Assignment1

Cameron Todd (5611230)

25 March 2017

Question 1 Import data set of from package VGAMdata and data set "xs.nz". Because we are plotting three categorical variables, we will use the mosaic plot with Education being the response variable and sexa and ethnicity being the explanatory variables

library("VGAMdata", xs.nz)

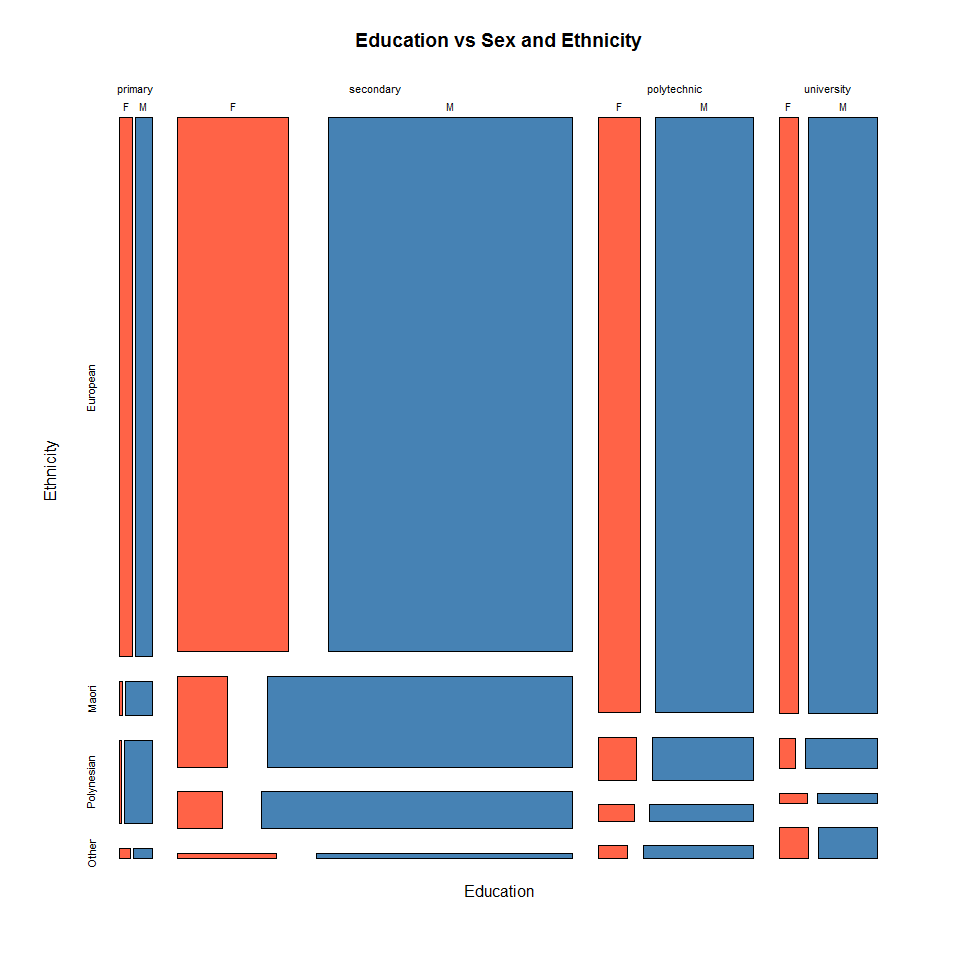
## Warning: package 'VGAMdata' was built under R version 3.3.3

eductable = table(xs.nz$educ,xs.nz$ethnicity,xs.nz$sex)  
print(eductable)

## , , = F  
##   
##   
## European Maori Polynesian Other  
## primary 171 2 3 3  
## secondary 1514 117 42 14  
## polytechnic 637 42 15 10  
## university 301 13 7 23  
##   
## , , = M  
##   
##   
## European Maori Polynesian Other  
## primary 224 23 58 5  
## secondary 3316 705 292 36  
## polytechnic 1479 111 43 37  
## university 1042 56 15 47

mosaicplot(eductable, ylab = "Ethnicity", xlab = "Education", col = c("tomato","steelblue"), main = "Education vs Sex and Ethnicity", scales=list(x=list(rot=45)))

## Warning: In mosaicplot.default(eductable, ylab = "Ethnicity", xlab = "Education",   
## col = c("tomato", "steelblue"), main = "Education vs Sex and Ethnicity",   
## scales = list(x = list(rot = 45))) :  
## extra argument 'scales' will be disregarded

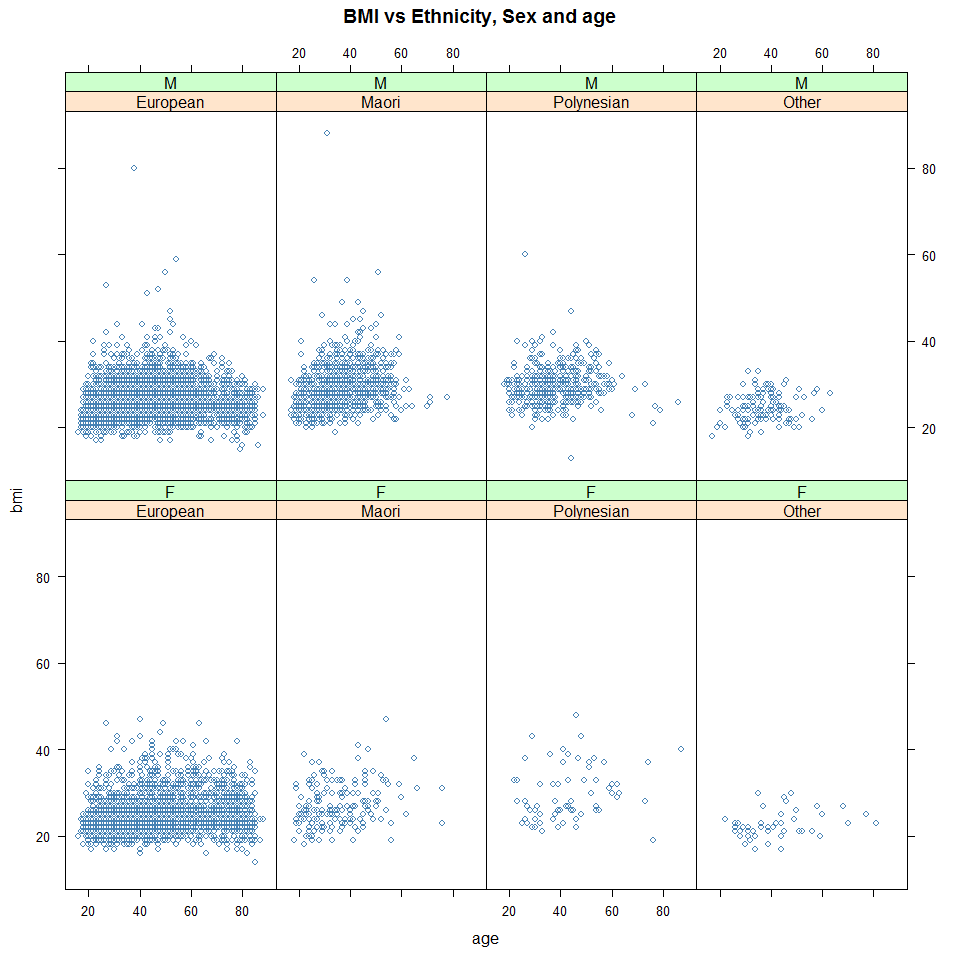
 Q1 (b) - Looking at the mosaic plot we can easily see that from the sample of people in the data, their is alot more with european ethnicity compared to Maori, Polynesian or other ethnicities accross all education categories. - Their is also a relationship between more europeans tend to go on and get a university or polytechnic qualification compared to the other cultures of Maori, and polynesia (as you can see by the height of the boxes in those columns). However you can see that more 'other' cultures tend to get university qualifications in NZ, this could be from our universities attracting the foreign students to study, whereas our NZ primary and secondary schools do not attract the same amount of 'other' culture students. - Another point to note would be that their are alot more males then females who have university qualifications or Polytechnic, however the sample does appear to have alot more males then females in total. - Their does appear to be a relationship between education and ethnicity, with more europeans going on to study further education such as university.

Q1(C)

library("lattice")

## Warning: package 'lattice' was built under R version 3.3.3

xs.nz <- transform(xs.nz, bmi = round(weight/height^2))  
xyplot(bmi ~ age | ethnicity \* sex, data = xs.nz, col = "steelblue", cex = 0.8, main = "BMI vs Ethnicity, Sex and age")

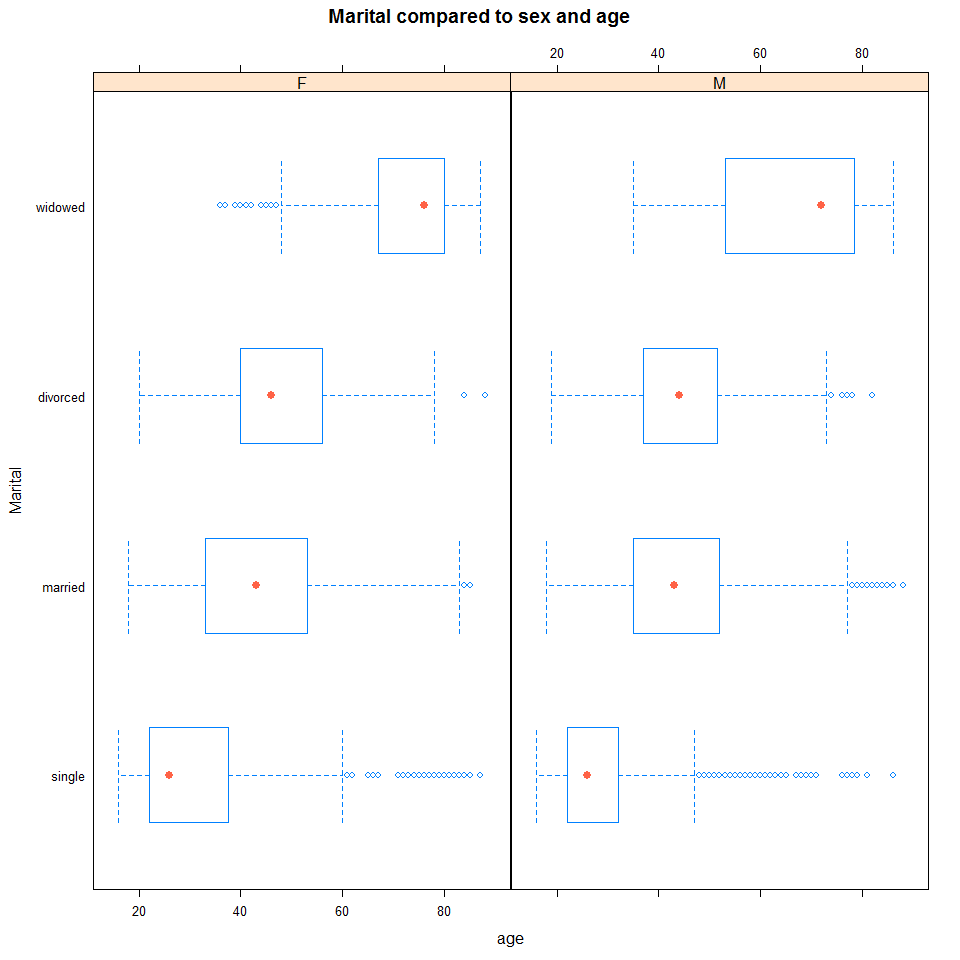


Q1 (D) - After analysing the plot their appears to be no relationship between BMI and age. However looking within the different ethnicities we can can see a weak relationship between the BMI and age of Maori male. Also looking within the different ethnicities we can see larger variances in BMI for European Maori and Polynesian males when comparing to the females from the same ethnicity. We can also see that 'other' ethnicities have a smaller average BMI then European, Maori and Polynesian, this could be different food cultures within the different ethnicity groups. Their is very little evidence here to drop the null.

Q1 (E) When comparing between the European males and european females, their appear to be no relationships. You could say their is a weak trend that as European males get older on average their BMI will decrease compared to the european females whose average BMI tend to remain fairly constant as they get older.

Q2

bwplot(marital ~ age | sex, data = xs.nz, col = "tomato",ylab = "Marital", xlab = "age", main = "Marital compared to sex and age")

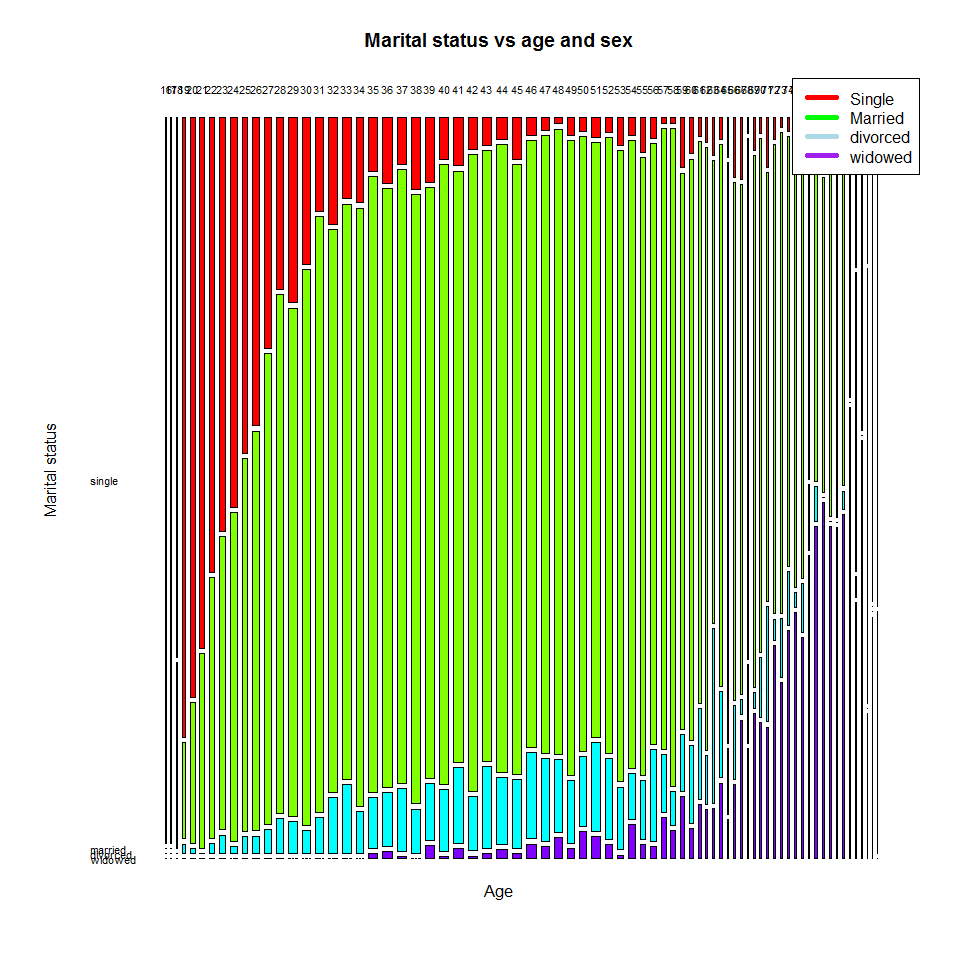
 Q2 (B) - The side by side boxplot is ok, as i can clearly see how different marital statuses have different median ages and the spread of those ages. I can see that the mjority of single people are between the ages of roughly 23 and 35 for Female and 23 and 33 for men, however the single category are both heavily skewed for males and females. Married and divorced are roughl the same average age for males and females between the ages of 35 and 55 years old. Widowed are generally 60+ for male and females with also some skewage particular in the female category. All these results are what i woudl of expected. Their appears to be a linear relationship between age and marital status, the scatters are reasonably constant.

Q2 (C) We are primarily interested in how the marital status changes as a function of age. At the moment our side by side box plot is pretty good as we can see their is a relationship between the average ages between the different groups of marital statuses. however we cannot see a clear linear function of how much they relate and the different probability densities within each group as the data appears quite skewed in the different marital categories.

Q2 (D) We would like to see more of a clear function between Marital status and age, it is a bit hard to look at the relationship without looking at the counts in each age bracket, i therefore split up the age variable into 4 categories and added a mosaic plot to understand the toal count in each of the 3 categorical variables. Sorry did not format axis correctly (ran out of time).

library(data.table)  
  
agebreaks <- c(0,25,45,65,100)  
agelabels= c("0-25","26-45","46-65","66+")  
  
setDT(xs.nz)[ , agegroups:= cut(age, breaks= agebreaks, right= FALSE, labels= agelabels)]

marital = table(xs.nz$marital,xs.nz$agegroups,xs.nz$sex)  
mosaicplot( age~ marital , data = xs.nz, ylab = "Marital status", xlab = "Age", col = rainbow(4), main = "Marital status vs age and sex", las = 1)  
legend("topright",c("Single","Married", "divorced", "widowed"), col=c("red","green","light blue","purple"),lwd=5 )

 Q3 (A)

library("gamair");data("mpg")

## Warning: package 'gamair' was built under R version 3.3.3

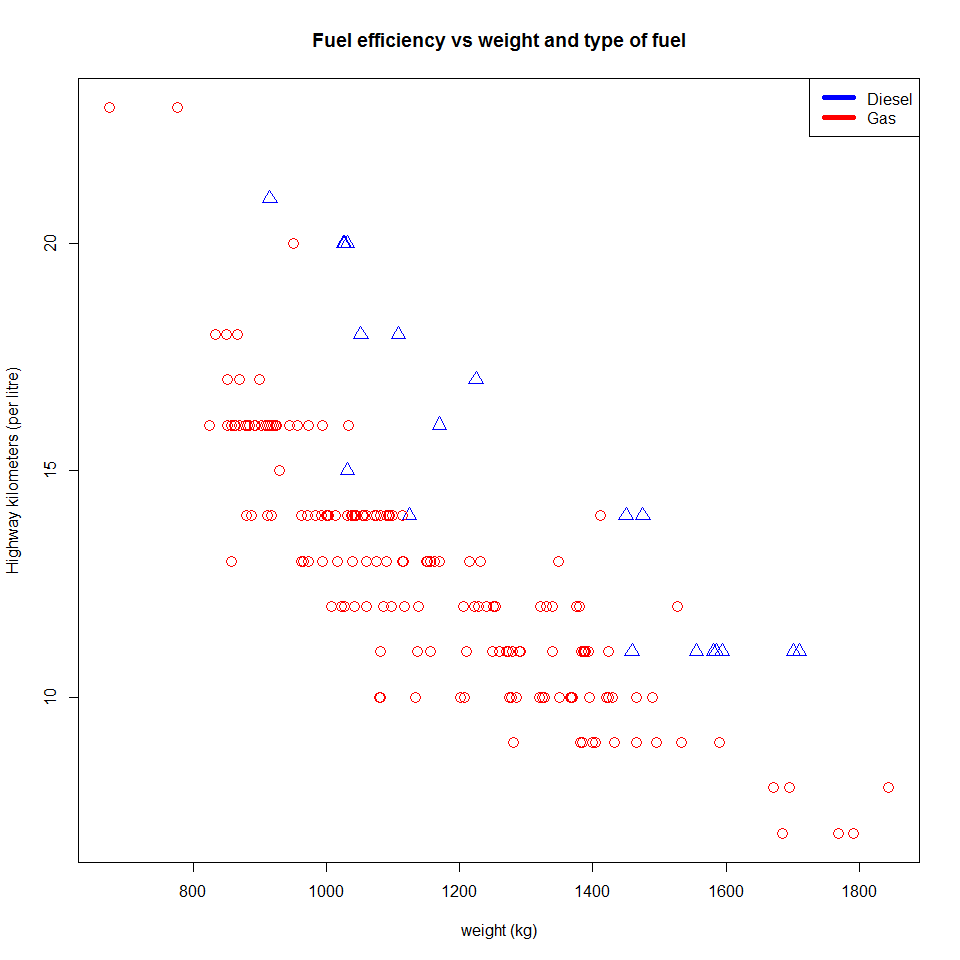
The key numbers i used to convert hw.kmpl were multiplying the miles to kilometers ratio of 1.6 and then dividing this figure by the ratio of gallons to litres of 3.79. I Also converted weight to kg, I assumed the original weight int eh data set was pounds so multiplied it by 0.45

mpg <- transform(mpg, hw.kmpl = round((hw.mpg \* 1.60934)/3.78541))  
mpg <- transform(mpg, weight.kg = round(weight \* 0.453592))  
summary(mpg[, c(27,28)])

## hw.kmpl weight.kg   
## Min. : 7.00 Min. : 675   
## 1st Qu.:11.00 1st Qu.: 973   
## Median :13.00 Median :1095   
## Mean :13.08 Mean :1159   
## 3rd Qu.:14.00 3rd Qu.:1331   
## Max. :23.00 Max. :1844

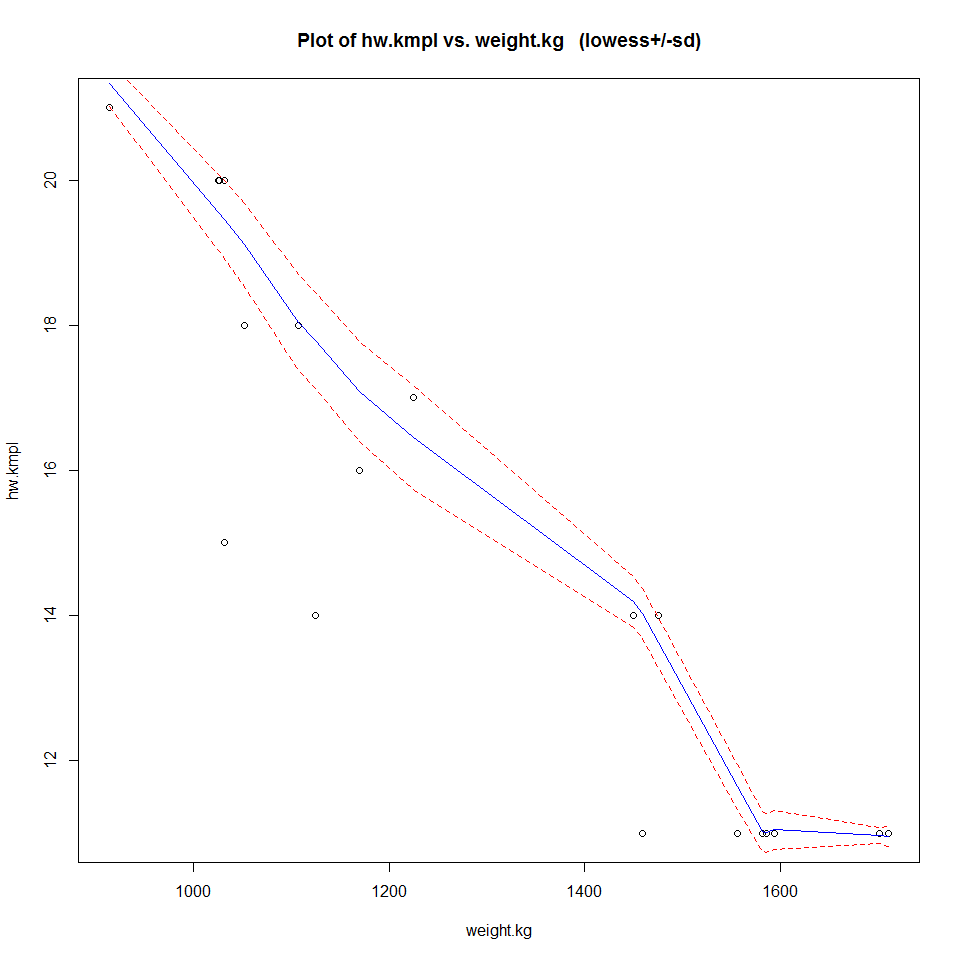
Q3 (B)

plot(hw.kmpl ~ weight.kg, data = mpg, xlab = "weight (kg)", ylab = "Highway kilometers (per litre)", cex=1.5, col= ifelse(fuel=="gas","red","blue"), pch=ifelse(fuel=="gas",1,2), main = "Fuel efficiency vs weight and type of fuel")  
legend("topright",c("Diesel","Gas"), col=c("blue","red"),lwd=5)

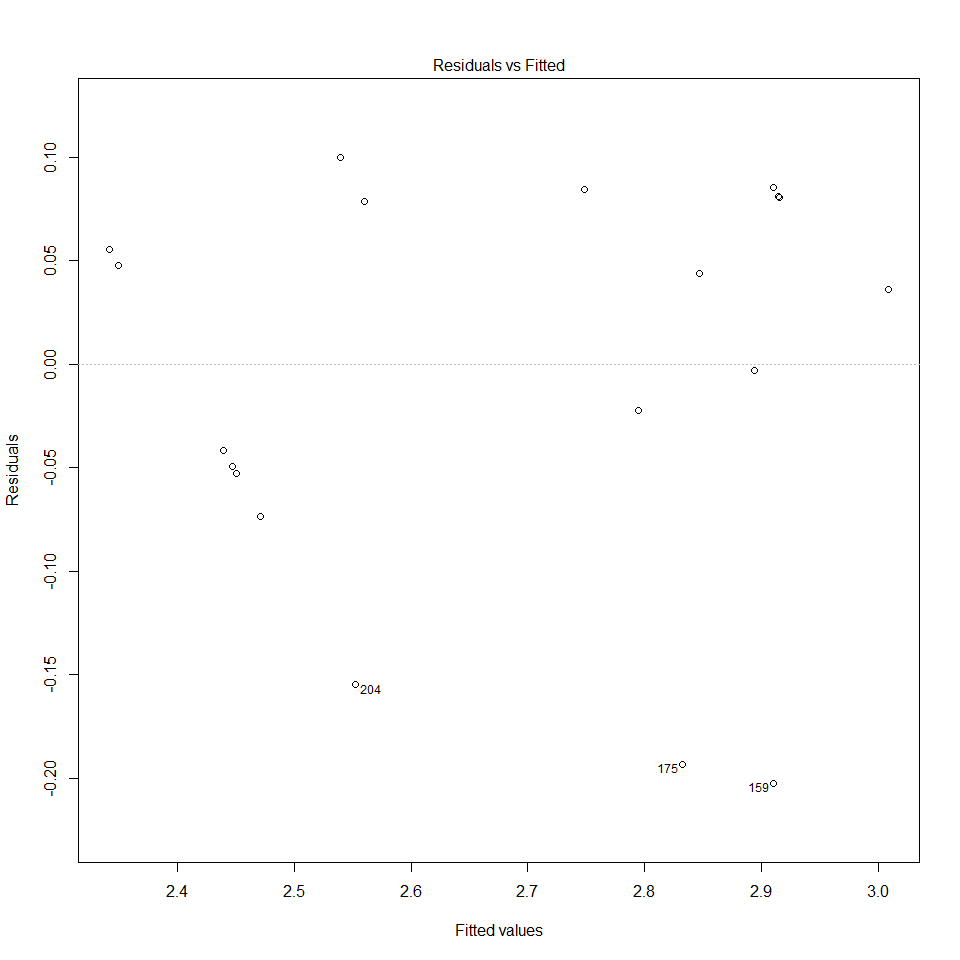
 (B) - The plot shows a negative linear relationship between the weight and the fuel efficiency of the vehicle (Highway kilometers per litre). - The plot also shows that the diesel vehicles tend to be a bit more fuel efficient compared to the gas vehicles. \_ Their could be slight curveture in the gas vehicles and may need a quadratic in the model. - Their appears to be no outliers, however sample item 19 is very light and efficent, this may be an outlier or error such as a motorbike, can explore this later. - The data appears to have reasonable constant scatter.

Q3 (C)

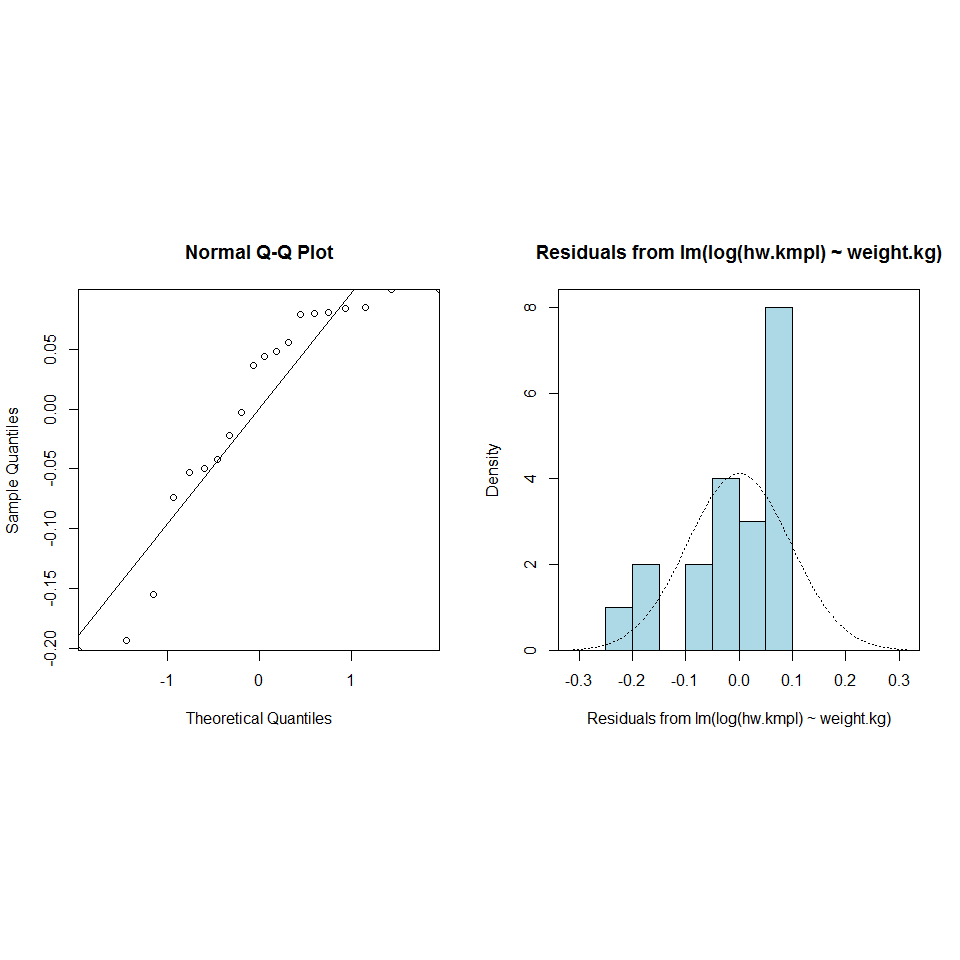
library("s20x")  
Ddata.df = subset(mpg, fuel=="diesel")  
trendscatter(hw.kmpl ~ weight.kg, data = Ddata.df)



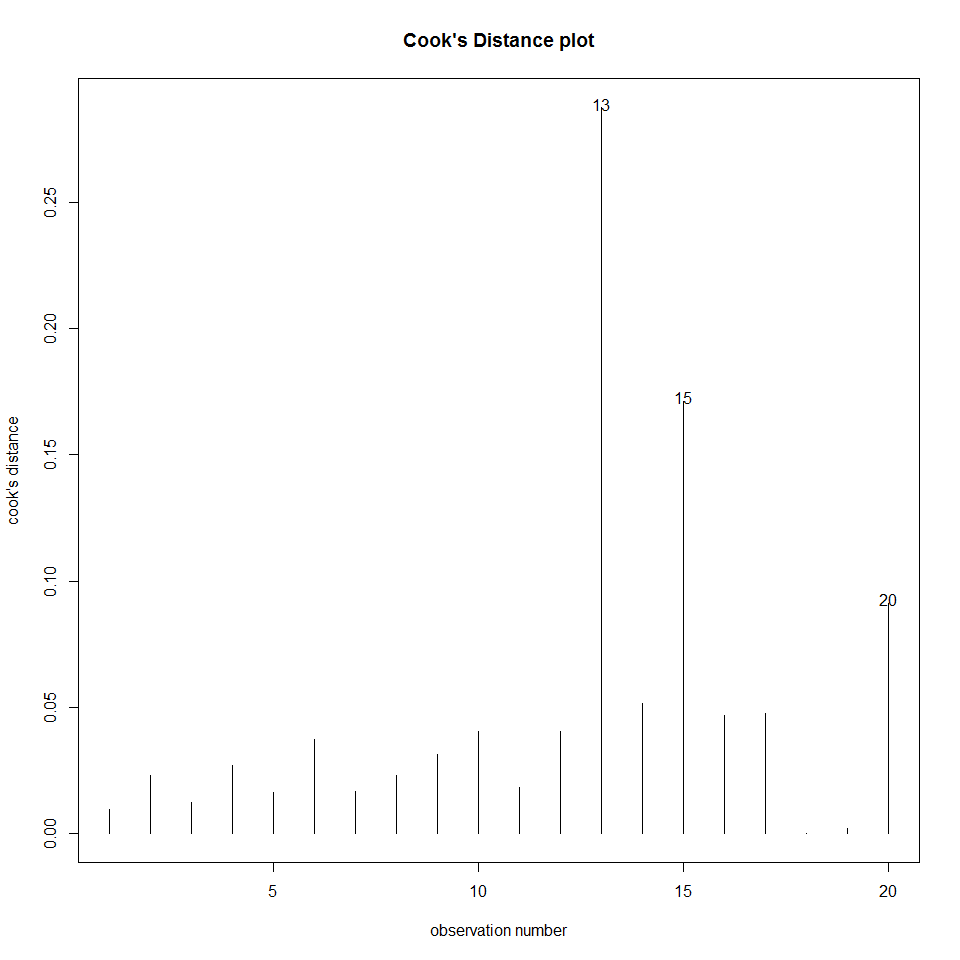
die.fit = lm(hw.kmpl ~ weight.kg, data = Ddata.df)  
dielog = lm(log(hw.kmpl) ~ weight.kg, data = Ddata.df)   
eovcheck(dielog)



normcheck(dielog)



cooks20x(dielog)



summary(dielog)

##   
## Call:  
## lm(formula = log(hw.kmpl) ~ weight.kg, data = Ddata.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.20226 -0.05013 0.03997 0.08060 0.09975   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.775e+00 1.150e-01 32.833 < 2e-16 \*\*\*  
## weight.kg -8.375e-04 8.579e-05 -9.762 1.29e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.09925 on 18 degrees of freedom  
## Multiple R-squared: 0.8411, Adjusted R-squared: 0.8323   
## F-statistic: 95.3 on 1 and 18 DF, p-value: 1.291e-08

After reviewing the trend scatter of the diesel engines plot. I could see some slight skewage in the data so i fitted a log model. All the assumptions check out, I was worried, if the data was normally distributed, however we only have a sample size of 20 and if sample size was bootstrapped, the normal distribution check would be fine. I had concerns about some of the sample items being outliers but the cooks test showed it to be all fine. The P-value (p~0) for the model shows a significant relationship between fuel efficiency and weight of the vehicle as expected. The validity of the model is pretty good with an R-squared value of .84, meaning the model explains 84% of all variation in the data. This means the model is suitable for predictions.

MAY HAVE TO REVISIT THIS AND BOOTSTRAP THE SAMPLE plot(dielog, which = 1:4)

Q3 (D)

exp(confint(dielog))

## 2.5 % 97.5 %  
## (Intercept) 34.2284104 55.484830  
## weight.kg 0.9989828 0.999343

100\*(exp(confint(dielog))-1)

## 2.5 % 97.5 %  
## (Intercept) 3322.8410368 5448.48304573  
## weight.kg -0.1017184 -0.06570272

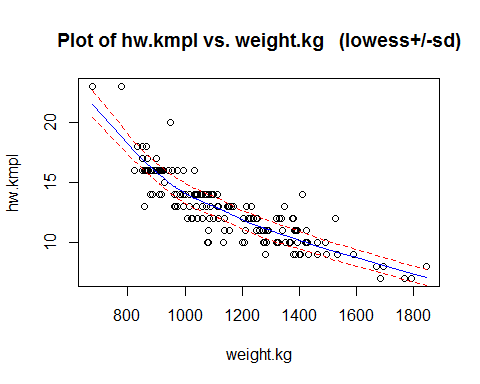
100\*(exp(confint(dielog))-1)\*23

## 2.5 % 97.5 %  
## (Intercept) 76425.343847 125315.110052  
## weight.kg -2.339523 -1.511163

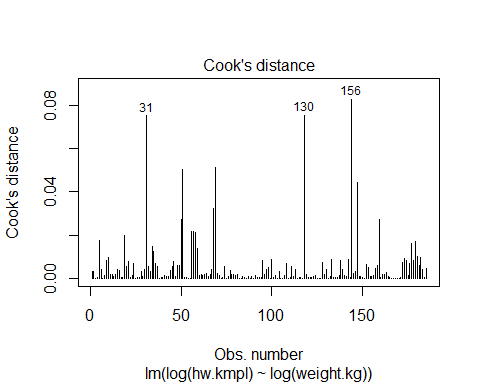
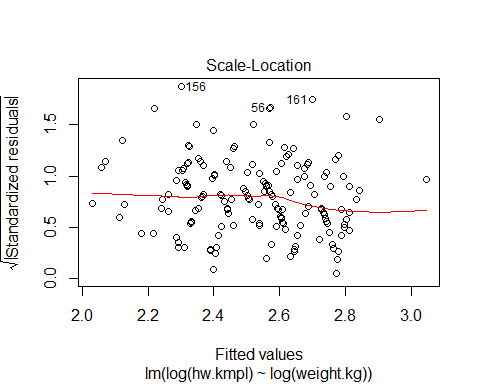
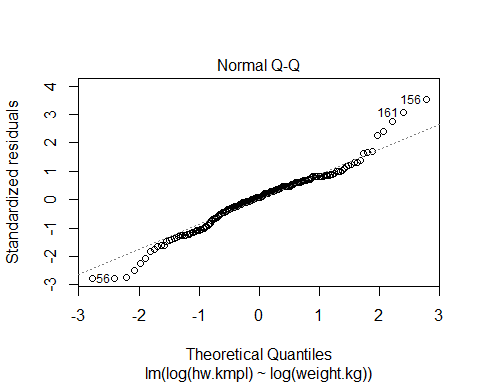
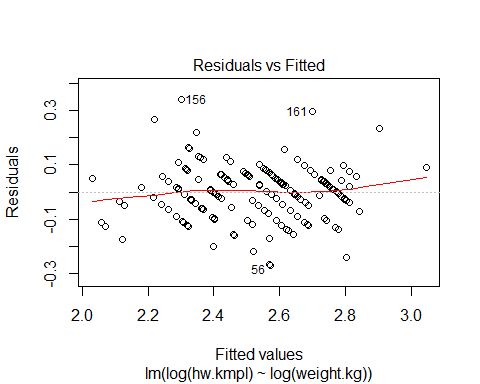
Comparing between the American driver and NZ driver using the above log model found in (c). The smoking variable stated in the question has no affect on the fuel efficiency and only the weight changes. Therefore an American driver who weighs 97kg compared to the NZ driver who weighs 64kg. Assuming the drivers are driving the same vehicle type and weight, we can expect for every 100kg added to the vehicle load, a median % decrease in fuel efficiency (km per litre on highway) of between 6.6% and 10.2%. Therefore the US driver weighing 23kg more then the NZ driver we can expect his car to between 1.5% and 2.3% less efficient then the NZ driver.

Q3 (E)

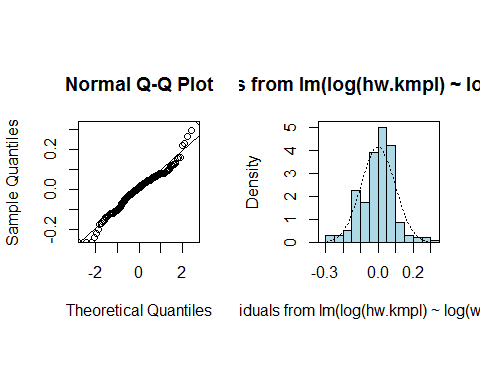
Gdata.df = subset(mpg, fuel=="gas")  
trendscatter(hw.kmpl ~ weight.kg, data = Gdata.df)



Gas.fit = lm(log(hw.kmpl) ~ log(weight.kg), data = Gdata.df)  
plot(Gas.fit, which = 1:4)



normcheck(Gas.fit)



summary(Gas.fit)

##   
## Call:  
## lm(formula = log(hw.kmpl) ~ log(weight.kg), data = Gdata.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.26813 -0.05691 0.00824 0.05782 0.33901   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.64258 0.26029 37.05 <2e-16 \*\*\*  
## log(weight.kg) -1.01248 0.03705 -27.32 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0969 on 183 degrees of freedom  
## Multiple R-squared: 0.8032, Adjusted R-squared: 0.8021   
## F-statistic: 746.7 on 1 and 183 DF, p-value: < 2.2e-16

Q3 (E) Thinking logically, it can be argued that efficiency of the vehicle (hw.kpl) will have a power law relationship with the weight of the vehicle, specifically a log log relationship. In particular, that fuel efficiency (kilometers per litre) would decrease linearly with the weight of the vehicle. After logging the variables, scatter plots showed the desired negative linear relationship between fuel efficiency and car weight and they had reasonable constant scatter. It is noted that after originally looking at the plot i had fitted a quadratic model before the loglog model but after seeing the R-squares was not as high i changed it to the log log model. However i do have concerns over the normality of the distribution, the QQ plot indicates that the differences have longer tails than normal and the plotted points are below the line for smaller and above the line for higher i.e. Not normally distributed

Q3 (F) The fitted model is:

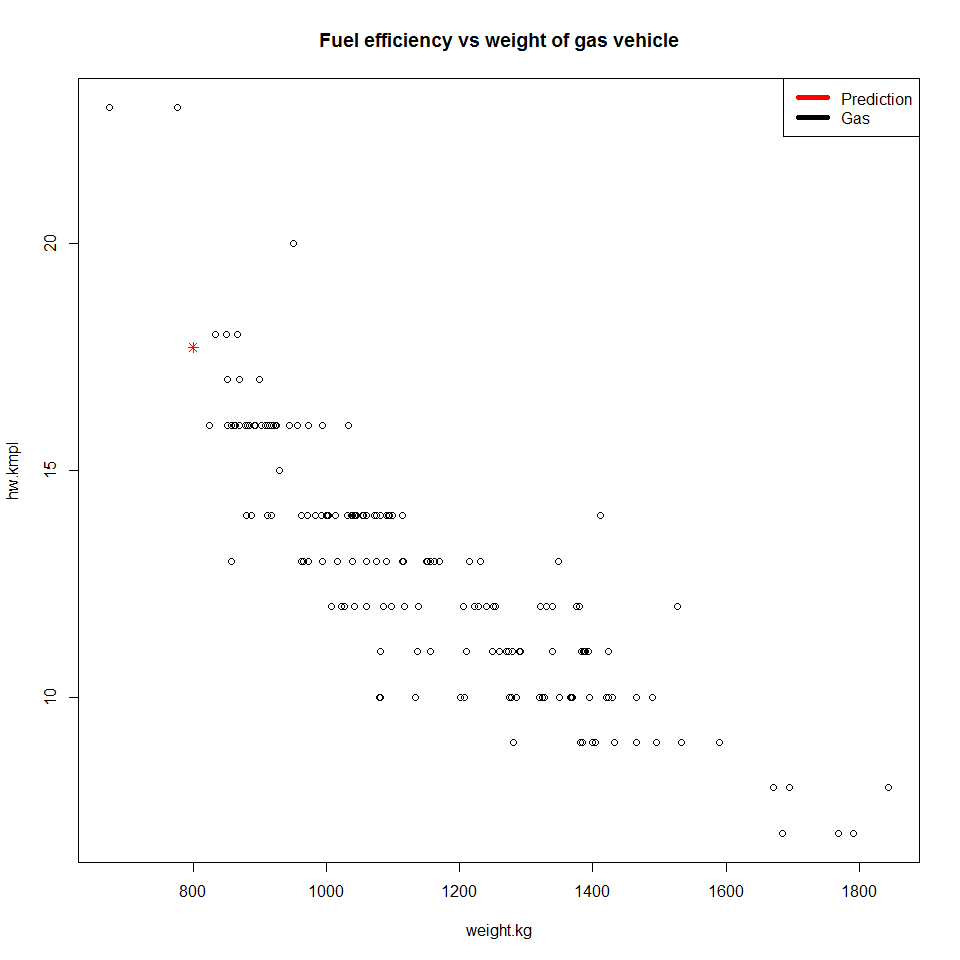
Pred.df = data.frame(weight.kg=800)  
exp(predict(Gas.fit,Pred.df,interval = "prediction"))

## fit lwr upr  
## 1 17.71771 14.60376 21.49565

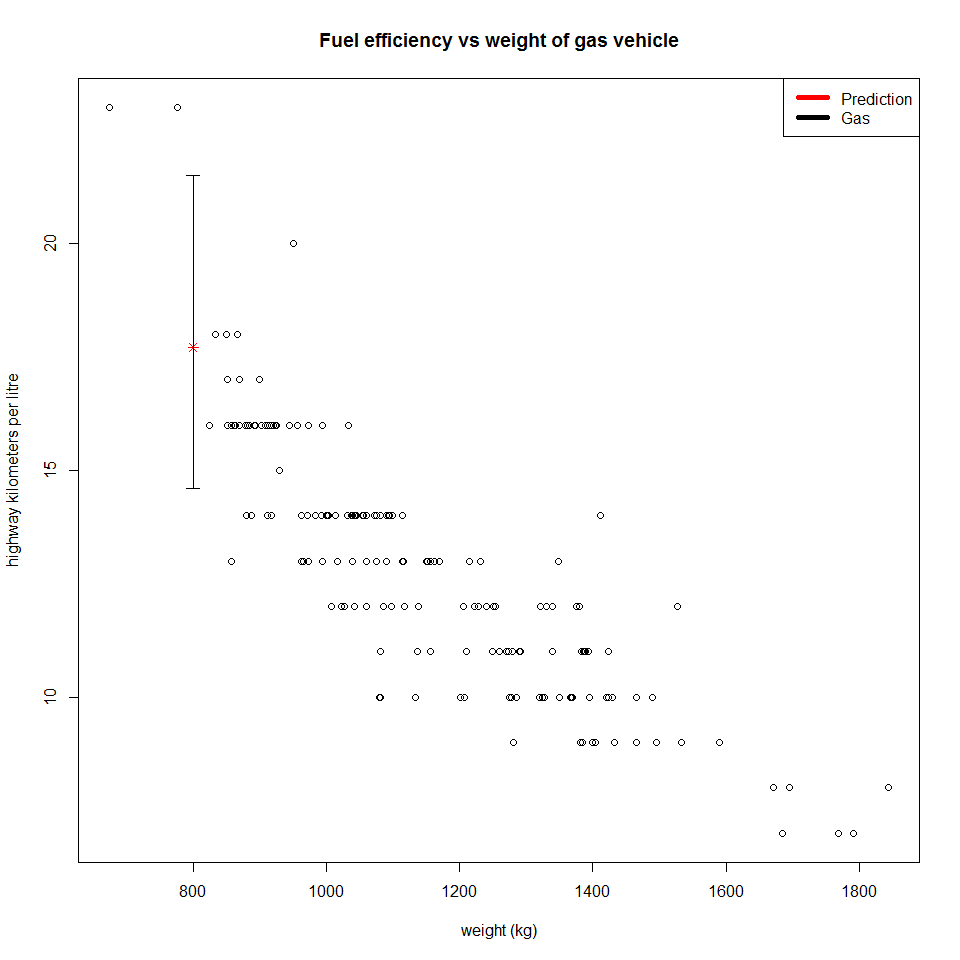
As you can see a vehicle weighing close to 800kg can have a median fuel economy of between 14.6 and 21.5 kilometers per litre.

Q3 (G)

n=nrow(Gdata.df)  
Gdatanew.df = Gdata.df[c(1:n,n),]  
Gdatanew.df[n+1,c("hw.kmpl","weight.kg")]=c(17.71,800)  
plot(hw.kmpl ~ weight.kg, data = Gdatanew.df,   
 col=ifelse(Gdatanew.df$weight.kg==800,"red","black"),  
 pch=ifelse(Gdatanew.df$weight.kg==800,8,1),cex=1, main="Fuel efficiency vs weight of gas vehicle")  
legend("topright",c("Prediction","Gas"), col=c("red","black"),lwd=5)



plot(hw.kmpl ~ weight.kg, data = Gdatanew.df, xlab = "weight (kg)", ylab = "highway kilometers per litre",  
 col=ifelse(Gdatanew.df$weight.kg==800,"red","black"),  
 pch=ifelse(Gdatanew.df$weight.kg==800,8,1),cex=1, main="Fuel efficiency vs weight of gas vehicle")  
legend("topright",c("Prediction","Gas"), col=c("red","black"),lwd=5 )  
  
segments(800,14.60376, x1 = 800, y1 = 21.49565, col = par("fg"), lty = par("lty"), lwd = par("lwd"))  
segments(790,14.60376, x1 = 810, y1 = 14.60376, col = par("fg"), lty = par("lty"), lwd = par("lwd"))  
segments(790,21.49565, x1 = 810, y1 = 21.49565, col = par("fg"), lty = par("lty"), lwd = par("lwd"))

 The preditction of an 800kg car looks reasonable when you compare to the log model in the plot, 800kg is within our data range and the R-squared of the model is high enough. I have manually coded the 95% confidence interval range at a car weight of 800kg and you can see the predicted value sits dead in the middle of this range. I think this is a reasonable prediction