HW10

Chapters 11 and 12

Your Name Here

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(readr)

options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Problem 1: Swiss Fertility Rates in 1888

The `swiss` data set comes with R; this description is taken from the R documentation:

> Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. … A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].
>
> `Fertility` Ig, 'common standardized fertility measure' `Agriculture` % of males involved in agriculture as occupation `Examination` % draftees receiving highest mark on army examination `Education` % education beyond primary school for draftees. `Catholic` % 'catholic' (as opposed to 'protestant'). `Infant.Mortality` live births who live less than 1 year.
>
> All variables but 'Fertility' give proportions of the population.
>
> (paraphrasing Mosteller and Tukey):
>
> Switzerland, in 1888, was entering a period known as the demographic transition; i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries.
>
> The data collected are for 47 French-speaking "provinces" at about 1888.
>
> Mosteller, F. and Tukey, J. W. (1977) Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading Mass.

Let's investigate the relationship between the `Fertility` rate (response variable) and the other socio-economic indicators, which are of interest to demographers.

```
swiss <- swiss %>%
  select(
    Agriculture,
    Examination,
    Education,
    Catholic,
    Infant.Mortality,
    Fertility
  )
head(swiss)
```

```
##              Agriculture Examination Education Catholic Infant.Mortality
## Courtelary          17.0          15        12     9.96             22.2
## Delemont            45.1           6         9    84.84             22.2
## Franches-Mnt        39.7           5         5    93.40             20.2
## Moutier             36.5          12         7    33.77             20.3
## Neuveville          43.5          17        15     5.16             20.6
## Porrentruy          35.3           9         7    90.57             26.6
##              Fertility
## Courtelary        80.2
## Delemont          83.1
## Franches-Mnt      92.5
## Moutier           85.8
## Neuveville        76.9
## Porrentruy        76.1
```

(a) Create a pairs plot of the data.

(b) Based on your pairs plot, comment on any characteristics of the data/observations that may cause problems.

(c) Use all subsets regression to identify a set of models with similar ability to model these data well.

(d) Obtain the model fits for the models you identified in part (c), and print the model summaries.

(e) Compute and display plots of the leverage, studentized residuals, and Cook's distance for each observation in the data set, based on the model from part (d) with the most explanatory variables.

(f) Fit your models from part (d) again after removing any observations that have been identified as potential outliers or high leverage observations. Print the model summaries.

(g) Summarize the results of your analysis. For each of the explanatory variables in the data set, summarize what your models have to say about the association of that variable with the response after accounting for the other explanatory variables in your models. Please address the extent to which your findings are dependent on the observations included and the explanatory variables used. (Note: we should also check residual plots and model conditions, but to save some time we will skip that in this problem.)

Problem 2: 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.

Here is the data set. I have applied transformations to all of the variables to fix issues with non-constant standard deviations; these transformations will be good enough for our purposes. Our response variables is `Native`, the (square root transformed) number of native species. In your interpretations of coefficient estimates confidence intervals, etc., throughout this problem, you don't need to keep track of the transformation that was used or the units. (I want you to focus on the statistical issues.)

```r
galapagos <- read_csv("http://www.evanlray.com/data/sleuth3/ex1220_galapagos.csv")
```

```
## Parsed with column specification:
## cols(
##   Island = col_character(),
##   Total = col_double(),
##   Native = col_double(),
##   Area = col_double(),
##   Elev = col_double(),
##   DistNear = col_double(),
##   DistSc = col_double(),
##   AreaNear = col_double()
## )
```

```r
galapagos_transformed <- galapagos %>%
  transmute(
    Area = log(Area),
    Elev = log(Elev),
    DistNear = log(DistNear),
    DistSc = log(DistSc + 1),
    AreaNear = log(AreaNear),
    Native = sqrt(Native)
  )
```

(a) Make a pairs plot of the transformed data

(b) Based on your pairs plot, which of the explanatory variables seem to be strongly associated with the number of native species?

(c) Use all subsets regression to identify a set of models with similar ability to model these data well.

(d) Obtain the model fits for the models you identified in part (c), and print the model summaries.

(e) Of course we should check the residuals plots, model conditions, and diagnostics of outliers and high leverage observations. To save some time, let's skip that for this problem (there are no major problems). 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.

(f) When I did the process in parts (c) and (d), `Elev` was not selected for inclusion in my models; this may be surprising because the scatter plots show it has a strong relationship with `Native`. Let's investigate this in more detail over the next few parts. Fit 3 models in this part: (i) a model with only `Area` as the explanatory variable; (ii) a model with only `Elev` as the explanatory variable; and (iii) a model with both `Area` and `Elev` as explanatory variables. Print the model summaries and find 95% confidence intervals for the variable coefficients based on each of the three models.

(g) Interpret the coefficient estimates for Area and Elev in each of the three models you fit in part (f). (You don't need to worry about the units/transformations, but you should otherwise take the usual care in your interpretations.)

(h) Create the added variable plots for Area and Elev, based on the model using both variables as explanatory variables.

(i) Explain how the added variable plot is created (what is on each axis?), and how the plot relates to the coefficient estimates from your model with both variables. How does this explain why `Elev` was not selected for inclusion in the models in the all subsets regression process?

(j) Interpret the confidence interval for the coefficient of Area, based on the model using Area as the only explanatory variable, and again based on the model using both Area and Elev.

(k) Find the variance inflation factor (using the function `vif` discussed on April 22) for the coefficient estimate of Area in the model using both explanatory variables.

(l) Verify that the width of the confidence interval for the coefficient of Area in the model with both explanatory variables is approximately equal to the width of the confidence interval from the model with only Area times the square root of the variance inflation factor. The width of a confidence interval is calculated as the upper confidence limit minus the lower confidence limit. (The results won't always work out this nicely – there is another term that goes into determining the width of the confidence interval that we are ignoring here. But this is good enough to provide intuition about what the VIF does, and the effect that adding a correlated explanatory variable into a model has on our uncertainty about other model parameters.)