```
## Homework #3 Chapter 4.1  Chapter 5.1,2,3

## Wallace O'Rear

##chapter 4 #1 Professor Salaries

profSals <- read.csv("Data/ProfessorSalaries.txt", sep="")

par(mfrow=c(1,1))
plot(x=profSals$Experience, y=profSals$ThirdQuartile, xlab="Experience", ylab="Third Quartile of Salary")
abline(lsfit(y=profSals$ThirdQuartile,x=profSals$Experience))

#linear model, no weights or transforms
salaries.lm <- lm(ThirdQuartile ~ Experience, data=profSals)
par(mfrow=c(2,2))
plot(salaries.lm)
summary(salaries.lm)

# Call:
#   lm(formula = ThirdQuartile ~ Experience, data = profSals)
#
# Residuals:
#   Min     1Q Median     3Q    Max
# -14150  -9430  -1428   9712  14370
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept) 104352.9     5619.4  18.570 7.29e-08 ***
#   Experience    1352.3      325.7   4.152   0.0032 **
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 11320 on 8 degrees of freedom
# Multiple R-squared:  0.683,   Adjusted R-squared:  0.6434
# F-statistic: 17.24 on 1 and 8 DF,  p-value: 0.0032


predict(salaries.lm, data.frame(Experience = 6),interval = "prediction")
#     fit      lwr      upr
# 1 112466.4 84544.64 140388.3


#now the weighted least squares model
salaries.weight.lm <- lm(ThirdQuartile ~ Experience, data= profSals, weights=sqrt(SampleSize))
summary(salaries.weight.lm)
# Call:
#   lm(formula = ThirdQuartile ~ Experience, data = profSals, weights = sqrt(SampleSize))
#
# Weighted Residuals:
#   Min     1Q Median     3Q    Max
# -28587 -19813   -863  22448  35950
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept) 104549.2     5726.8  18.256 8.33e-08 ***
#   Experience    1262.7      332.8   3.794  0.00528 **
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 25520 on 8 degrees of freedom
# Multiple R-squared:  0.6428,  Adjusted R-squared:  0.5982
# F-statistic:  14.4 on 1 and 8 DF,  p-value: 0.005279


plot(salaries.weight.lm)
predict(salaries.weight.lm, data.frame(Experience = 6),interval = "prediction")
#     fit      lwr      upr
# 1 112125.4 52437.59 171813.2


##chapter 5 #1 Overdue bills
par(mfrow=c(1,1))

overdue <- read.csv("Data/overdue.txt", sep="")
overdue$TYPE <- ""
overdue$TYPE[1:48] <- "RESIDENTIAL"
overdue$TYPE[49:96] <- "COMMERCIAL"
overdue$TYPE <- as.factor(overdue$TYPE)

plot(x=overdue$BILL, y=overdue$LATE, pch=ifelse(overdue$TYPE=="RESIDENTIAL", 18,22), ylab = "# of days late",
xlab="Amount of overdue bill in $'s")
legend(200, 20,legend=c("RESIDENTIAL","COMMERCIAL"),pch=c(18,22), cex=0.8)

#from the plot, it looks like it will be a multiple lines model, the residential and commercial clearly have
different slopes
```

```
#but let's fit a regular lm so we can run the anova test and show that type will have an impact

overdue.lm <- lm(LATE ~ BILL, data=overdue)
summary(overdue.lm)

overdue.lm2 <- lm(LATE ~ BILL + TYPE + TYPE:BILL, data=overdue)
summary(overdue.lm2)
# Call:
#   lm(formula = LATE ~ BILL + TYPE + TYPE:BILL, data = overdue)
#
# Residuals:
#   Min       1Q   Median       3Q      Max
# -12.1211  -2.2163   0.0974   1.9556   8.6995
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)          101.758184   1.198504   84.90   <2e-16 ***
#   BILL                  -0.190961   0.006285  -30.38   <2e-16 ***
#   TYPERESIDENTIAL      -99.548561   1.694940  -58.73   <2e-16 ***
#   BILL:TYPERESIDENTIAL   0.356644   0.008888   40.12   <2e-16 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 3.371 on 92 degrees of freedom
# Multiple R-squared:  0.9803,  Adjusted R-squared:  0.9796
# F-statistic:  1524 on 3 and 92 DF,  p-value: < 2.2e-16

anova(overdue.lm,overdue.lm2)
## just as suspected, the lm2 model that takes type into account looks way better
#run the plots again and add the lines
plot(x=overdue$BILL, y=overdue$LATE, pch=ifelse(overdue$TYPE=="RESIDENTIAL", 18,22), ylab = "# of days late",
xlab="Amount of overdue bill in $'s")
legend(200, 20,legend=c("RESIDENTIAL","COMMERCIAL"),pch=c(18,22), cex=0.8)
abline(101.758,-0.191, col="red")
abline(101.758-99.55,-0.191+0.357,col="blue")
legend(50,60,legend = c("Res reg line","comm reg line"), col=c("blue","red"),lwd=2, cex=0.7)


##chapter 5 #2 Houston Chronicle
par(mfrow=c(1,1))

houston <- read.csv("Data/HoustonChronicle.csv")
houston$Year<-as.factor(houston$Year)
plot(x=houston$X.Low.income.students, y=houston$X.Repeating.1st.Grade,
     pch=ifelse(houston$Year==1994, 12,16), ylab = "% of Students Repeating First Grade", xlab="% of Low-Income
Students")
legend(0, 18,legend=c("1994","2004"),pch=c(12,16), cex=0.8)
abline(lsfit(x=houston$X.Low.income.students,y=houston$X.Repeating.1st.Grade))

houston.lm <- lm(X.Repeating.1st.Grade ~ X.Low.income.students, data=houston) #reduced model

summary(houston.lm)
## below is the output of the reduced model.

# Call:
#   lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students, data = houston)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -8.9845 -2.5072 -0.4184  1.8505 11.1067
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)            2.91419    0.83836   3.476 0.000709 ***
#   X.Low.income.students  0.07550    0.01823   4.141 6.47e-05 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 3.821 on 120 degrees of freedom
# Multiple R-squared:  0.125,   Adjusted R-squared:  0.1177
# F-statistic: 17.14 on 1 and 120 DF,  p-value: 6.472e-05

# You can see that the slope is significant, but it's really close to 0, 1% change in low-income students attributes
to a .07% change in % of students
# failing the 1st grade.

houston.lm2 <- lm(X.Repeating.1st.Grade ~ X.Low.income.students + Year:X.Low.income.students,data=houston) #model for
part b
houston.lm3 <- lm(X.Repeating.1st.Grade ~ X.Low.income.students + Year + Year:X.Low.income.students,data=houston)
#full model

anova(houston.lm,houston.lm2,houston.lm3)

# Analysis of Variance Table
#
```

```
# Model 1: X.Repeating.1st.Grade ~ X.Low.income.students
# Model 2: X.Repeating.1st.Grade ~ X.Low.income.students + Year:X.Low.income.students
# Model 3: X.Repeating.1st.Grade ~ X.Low.income.students + Year + Year:X.Low.income.students
# Res.Df    RSS Df Sum of Sq      F Pr(>F)
# 1    120 1751.9
# 2    119 1745.1  1    6.7886 0.4592 0.4993
# 3    118 1744.4  1    0.7233 0.0489 0.8253

#after running the anova, there looks like for both parts b and c, there is no association with anything.

##chapter 5 #3 Chateau Latour
wine <- read.csv("Data/Latour.txt",sep="")

#a)
wine.lm.red <- lm(Quality ~ EndofHarvest, data=wine)
wine.lm.full <- lm(Quality ~ EndofHarvest + Rain + Rain:EndofHarvest, data=wine)

summary(wine.lm.red)
# Call:
#   lm(formula = Quality ~ EndofHarvest, data = wine)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -2.2374 -0.7618 -0.1888  0.9510  1.9267
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)   6.43784    0.86005   7.485 2.96e-09 ***
#   EndofHarvest -0.08206    0.02110  -3.889 0.000352 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1.13 on 42 degrees of freedom
# Multiple R-squared:  0.2648,  Adjusted R-squared:  0.2473
# F-statistic: 15.13 on 1 and 42 DF,  p-value: 0.0003522

summary(wine.lm.full)
# Call:
#   lm(formula = Quality ~ EndofHarvest + Rain + Rain:EndofHarvest,
#      data = wine)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -1.6833 -0.5703  0.1265  0.4385  1.6354
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)        5.16122    0.68917   7.489 3.95e-09 ***
#   EndofHarvest      -0.03145    0.01760  -1.787   0.0816 .
# Rain               1.78670    1.31740   1.356   0.1826
# EndofHarvest:Rain -0.08314    0.03160  -2.631   0.0120 *
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.7578 on 40 degrees of freedom
# Multiple R-squared:  0.6848,  Adjusted R-squared:  0.6612
# F-statistic: 28.97 on 3 and 40 DF,  p-value: 4.017e-10

anova(wine.lm.red,wine.lm.full)
# Analysis of Variance Table
#
# Model 1: Quality ~ EndofHarvest
# Model 2: Quality ~ EndofHarvest + Rain + Rain:EndofHarvest
# Res.Df    RSS Df Sum of Sq      F    Pr(>F)
# 1     42 53.587
# 2     40 22.970  2    30.616 26.657 4.388e-08 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## looking at the summaries of the models, and the anova table, the interaction term is
## statistically significant.

#b)

#i) No Rain




#ii) W/ Rain
```