Homework 5A

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; continued from Homework 4B)

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) Use all subsets regression to identify a set of models with similar ability to model these data well, based on your transformed variables. If necessary, perform this step both with and without the outliers or influential observations included (you identified these point in Homework 4B).

(b) Obtain the model fits for all models you identified in part (d) as explaining the data about as well as each other, and print the model summaries.

(c) Summarize what your analysis has to say about the association of each of the explanatory variables in the data set with the response, after accounting for the explanatory variables in your models. Indicate which of your findings are consistent across the various models considered and which depend on the details of your analysis.

(d) In this part we'll think through what's going on in a model that includes only your transformed `Elev` variable and your transformed `Area` variable.

i. Fit a model that has your (potentially transformed) `Native` as the response and your (potentially transformed) `Elev` and `Area` variables as explanatory variables. Print the model summary and also use the `avPlots` function to create added variables plots for these variables.

ii. Fit a model that has (potentially transformed) `Elev` as the response and (potentially transformed) `Area` as the only explanatory variable. Add the residuals from this model to your data set with transformed variables.

iii. Fit a model that has (potentially transformed) `Native` as the response and (potentially transformed) `Area` as the only explanatory variable. Add the residuals from this model to your data set with transformed variables.

iv. Make a plot that has the residuals from part ii on the horizontal axis and the residuals from part iii on the vertical axis. Compare this plot to the added variable plot for `Elev` from part i.

v. Fit a linear model that has the residuals from part iii as the response and the residuals from part ii as the explanatory variable. Print out the model summary. Compare the coefficient estimate for the slope to the coefficient estimate for `Elev` from your model in part i.