STAT 420 – Homework 7

1. Plasma Data

The dataset `plasma` comes from a study on 25 healthy children to check what normal plasma levels of a polyamine in those children should look like. The explanatory variable (X) is the child's age in years, and the response variable (Y) is the plasma level. Note that $x = 0$ represents a newborn child.

a. Create a scatterplot of the two variables. Fit the least-squares, simple linear regression model, overlay the estimated regression line onto the plot, and print the summary statistics. Comment on whether the line seems to fit the trend of the points.

b. Create the residual plot for the simple linear regression line. Comment on the model assumptions which are able to be verified in that plot.

c. Create a scatterplot of the two variables, but this time the response should be the log-transformed values of plasma level. Fit the model that belongs to this transformation, and print the summary statistics. Comment on these results in comparison to those in part a.

d. Create the residual plot for the log-transform linear regression line. Comment on the model assumptions which are able to be verified in that plot, comparing it to the residual plot in part b.

e. Generate a Box-Cox plot for plausible values of λ for a transformed response. Is $\lambda = 0$ (and hence, a log-transformation) among the recommended transformations in the Box-Cox plot? Explain your answer.

f. What is the value of λ "most recommended" by the Box-Cox method? Create a scatterplot of the recommended transformed response and age. Fit the model that belongs to this transformation, and print the summary statistics. Comment on these results in comparison to those in part c.

g. Create the residual plot for the Box-Cox transformed regression line. Comment on the model assumptions which are able to be verified in that plot, comparing it to the residual plot in part d.

h. Check the normality assumption of the model created in part f.

2. Plasma Data

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 noteworthy items.

c. Fit a linear model with Employed as the response. Calculate the variance inflation factor for each of the predictors. Do any of the VIF values suggest multicollinearity?

d. What proportion of observed variation in Population is explained by a linear relationship with the other predictors?

e. Calculate the partial correlation coefficient for Population and Employed with the effects of the other predictors removed.

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

g. Use an F-test to compare the models in parts b and e. Which is preferable?