Assignment7

library(UsingR)

## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

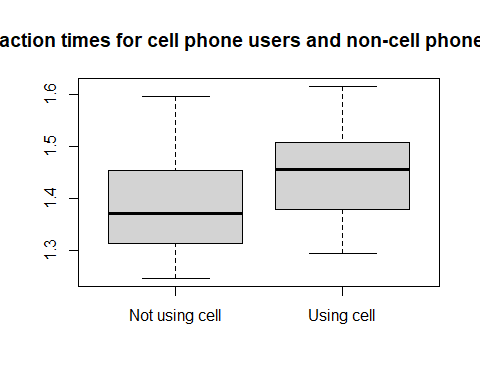
##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

##   
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':  
##   
## cancer

#3.9  
attach(reaction.time)  
nocell = reaction.time[control == 'C',]  
yescell = reaction.time[control == 'T',]  
boxplot(nocell$time, yescell$time, names=c("Not using cell", "Using cell"), main="Reaction times for cell phone users and non-cell phone users")



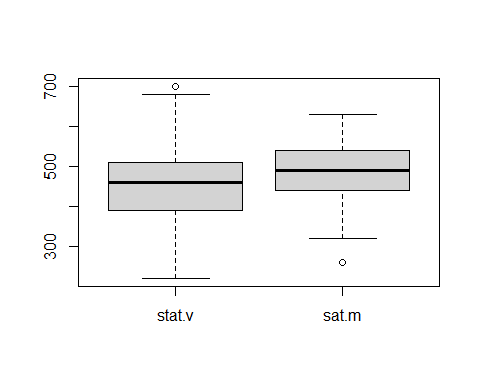
detach(reaction.time)

3.9 discussion: The median reaction times for people not using a cell phone is much less than for people who were using their cell phone (1.38 vs 1.47 sec). The maximum values are surprisingly similar (1.6 sec for non-cell users vs 1.61 sec for cell users), but the spread (1.21 sec – 1.6 sec for non-cell users vs 1.3 sec - 1.61 sec for cell users) shows that the people who did not use their phone while driving had much better reaction times.

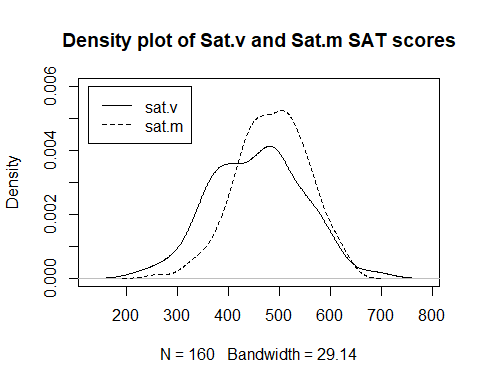
#3.11  
attach(stud.recs)  
summary(stud.recs, na.rm=T)

## seq.1 seq.2 seq.3 sat.v   
## Min. : 36.00 Min. :21.00 Min. : 16.00 Min. :220.0   
## 1st Qu.: 68.75 1st Qu.:65.00 1st Qu.: 55.75 1st Qu.:390.0   
## Median : 78.00 Median :72.00 Median : 67.00 Median :460.0   
## Mean : 77.74 Mean :71.22 Mean : 66.11 Mean :455.8   
## 3rd Qu.: 86.00 3rd Qu.:80.00 3rd Qu.: 78.00 3rd Qu.:510.0   
## Max. :100.00 Max. :99.00 Max. :100.00 Max. :700.0   
##   
## sat.m letter.grade num.grade   
## Min. :260.0 A :25 Min. :0.000   
## 1st Qu.:440.0 P :20 1st Qu.:1.000   
## Median :490.0 D :18 Median :2.700   
## Mean :485.9 C :16 Mean :2.173   
## 3rd Qu.:540.0 A- :13 3rd Qu.:3.700   
## Max. :630.0 B :13 Max. :4.000   
## (Other):55 NA's :36

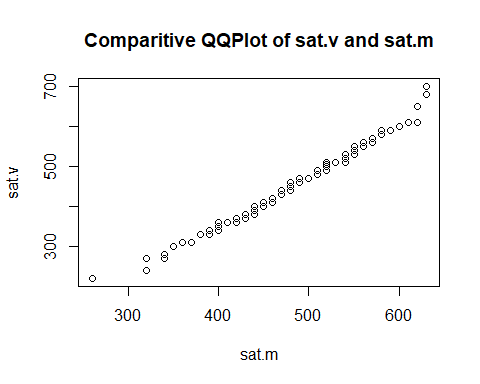
boxplot(sat.v, sat.m, names=c("stat.v", "sat.m"))



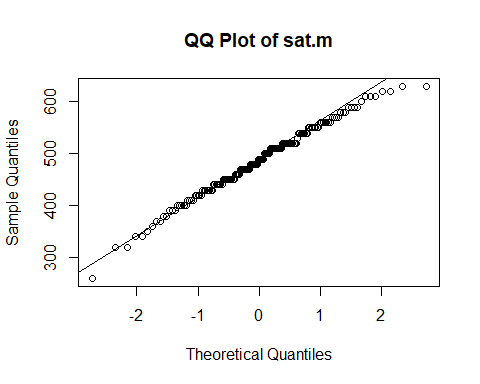
plot(density(sat.v, na.rm=T), ylim=c(0,0.006), main="Density plot of Sat.v and Sat.m SAT scores")  
lines(density(sat.m, na.rm=T), lty=2)  
legend(125, 0.006, c("sat.v", "sat.m"), lty=c(1,2))



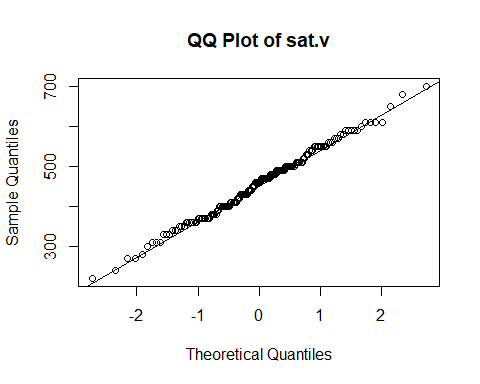
qqplot(sat.m, sat.v, main="Comparitive QQPlot of sat.v and sat.m")



qqnorm(sat.m, main="QQ Plot of sat.m")  
qqline(sat.m)



qqnorm(sat.v, main="QQ Plot of sat.v")  
qqline(sat.v)



detach(stud.recs)

3.11 discussion: The data sets have nearly the same center (sat.v = 460, sat.m=490), but the density plot for sat.v does not have as smooth of a distribution as for sat.m.

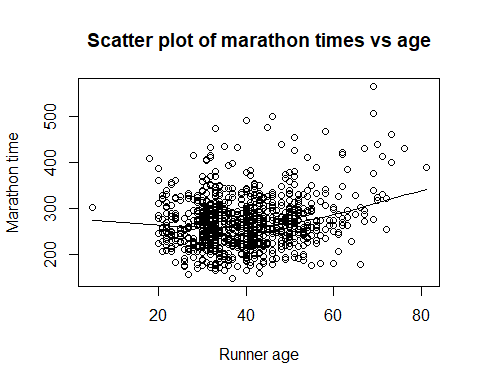
Using the qqplot, we can see that the data sets appear to have the same shape. In this plot, the observations are nearly overlapping and follow the same direction and slope.

Sat.m appears to have a slight curve beginning in the upper quantile range, so it might not come from a normal distribution. Sat.v appears to be fairly straight, so this could be from a normal distribution.

#3.18  
cor(nym.2002$age, nym.2002$time)

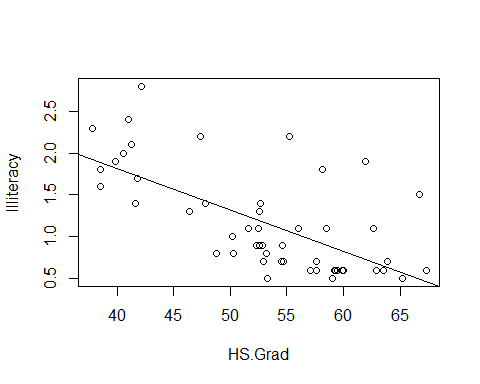
## [1] 0.1898672

scatter.smooth(nym.2002$age, nym.2002$time, ylab= "Marathon time", xlab= "Runner age", main="Scatter plot of marathon times vs age")

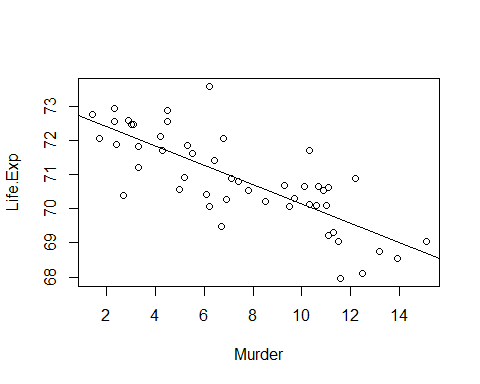


3.18 discussion: I expect that that the middle-aged people will have better times (due to experience). Higher aged people might not do very well because of general age-health issues. I was right, there is a statistical significance (p-value = 0.1898672) between age and marathon time and the scatter plot shows a higher amount of fast racers .

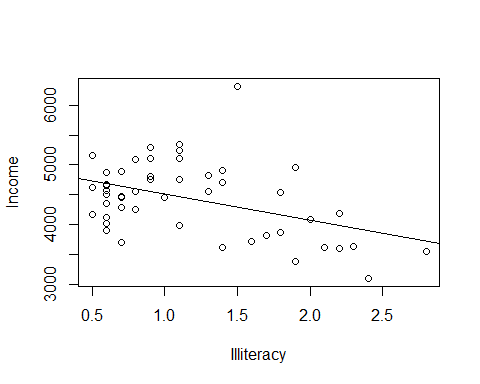
#3.24.1  
x77 <- data.frame(state.x77)  
x77.reg = lm(Illiteracy ~ HS.Grad, data=x77)  
plot(Illiteracy ~ HS.Grad, data=x77)  
abline(x77.reg)



#3.24.2  
x77.reg = lm(Life.Exp ~ Murder, data=x77)  
plot(Life.Exp ~ Murder, data=x77)  
abline(x77.reg)

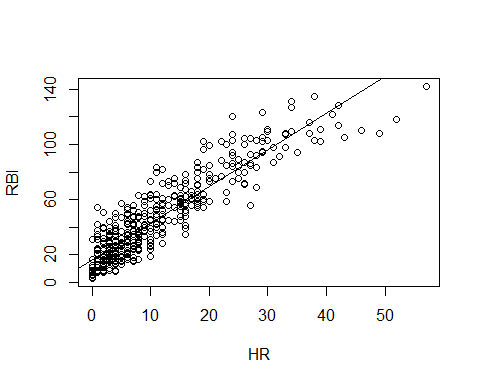


#3.24.3  
x77.reg = lm(Income ~ Illiteracy, data=x77)  
plot(Income ~ Illiteracy, data=x77)  
abline(x77.reg)



attach(batting)  
batt.reg = lm(RBI ~ HR, data=batting)  
plot(RBI ~ HR, data=batting)  
abline(batt.reg)

3.24 discussion: (1)The illiteracy rate decreases as the percentage of HS graduates increases. (2)Murder rates decrease as life expectancy increases. (3)Income decreases as the illiteracy rate increases. The relationship between life expectancy and murder rate is interesting to me – For most people, I would assume that in the process of getting older you tend to not care as much about what other people think. This could cause for things being said that generally should not be and thus creating a possible reason for a murder to take place. Are murderers having sympathy for older folks or is there some other reason?



predict(batt.reg, data.frame(HR=33)) #prediction

## 1   
## 104.1099

98-predict(batt.reg, data.frame(HR=33)) #residual

## 1   
## -6.1099

detach(batting)

#3.32  
library(MASS)  
attach(UScereal)

## The following object is masked from package:UsingR:  
##   
## fat

cereal.table = table(mfr, shelf)  
options('digits' =2)  
prop.table(cereal.table, 1)

## shelf  
## mfr 1 2 3  
## G 0.27 0.32 0.41  
## K 0.19 0.33 0.48  
## N 0.67 0.00 0.33  
## P 0.22 0.11 0.67  
## Q 0.00 0.60 0.40  
## R 0.80 0.00 0.20

prop.table(cereal.table, 2)

## shelf  
## mfr 1 2 3  
## G 0.333 0.389 0.310  
## K 0.222 0.389 0.345  
## N 0.111 0.000 0.034  
## P 0.111 0.056 0.207  
## Q 0.000 0.167 0.069  
## R 0.222 0.000 0.034

detach(UScereal)

3.24 discussion: There are some obvious differences between manufacturers. G and K seem to do this best given that their products are almost evenly distributed across each shelf. They also hold the highest percent of product on any shelf. All of the other manufacturers, except P, have at least one shelf where they do not have any product.