> library(Stat2Data)

> library(MASS)

> #Question 2

> bears <- read.table("/home/thomas/git/datascience/MSDataSci/MATH550/data/Bears.csv",header=TRUE,sep=",")

> attach(bears)

The following objects are masked from bears (pos = 3):

ChestGirth, Weight

The following objects are masked from bears (pos = 4):

ChestGirth, Weight

The following objects are masked from bears (pos = 5):

ChestGirth, Weight

> #Part A

> plot(ChestGirth, Weight, pch = 16, cex = 1.3, col = "blue", main = "Weight plotted against Chest Girth", xlab = "Chest Girth (inches)", ylab = "Weight (pounds)")

> lm(Weight ~ ChestGirth)

Call:

lm(formula = Weight ~ ChestGirth)

Coefficients:

(Intercept) ChestGirth

-278.75 12.97

> abline(lm(Weight ~ ChestGirth))

> modelA <- lm(Weight ~ ChestGirth)

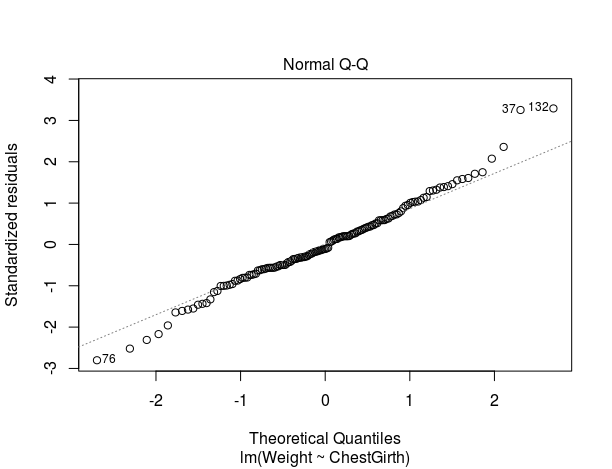


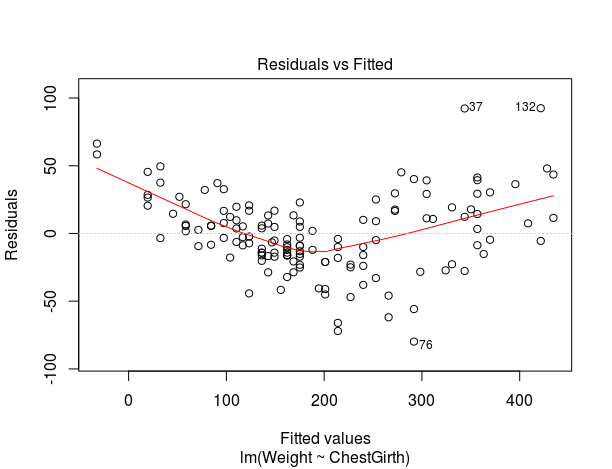
> plot(modelA)

Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:



**The results do not seem entirely adequate. It is easy in the data and in the residuals plot that a line does not pass through the points evenly. It seems to over predict the middle of the data and under predict the extremes. This organization in the residuals alludes to a different model.**

> #Part B

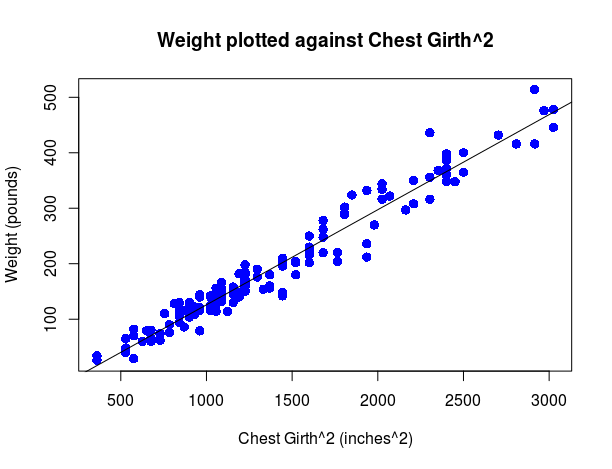
> ChestGirth2 <- ChestGirth^2

> plot(ChestGirth2, Weight, pch = 16, cex = 1.3, col = "blue", main = "Weight plotted against Chest Girth^2", xlab = "Chest Girth^2 (inches^2)", ylab = "Weight (pounds)")

> lm(Weight ~ ChestGirth2)

Call:

lm(formula = Weight ~ ChestGirth2)



Coefficients:

(Intercept) ChestGirth2

-45.4369 0.1714

> abline(lm(Weight ~ ChestGirth2))

> modelB <- lm(Weight ~ ChestGirth2)

> par(mfrow = c(2,2))

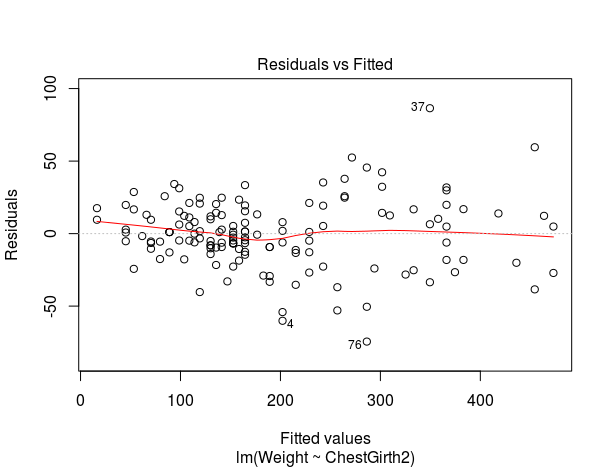
> plot(modelB)

Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:



**This model fits much better than the simple linear model before. However, it looks like there may be a problem with constant variance. Also, intuition is telling me that a quadratic may not be the best model because it is empirical. It has no physical meaning to the data.**

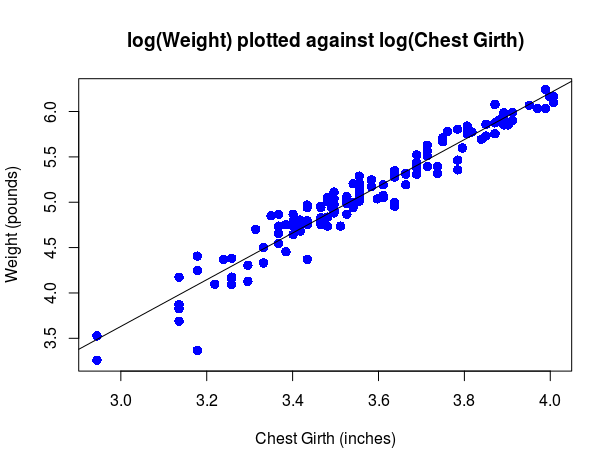
> #Part C

> logChestGirth <- log(ChestGirth)

> logWeight <- log(Weight)

> plot(logChestGirth, logWeight, pch = 16, cex = 1.3, col = "blue", main = "log(Weight) plotted against log(Chest Girth)", xlab = "Chest Girth (inches)", ylab = "Weight (pounds)")

> lm(logWeight ~ logChestGirth)

Call:

lm(formula = logWeight ~ logChestGirth)

Coefficients:

(Intercept) logChestGirth

-4.092 2.574

> abline(lm(logWeight ~ logChestGirth))

> modelC <- lm(logWeight ~ logChestGirth)

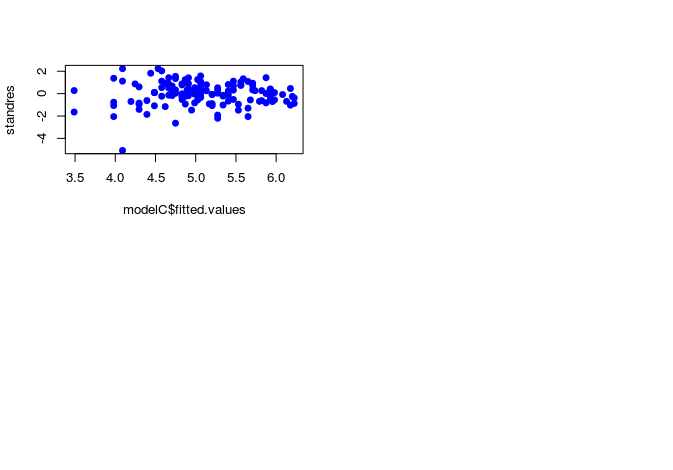
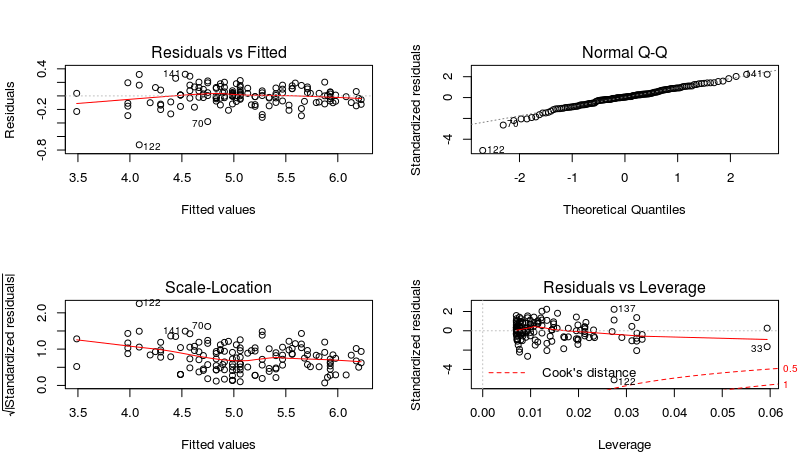
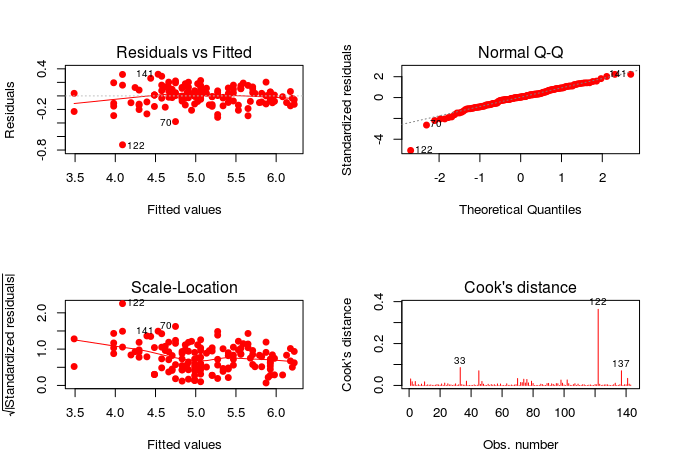
> par(mfrow = c(2,2))

> plot(modelC)

> plot(modelC,pch=19,col=10,which=c(1,2,3,4))

> standres<-stdres(modelC)

> plot(standres~modelC$fitted.values,pch=19,col=12)

**Point 122 is showing up as an unusual point. Relative to the other points, it has a large absolute standard residual. I would imagine that in the field, this could represent a malnourished bear. Perhaps one that is dying or just out of hibernation whose chest may be a normal size because of their ribcage, but low in weight.**

> #Part D

> round(confint(modelC,level=0.95),3)

2.5 % 97.5 %

(Intercept) -4.467 -3.717

logChestGirth 2.470 2.679

> results <- predict(modelC,newdata=data.frame(logChestGirth=c(log(38))),interval="prediction",level=0.95)

> exp(results)

fit lwr upr

1 194.9615 146.4104 259.6126

> #Question 3

> #Part A

> data(PalmBeach)

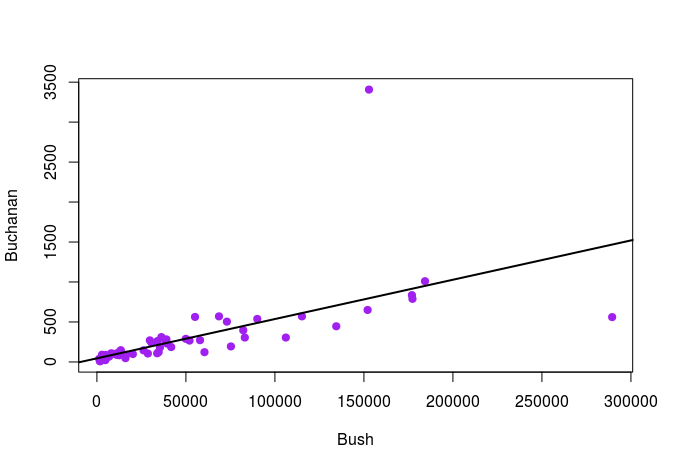
> palm<-PalmBeach

> palm.lm<-lm(Buchanan~Bush,data=palm)

> plot(Buchanan~Bush,data=palm,col="purple",pch=19)

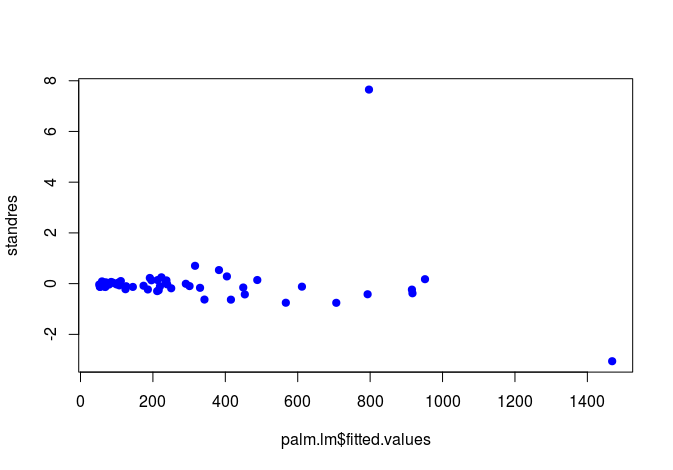
> plot(Buchanan~Bush,data=palm,col="purple",pch=19)

> abline(palm.lm,lwd=2)



> standres<-stdres(palm.lm)

> plot(standres~palm.lm$fitted.values,pch=19,col=12)

****

**The Standard Residuals plot immediately draws attention to two outliers. Palm and Dade County. The rest of the points naturally look related except for these two.**

> #Part B

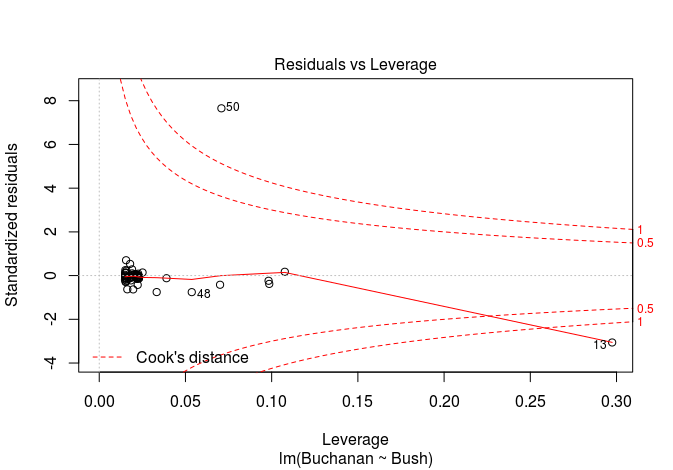
> plot(palm.lm)

Hit <Return> to see next plot:

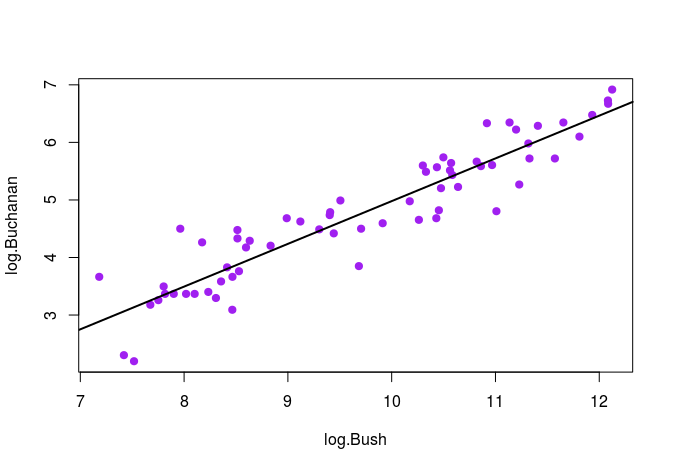
Hit <Return> to see next plot:

Hit <Return> to see next plot:

Hit <Return> to see next plot:

****

**Number 13 has a very high leverage relative to the other points. It is because this point is located by itself so much farther away from all of the other points that it has the ability to drastically pull the fitting line in one direction or another. In this case, its lower than expected y value pulls the line downwards.**



> #Part C

> palm.removed <- palm[- c(13,50),]

> log.Buchanan <- log(palm.removed$Buchanan)

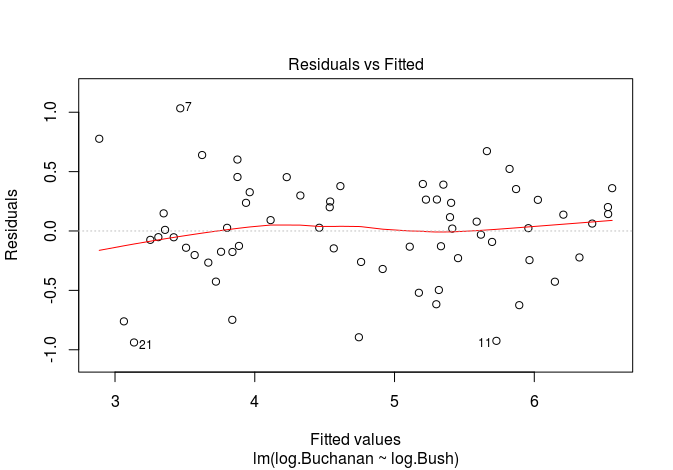
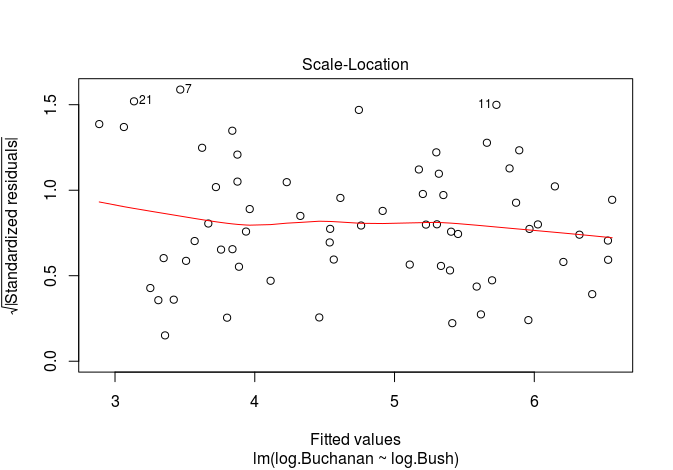
> log.Bush <- log(palm.removed$Bush)

> plot(log.Buchanan~log.Bush,col="purple",pch=19)

> palm.lm2<-lm(log.Buchanan~log.Bush)

> abline(palm.lm2,lwd=2)

> plot(palm.lm2)



**The new residual plots are looking much better with a log log transformation.**

**The residual line has flattened and the variance has calmed down a bit. Overall I would be happy with this model as compared to the previous linear model without the outliers removed.**

> #Part D

> round(confint(palm.lm2,level=0.95),3)

2.5 % 97.5 %

(Intercept) -3.170 -1.724

log.Bush 0.669 0.816

> results.50 <- predict(palm.lm2,newdata=data.frame(log.Bush=c(log(palm$Bush[50]))),interval="prediction",level=0.95)

> exp(results.50)

fit lwr upr

1 612.9178 260.1905 1443.82

> results.13 <- predict(palm.lm2,newdata=data.frame(log.Bush=c(log(palm$Bush[13]))),interval="prediction",level=0.95)

> exp(results.13)

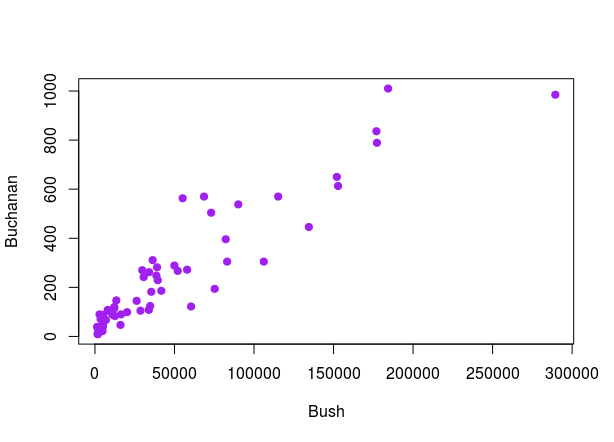
fit lwr upr

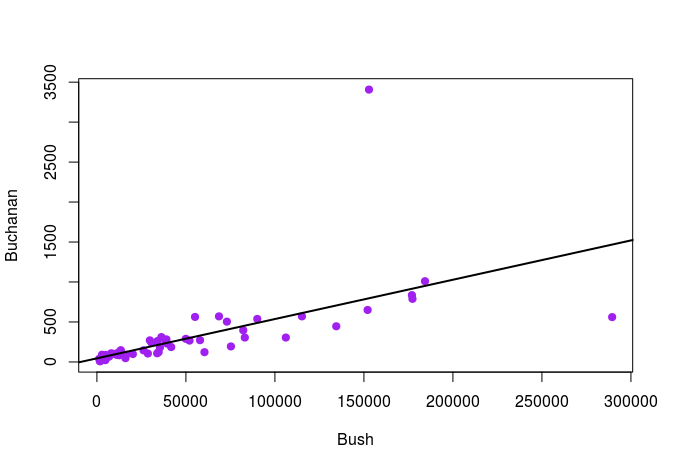
1 984.8219 413.7947 2343.853

> palm$Buchanan[50] <- exp(results.50[1])

> palm$Buchanan[13] <- exp(results.13[1])

> plot(Buchanan~Bush,data=palm,col="purple",pch=19)





Original Predicted

Now we can see that the predicted points from the model make much more sense than the outliers that we started with