STAT 420 Spring 2017
2GR, 2UG

Homework 9

(Due Friday, April 14, by 4:00 p.m.)

Please submit your assignment *on paper*, following the Guidelines for Homework Write-Ups and Submissions. Please include your name (with your last name underlined), and your NetID at the top of the first page.

Note: For this entire Homework 9, you are to use R / R Studio for all computations and plots. Please copy and paste your results and printouts into Word or other word processor.

1. The dataset `Longley` comes from a study of macroeconomic data over a 16 year period from 1947 to 1962.

```
> library(faraway)
> data(longley)
> ?longley
```

(a) Create a correlation matrix to show the correlation between each of the variables in the data set. Run the following code first to control the number of significant digits displayed.

```
> options(digits=3)
```

(b) Create a scatterplot matrix of all variables in the data set. Using these plots and the results of part (a), comment on any pairs of variables that might be linearly related.

(c) Fit a linear model with Employed as the response and all other variables as explanatory. Calculate the variance inflation factor (VIF) for each of the predictors in the full model. Do any of the VIF values suggest multicollinearity?

(d) Calculate the partial correlation coefficient for Population and Employed with the effects of the other predictors removed. Construct the added variable plot for Population as well. Does it seem that Population should remain in the full model?

(e) Fit a new model with Employed as the response and the predictors from the model in part (c) which were significant. Calculate the variance inflation factor (VIF) for each of the predictors. Do any of the VIFs suggest multicollinearity?

(f) Use an F-test to compare the models in parts (c) and (e). Which model is preferable and why?

2. Data set `mammals` contains the average body weight in kg (x) and the average brain weight in g (y) for 62 species of land mammals.

```
> library(MASS)
> data(mammals)
```

[Note: The data are also stored in the posted " mammals.csv "]

Researchers such as Sprent (1972) and Gould (1996) have noted that the following relationship seems to work well:

$$\text{brain weight} = \gamma_0 (\text{body weight})^{\beta_1} (\varepsilon).$$

This model asserts that brain weight is proportional to body weight raised to the β_1 power, with a multiplicative error ε. Obviously, this model can be linearized if we take the logarithm of both x and y. That is,

$$\log(\text{brain weight}) = \log(\gamma_0) + \beta_1 \log(\text{body weight}) + \log(\varepsilon).$$

(a) Plot the average brain weight (y) vs. the average body weight (x).

The log rule: if the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable by its logarithm is likely to be helpful.

(b) Since the body weights do range over more than one order of magnitude and are strictly positive, we will use $\log(\text{body weight})$ as our predictor. Use the Box-Cox method to verify that $\log(\text{brain weight})$ is a "recommended" transformation of the response variable. That is, verify that $\lambda = 0$ is among the "recommended" values of λ when considering

$$g_\lambda(y) = \beta_0 + \beta_1 \log(\text{body weight}) + \varepsilon.$$

Please include the relevant plot in your results, using an appropriate zoom onto the relevant values.

(c) Plot $\log(\text{brain weight})$ vs. $\log(\text{body weight})$. Does linear relationship seem to be appropriate here? Fit the model

$$\log(\text{brain weight}) = \beta_0 + \beta_1 \log(\text{body weight}) + \varepsilon.$$

and use it to predict the average brain weight of a Siberian tiger (average body weight 254 kg). Construct a 95% prediction interval.