   47.4     0.18
## 5       Coamano     2      1  1.05  130      1.9    1.9   903.82
## 6  Daphne Major    18     11  0.34  119      8.0    8.0     1.84
```

(a) Make a pairs plot of the data including only the variables you will use in your analysis (with the response variable last).

(b) Identify a set of transformations for all variables in the model so that the regression conditions appear to be fairly well satisfied. In doing this, consider pairs plots of the transformed data and plots of residuals vs explanatory variables in a regression model including all transformed explanatory variables. (You should have both of these types of plots.) After this step, you should feel fairly confident that any models you fit will either have approximately linear relationships among transformed variables, or know how you will handle non-linearity by adding polynomial terms in the model. You should also feel feel confident that the variance of residuals is fairly constant across values of explanatory variables. Note that `DistSc` includes some 0 values. A common trick in cases like this is to add 1 to the observed values of that variable before doing transformations (this means that things like a log transformation are an option).

(c) Check for influential observations, outliers or high leverage observations.

Note, you will finish this analysis in Homework 5A.