Homework 10

We will carry out a regression exercise using R's built-in data set **Loblolly**; the variable **age** (age of loblolly pine in years) will be the explanatory variable x, and the variable **height** (height of the loblolly pine in feet) will be the response variable y. I like this data set because it resembles a completely randomized design, as though 70 trees in a stand were each selected to be sampled exactly once at a given year for measurement. Using the commands below, study the dataset and create a scatterplot of the variables. Does the relationship of the response to the explanatory variable appear linear? Is variation in the response equivalent for each value of x?

```
#Run this line first to learn about the data set
?Loblolly
#Now generate the scatterplot
attach(Loblolly)
plot(age,height)
```

Now confirm your answers using diagnostic plots from a simple linear regression model. Does the residual plot confirm your concerns about non-linearity? Non-equal variances? Explain. I did not generate a normal plot here since we do not yet have a good model fit.

```
Loblolly_SLR=lm(height~age,data=Loblolly)
plot(Loblolly_SLR,1)
```

The problems with this data set are tricky to fix, since the standard transformations do not work so well. Using the Box-Cox method, the optimal transformation for y appears to be $y^{5/4}$. Let us study the regression of $y^{5/4}$ on x. Based on the scatterplot, what is your impression of how well the transformation worked–is the relationship now linear? Are variances equal? Do the diagnostic plots from the regression back up your opinion? Of course, the slope parameter is significant here given the strong relationship between height and age; can you interpret it?

```
#Run this statement by itself so you can inspect the plot
plot(Loblolly$age,Loblolly$height^1.25)
#Regression and diagnositcs
BCLoblolly_SLR=lm(height^1.25~age,data=Loblolly)
par(mfrow=c(1,2))
plot(BCLoblolly_SLR,c(1,2))
par(mfrow=c(1,1))
```