COMP 4442 Midterm, Winter 2022

Krista Miller

knitr::opts\_chunk$set(echo = TRUE)  
  
library(leaps)

## Warning: package 'leaps' was built under R version 4.0.5

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble 3.1.0 v dplyr 1.0.5  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(GGally)

## Warning: package 'GGally' was built under R version 4.0.5

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(ggpubr)

## Warning: package 'ggpubr' was built under R version 4.0.5

library("dplyr")   
library(readxl)  
library(nortest)  
library(lawstat)

## Warning: package 'lawstat' was built under R version 4.0.5

There are six questions on this midterm, all of which have multiple parts. Please be sure to provide answers to all parts of each question. Each question has an associated .csv file, which you will load into memory at the beginning of the question. All of the included data sets are simulated, so any results should not be taken as evidence for or against the existence of anything in the real world. The data were simulated to minimize the ambiguity and messiness that typifies real data. If you feel that something is ambiguous in a way that impedes your ability to answer the questions, please let me know.

I believe in you!

## Question 1: Basic ANOVA - 10 points total

A tire manufacturing company wants to know if different formulations of tire rubber result in differences in tire durability. They are interested in four different rubber formulations (“form”). To test this, 20 tires of each rubber formulation are selected for testing. Aside from the rubber formulation, all 80 tires in this experiment are otherwise exactly the same. The durability of each tire is tested using a durability machine, which mimics the forces and stress a tire is exposed to when installed in a standard sedan driving down a flat asphalt road at 60 miles per hour. The machines tracks how many miles the tire has “traveled” based on the number of rotations of the tire. The durability test stops when the tire’s structure fails, which is when the durability machine records the number of traveled miles (“miles”). The data from this hypothetical experiment is contained in the Q1data.csv file.

Run the code chunk below to load the data into memory before beginning your work on this question.

tires <- read.csv("Q1data.csv", header=TRUE, sep=",") # Loads the CSV file into memory. You may need to adapt this line to work on your computer  
  
tires$form <- as.factor(tires$form)  
  
str(tires)

## 'data.frame': 80 obs. of 2 variables:  
## $ miles: num 81171 81419 79781 79093 81212 ...  
## $ form : Factor w/ 4 levels "a","b","c","d": 1 1 1 1 1 1 1 1 1 1 ...

# Q1, Part 1: Assessing the normality of groups assumption (2 points)

You will assess the assumption of normality in two ways: quantitatively and visually.

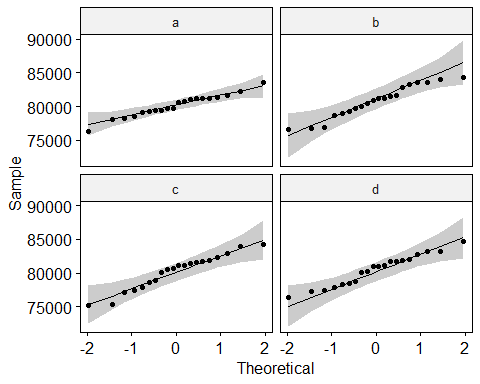
In this first code chunk, please conduct an appropriate *quantitative* assessment of the normality assumption and display the results.

# Code for your quantitative assessment of the normality assumption goes here  
  
tires%>%group\_by(form)%>%summarize(pval=shapiro.test(miles)$p)

## # A tibble: 4 x 2  
## form pval  
## <fct> <dbl>  
## 1 a 0.916  
## 2 b 0.309  
## 3 c 0.349  
## 4 d 0.623

In this second code chunk, please conduct an appropriate *visual* assessment of the normality assumption and display the visualization/s you create.

# Code for your visual assessment of the normality assumption goes here  
  
ggqqplot(tires,x="miles",facet.by = "form")



# Q1, Part 2: Assessing the equality of variances of groups assumption (2 points)

You will assess the assumption of equality of variances in two ways: quantitatively and visually.

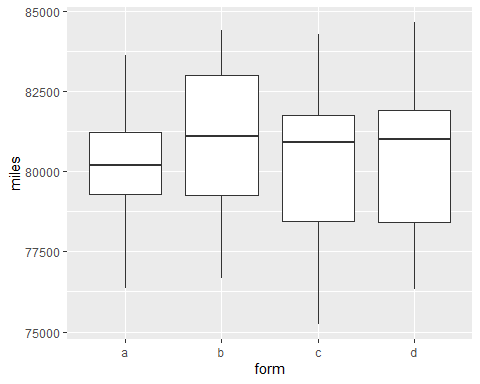
In this first code chunk, please conduct an appropriate *quantitative* assessment of the equality of variances assumption and display the results.

# Code for your quantitative assessment of the equality of variances assumption goes here  
  
levene.test(tires$miles,tires$form)

##   
## Modified robust Brown-Forsythe Levene-type test based on the absolute  
## deviations from the median  
##   
## data: tires$miles  
## Test Statistic = 0.96439, p-value = 0.414

In this second code chunk, please conduct an appropriate *visual* assessment of the equality of variances assumption and display the visualization/s you create.

# Code for your visual assessment of the equality of variances assumption goes here  
  
ggplot(tires, aes(x=form, y= miles)) + geom\_boxplot()



# Q1, Part 3: Fitting the ANOVA model (2 points)

Now, you will conduct an ANOVA on the tires data set that can provide an answer to the research question: do different formulations of tire rubber have different durability? Please be sure to display the results of your analysis.

# Code for your ANOVA goes here  
  
tires.aov <- aov(miles~form, data=tires)  
  
# Don't forget to display the results using the summary() function!  
  
summary(tires.aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## form 3 5340001 1780000 0.345 0.793  
## Residuals 76 392461623 5163969

# Q1, Part 4: Interpreting the ANOVA results (4 points)

1. What is the null hypothesis being tested by the ANOVA you conducted? Based on the results of your analysis, do you reject or fail to reject this null hypothesis?

Your answer here: The null hypothesis is that the tire groups (forms) have equal means in terms of tire durability (miles).

1. What do the results of your ANOVA suggest about the research question? That is, what is your answer to the tire manufacturer’s research question about tire durability?

Your answer here: The high p-value (0.793) gives strong evidence supporting this conclusion: I fail to reject the null hypothesis and conclude that there is not a statistically significant difference in the mean miles traveled among the tire forms.

## Question 2: Multifactor ANOVA - 10 points

A health researcher designed an experiment to test the effects of two medications, Lowesterol and Lipidown, on LDL cholesterol levels of people who had been diagnosed as having high cholesterol but no other health problems. He recruited 160 participants, all of whom took two pills each day for 90 days. For 40 participants, both pills were placebos. For 40 participants, one pill contained Lowesterol and the other pill was a placebo. For 40 participants, one pill contained Lipidown and the other pill was a placebo. For the last 40 participants, one pill contained Lowesterol and the other contained Lipidown. After 90 days, each participant gave a blood sample and the LDL level in their blood was recorded. The data from this hypothetical experiment is contained in the Q2data.csv file.

Run the code chunk below to load the data into memory before beginning your work on this question.

drugs <- read.csv("Q2data.csv", header=TRUE, sep=",") # Loads the CSV file into memory. You may need to adapt this line to work on your computer  
  
drugs$lowesterol <- as.factor(drugs$lowesterol)  
drugs$lipidown <- as.factor(drugs$lipidown)  
  
str(drugs)

## 'data.frame': 160 obs. of 3 variables:  
## $ lowesterol: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ lipidown : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ post.ldl : num 172 170 173 175 166 ...

# Q2, Part 1: Fitting the factorial ANOVA model (4 points)

Now, you will conduct a two-way ANOVA with an interaction on the drug data. Use post.ldl as the outcome. Please be sure to display the results of your analysis.

# Code for your ANOVA goes here  
  
drugs.aov <- aov(post.ldl~lowesterol \* lipidown, data= drugs)  
  
# Don't forget to display the results using the summary() function!  
  
summary(drugs.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## lowesterol 1 186 186 2.442 0.120   
## lipidown 1 18539 18539 243.589 <2e-16 \*\*\*  
## lowesterol:lipidown 1 122 122 1.608 0.207   
## Residuals 156 11873 76   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Q2, Part 2: Interpreting the factorial ANOVA model (6 points)

Use the output from the factorial ANOVA to answer the following three questions.

1. Is the main effect of Lowesterol significant?

Your answer here (yes/no): no

1. Is the main effect of Lipidown significant?

Your answer here (yes/no): yes

1. Is the interaction between Lipidown and Lowesterol significant?

Your answer here (yes/no): no

## Question 3: Multiple Regression - 20 points total

A security firm contracted by a shopping center wants to examine the factors that contribute to “loss” (theft of money or goods by customers or employees of a store) in the 200 stores in the shopping center. They have four pieces of information reported by the shopping center about each store: the amount of loss in dollars (“loss”, continuous), the area of the store in square feet (“area”, continuous), the average number of people who walk into the store on a weekly basis (“traffic”, continuous), and whether the store is primarily a retail store (retail=1) or a service-oriented store (retail=0). The data from this hypothetical study is contained in the Q3data.csv file.

Run the code chunk below to load the data into memory before beginning your work on this question.

mall <- read.csv("Q3data.csv", header=TRUE, sep=",") # Loads the CSV file into memory. You may need to adapt this line to work on your computer  
  
mall$retail <- as.factor(mall$retail)  
  
str(mall)

## 'data.frame': 200 obs. of 4 variables:  
## $ loss : num 1870 1532 1438 1537 1628 ...  
## $ retail : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 1 2 2 ...  
## $ area : int 1681 1448 1322 1356 1429 1568 1644 1880 1445 1727 ...  
## $ traffic: int 373 149 217 347 384 228 337 365 142 307 ...

# Q3, Part 1: Fitting the regression model (2 points)

Now, you will conduct a multiple regression analysis. The outcome for this regression will be loss, and the predictors will be retail, area, and traffic. Be sure to display the results of your analysis.

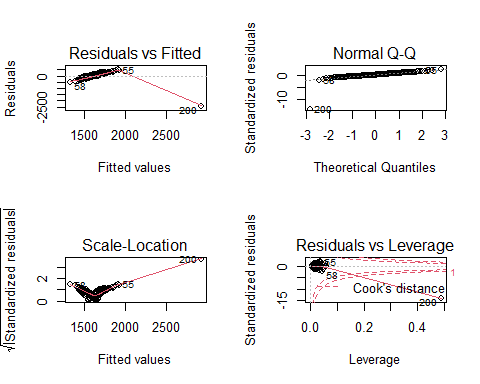
# Code for your regression goes here  
  
mall.reg <- lm(loss ~ retail+area+traffic, data= mall)  
  
# Don't forget to display the results using the summary() function!  
  
summary(mall.reg)

##   
## Call:  
## lm(formula = loss ~ retail + area + traffic, data = mall)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2427.2 -103.7 0.8 117.4 499.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 966.02901 95.02144 10.166 < 2e-16 \*\*\*  
## retail1 51.54747 40.51284 1.272 0.205   
## area 0.36681 0.04812 7.623 1.04e-12 \*\*\*  
## traffic 0.26017 0.17359 1.499 0.136   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 242.5 on 196 degrees of freedom  
## Multiple R-squared: 0.2437, Adjusted R-squared: 0.2321   
## F-statistic: 21.05 on 3 and 196 DF, p-value: 7.283e-12

# Q3, Part 2: Checking diagnostic plots (4 points)

Please display the diagnostic plots for the model you fit in the previous part of this question and answer the question below:

# Code for your regression diagnostic plots  
  
par(mfrow = c(2, 2))  
  
plot(mall.reg)



1. What is the most obvious problem that all of the diagnostic plots for this model share?

Your answer here: point 200 is influential in all of the diagnostic plots

1. What would be a good solution to this specific problem?

Your answer here: A good solution is to remove this potentially influential point, and then see how the removal of point #200 changes the model estimates and diagnostic plots.

# Q3, Part 3: Re-fitting the regression model (4 points)

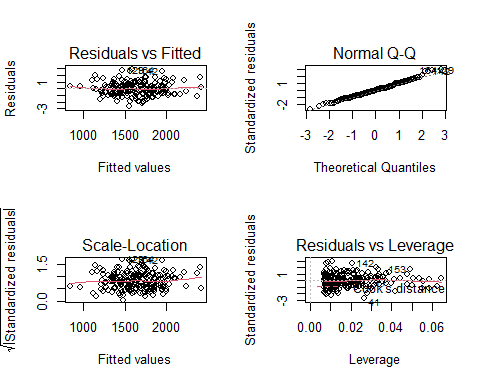
Now, implement the solution you proposed in the last part and re-fit the regression model. Be sure to display the results of your updated analysis.

# Code for any changes you make to implement your solution here  
mall.remove <- mall[-c(200),]  
  
  
# Code for your re-fitted regression goes here  
  
mall.reg.change <- lm(loss ~ retail+area+traffic, data= mall.remove)  
  
# Don't forget to display the results of your analysis using the summary() function!  
  
summary(mall.reg.change)

##   
## Call:  
## lm(formula = loss ~ retail + area + traffic, data = mall.remove)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.58889 -0.70121 -0.00033 0.57188 2.85916   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.5986718 0.4728421 7.611 1.14e-12 \*\*\*  
## retail1 4.9477571 0.1638849 30.190 < 2e-16 \*\*\*  
## area 0.9996022 0.0002662 3755.487 < 2e-16 \*\*\*  
## traffic 0.5005116 0.0007033 711.670 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9777 on 195 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 4.868e+06 on 3 and 195 DF, p-value: < 2.2e-16

Next, display the updated diagnostic plots and answer the question below

# Display your updated diagnostic plots here  
  
par(mfrow = c(2, 2))  
  
plot(mall.reg.change)



Did your solution to the problem you identified in Q3, Part 2 noticeably improve the diagnostic plots of the model?

Your answer here: Yes

# Q3, Part 4: Interpreting the re-fitted regression model (10 points)

1. Interpret the estimated intercept in the context of the predicted outcome and the predictors.

Your answer here: Without being a primarily retail store, an increase in store area, or increase in traffic, the loss is predicted to be $3.59.

1. Interpret the coefficient associated with retail:

Your answer here: For a primarily retail store, (and no increase in store area or increase in traffic), the loss is predicted to be $8.53.

1. What is the predicted amount of loss for a non-retail store that has an area of 1000 square feet and average weekly traffic of 200?

= 3.59+ 0(4.94) +1000(0.999)+ 200(0.500)

Predicted amount of loss= $1,102.59

## Question 4: Automated model selection - 30 points total

The data set Q4data.csv contains nine variables: y, x1, x2, x3, x4, x5, x6, x7, and x8. All of these variables are continuous.

Run the code chunk below to load the data into memory before beginning your work on this question.

many.var <- read.csv("Q4data.csv", header=TRUE, sep=",") # Loads the CSV file into memory. You may need to adapt this line to work on your computer  
  
str(many.var)

## 'data.frame': 200 obs. of 9 variables:  
## $ y : num 238 325 260 367 297 ...  
## $ x1: int 7 15 11 9 14 12 9 -3 15 5 ...  
## $ x2: int 6 13 11 7 10 14 12 10 6 9 ...  
## $ x3: int 6 1 11 8 11 12 10 12 15 5 ...  
## $ x4: int 15 8 0 12 12 12 3 12 12 14 ...  
## $ x5: int 8 17 7 16 7 13 11 13 13 7 ...  
## $ x6: int 9 11 10 14 14 15 10 4 18 6 ...  
## $ x7: int 7 11 10 19 7 12 10 7 10 7 ...  
## $ x8: int 10 10 13 15 12 13 9 11 15 13 ...

# Q4, Part 1: Forward selection - 10 points

First, you will use forward selection to select a model. The outcome will be y and the pool of potential predictors will include x1, x2, x3, x4, x5, x6, x7, and x8. Be sure to include trace=1 as part of your use of the function. After this, display the model selected using forward selection.

# Code for any preliminary steps goes here  
  
lm.null<- lm(y~1, data= many.var)  
  
# Code for your forward selection goes here  
  
forward.model<- step(lm.null, direction='forward', scope= ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, trace=1)

## Start: AIC=1657.33  
## y ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + x7 1 367774 418442 1533.2  
## + x2 1 148023 638194 1617.6  
## + x1 1 101922 684294 1631.6  
## + x4 1 32421 753796 1650.9  
## <none> 786216 1657.3  
## + x3 1 7021 779196 1657.5  
## + x6 1 4784 781433 1658.1  
## + x8 1 4138 782079 1658.3  
## + x5 1 185 786031 1659.3  
##   
## Step: AIC=1533.2  
## y ~ x7  
##   
## Df Sum of Sq RSS AIC  
## + x2 1 181931 236512 1421.1  
## + x1 1 87579 330863 1488.2  
## + x4 1 85084 333358 1489.7  
## + x3 1 12164 406278 1529.3  
## <none> 418442 1533.2  
## + x5 1 4072 414371 1533.2  
## + x8 1 1114 417328 1534.7  
## + x6 1 1 418441 1535.2  
##   
## Step: AIC=1421.09  
## y ~ x7 + x2  
##   
## Df Sum of Sq RSS AIC  
## + x4 1 103663 132848 1307.7  
## + x1 1 94439 142072 1321.2  
## + x3 1 20259 216253 1405.2  
## + x5 1 5408 231104 1418.5  
## <none> 236512 1421.1  
## + x6 1 1215 235296 1422.1  
## + x8 1 49 236463 1423.0  
##   
## Step: AIC=1307.73  
## y ~ x7 + x2 + x4  
##   
## Df Sum of Sq RSS AIC  
## + x1 1 110893 21955 949.68  
## + x3 1 14405 118444 1286.78  
## <none> 132848 1307.73  
## + x5 1 1170 131678 1307.96  
## + x6 1 565 132283 1308.88  
## + x8 1 3 132845 1309.72  
##   
## Step: AIC=949.68  
## y ~ x7 + x2 + x4 + x1  
##   
## Df Sum of Sq RSS AIC  
## + x3 1 21750.3 204.5 16.46  
## + x5 1 384.6 21570.2 948.15  
## + x8 1 227.6 21727.1 949.60  
## <none> 21954.8 949.68  
## + x6 1 47.1 21907.7 951.25  
##   
## Step: AIC=16.46  
## y ~ x7 + x2 + x4 + x1 + x3  
##   
## Df Sum of Sq RSS AIC  
## + x5 1 2.64717 201.87 15.856  
## <none> 204.51 16.462  
## + x8 1 0.64123 203.87 17.834  
## + x6 1 0.34763 204.16 18.121  
##   
## Step: AIC=15.86  
## y ~ x7 + x2 + x4 + x1 + x3 + x5  
##   
## Df Sum of Sq RSS AIC  
## <none> 201.87 15.856  
## + x8 1 0.67420 201.19 17.187  
## + x6 1 0.42727 201.44 17.432

The model selected by forward selection:

# Display the model selected using forward selection here!  
  
summary(forward.model)

##   
## Call:  
## lm(formula = y ~ x7 + x2 + x4 + x1 + x3 + x5, data = many.var)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.26021 -0.77847 -0.03343 0.68273 2.73044   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.20153 0.39170 10.726 <2e-16 \*\*\*  
## x7 10.03005 0.01514 662.369 <2e-16 \*\*\*  
## x2 7.03129 0.01538 457.194 <2e-16 \*\*\*  
## x4 4.97113 0.01523 326.424 <2e-16 \*\*\*  
## x1 5.01762 0.01495 335.648 <2e-16 \*\*\*  
## x3 2.00461 0.01402 142.933 <2e-16 \*\*\*  
## x5 0.02313 0.01454 1.591 0.113   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.023 on 193 degrees of freedom  
## Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997   
## F-statistic: 1.252e+05 on 6 and 193 DF, p-value: < 2.2e-16

# Q4, Part 2: Backward selection - 10 points

Next, you will use backward selection to select a model. The outcome will be y and the pool of potential predictors will include x1, x2, x3, x4, x5, x6, x7, and x8. Be sure to include trace=1 or trace=TRUE as part of your use of the function. After this, display the model selected using backward selection.

# Code for any preliminary steps goes here  
  
lm.full<- lm(y ~ ., data=many.var)  
  
# Code for your backward selection goes here  
  
backward.model<- step(lm.full, direction ="backward", trace=1)

## Start: AIC=18.71  
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8  
##   
## Df Sum of Sq RSS AIC  
## - x6 1 0 201 17.19  
## - x8 1 1 201 17.43  
## <none> 201 18.71  
## - x5 1 3 203 19.45  
## - x3 1 21110 21311 949.73  
## - x4 1 111048 111249 1280.24  
## - x1 1 117278 117479 1291.14  
## - x2 1 216864 217065 1413.93  
## - x7 1 451763 451964 1560.61  
##   
## Step: AIC=17.19  
## y ~ x1 + x2 + x3 + x4 + x5 + x7 + x8  
##   
## Df Sum of Sq RSS AIC  
## - x8 1 1 202 15.86  
## <none> 201 17.19  
## - x5 1 3 204 17.83  
## - x3 1 21156 21357 948.17  
## - x4 1 111261 111462 1278.62  
## - x1 1 117696 117897 1289.85  
## - x2 1 217865 218066 1412.85  
## - x7 1 457472 457673 1561.12  
##   
## Step: AIC=15.86  
## y ~ x1 + x2 + x3 + x4 + x5 + x7  
##   
## Df Sum of Sq RSS AIC  
## <none> 202 15.86  
## - x5 1 3 205 16.46  
## - x3 1 21368 21570 948.15  
## - x4 1 111447 111649 1276.96  
## - x1 1 117834 118036 1288.09  
## - x2 1 218627 218829 1411.55  
## - x7 1 458884 459086 1559.74

The model selected by backward selection:

# Display the model selected using backward selection here!  
  
summary(backward.model)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x7, data = many.var)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.26021 -0.77847 -0.03343 0.68273 2.73044   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.20153 0.39170 10.726 <2e-16 \*\*\*  
## x1 5.01762 0.01495 335.648 <2e-16 \*\*\*  
## x2 7.03129 0.01538 457.194 <2e-16 \*\*\*  
## x3 2.00461 0.01402 142.933 <2e-16 \*\*\*  
## x4 4.97113 0.01523 326.424 <2e-16 \*\*\*  
## x5 0.02313 0.01454 1.591 0.113   
## x7 10.03005 0.01514 662.369 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.023 on 193 degrees of freedom  
## Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997   
## F-statistic: 1.252e+05 on 6 and 193 DF, p-value: < 2.2e-16

# Q4, Part 3: Best subsets selection - 10 points

Finally, you will use best subsets selection to select a model. The outcome will be y and the pool of potential predictors will include x1, x2, x3, x4, x5, x6, x7, and x8. *Be sure to display a table (filled with either stars or TRUE/FALSE values) that shows which predictors were included in the best models of each size* and *display a plot showing the BIC values of the best models of each size*. After this, display the model selected using best subsets selection.

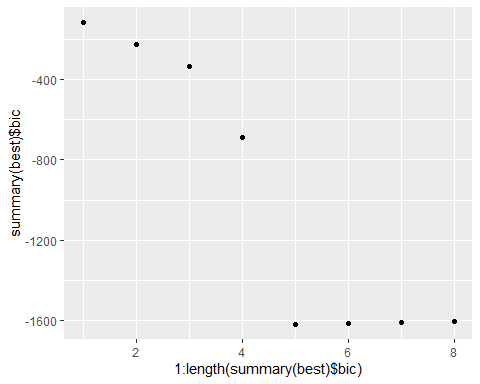
# Code for any preliminary steps goes here  
best<- regsubsets(y~., data=many.var, method= "exhaustive", nvmax=8, nbest=1)  
# Code for your best subsets selection goes here, including your table and plot as requested in the question prompt  
summary(best) #star table

## Subset selection object  
## Call: regsubsets.formula(y ~ ., data = many.var, method = "exhaustive",   
## nvmax = 8, nbest = 1)  
## 8 Variables (and intercept)  
## Forced in Forced out  
## x1 FALSE FALSE  
## x2 FALSE FALSE  
## x3 FALSE FALSE  
## x4 FALSE FALSE  
## x5 FALSE FALSE  
## x6 FALSE FALSE  
## x7 FALSE FALSE  
## x8 FALSE FALSE  
## 1 subsets of each size up to 8  
## Selection Algorithm: exhaustive  
## x1 x2 x3 x4 x5 x6 x7 x8   
## 1 ( 1 ) " " " " " " " " " " " " "\*" " "  
## 2 ( 1 ) " " "\*" " " " " " " " " "\*" " "  
## 3 ( 1 ) " " "\*" " " "\*" " " " " "\*" " "  
## 4 ( 1 ) "\*" "\*" " " "\*" " " " " "\*" " "  
## 5 ( 1 ) "\*" "\*" "\*" "\*" " " " " "\*" " "  
## 6 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " "\*" " "  
## 7 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " "\*" "\*"  
## 8 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"

subsets<- summary(best)$which  
subsets #TRUE/FALSE table

## (Intercept) x1 x2 x3 x4 x5 x6 x7 x8  
## 1 TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE  
## 2 TRUE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE  
## 3 TRUE FALSE TRUE FALSE TRUE FALSE FALSE TRUE FALSE  
## 4 TRUE TRUE TRUE FALSE TRUE FALSE FALSE TRUE FALSE  
## 5 TRUE TRUE TRUE TRUE TRUE FALSE FALSE TRUE FALSE  
## 6 TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE FALSE  
## 7 TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE  
## 8 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

qplot(1:length(summary(best)$bic),summary(best)$bic) # Visualization of BICs



best.subset.bic<-which(summary(best)$bic==min(summary(best)$bic))[1] # Min function searches BICs and shows model number  
best.subset.bic #model of length 5 has minimum BIC- corresponds to x1, x2, x3, x4, x7

## [1] 5

varnames<-attr(subsets,"dimnames")[[2]]  
varnames.best <- varnames[subsets[best.subset.bic,]]   
varnames.best

## [1] "(Intercept)" "x1" "x2" "x3" "x4"   
## [6] "x7"

The model selected by best subsets selection:

# Display the model selected using best subsets selection here!  
bestsubsets.model = lm(y~x1+x2+x3+x4+x7, data=many.var)   
summary(bestsubsets.model)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x4 + x7, data = many.var)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.37981 -0.70104 -0.04696 0.69458 2.84992   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.36813 0.37892 11.53 <2e-16 \*\*\*  
## x1 5.01890 0.01499 334.90 <2e-16 \*\*\*  
## x2 7.03119 0.01544 455.40 <2e-16 \*\*\*  
## x3 2.00733 0.01397 143.64 <2e-16 \*\*\*  
## x4 4.97401 0.01518 327.66 <2e-16 \*\*\*  
## x7 10.02865 0.01518 660.80 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.027 on 194 degrees of freedom  
## Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997   
## F-statistic: 1.491e+05 on 5 and 194 DF, p-value: < 2.2e-16

## Question 5: Nested model selection - 20 points total

The data set Q5data.csv contains nine variables: y, x1, x2, x3, x4, x5, x6, x7, and x8. All of these variables are continuous. This is the same data set used in Question 4, but please reload the data set under a new name to ensure no “cross-contamination” between questions.

Run the code chunk below to load the data into memory before beginning your work on this question.

Q5.var <- read.csv("Q5data.csv", header=TRUE, sep=",") # Loads the CSV file into memory. You may need to adapt this line to work on your computer  
  
str(Q5.var)

## 'data.frame': 200 obs. of 9 variables:  
## $ y : num 238 325 260 367 297 ...  
## $ x1: int 7 15 11 9 14 12 9 -3 15 5 ...  
## $ x2: int 6 13 11 7 10 14 12 10 6 9 ...  
## $ x3: int 6 1 11 8 11 12 10 12 15 5 ...  
## $ x4: int 15 8 0 12 12 12 3 12 12 14 ...  
## $ x5: int 8 17 7 16 7 13 11 13 13 7 ...  
## $ x6: int 9 11 10 14 14 15 10 4 18 6 ...  
## $ x7: int 7 11 10 19 7 12 10 7 10 7 ...  
## $ x8: int 10 10 13 15 12 13 9 11 15 13 ...

# Q5, Part 1: Identifying nested models - 10 points

I fitted five regression models using different sets of predictors. Run the code chunk below to estimate and view the models I fitted. Review the output for these models and answer the questions below.

model.1 = lm(y~x1, data=Q5.var)  
model.2 = lm(y~x1+x2, data=Q5.var)  
model.3 = lm(y~x1+x3, data=Q5.var)  
model.4 = lm(y~x1+x2+x3, data=Q5.var)  
model.5 = lm(y~x1+x2+x3+x1:x2+x1:x3+x2:x3+x1:x2:x3, data=Q5.var)  
  
summary(model.1)

##   
## Call:  
## lm(formula = y ~ x1, data = Q5.var)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -126.218 -45.454 -6.811 43.005 144.496   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 253.6875 9.2338 27.474 < 2e-16 \*\*\*  
## x1 4.6256 0.8518 5.431 1.64e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 58.79 on 198 degrees of freedom  
## Multiple R-squared: 0.1296, Adjusted R-squared: 0.1252   
## F-statistic: 29.49 on 1 and 198 DF, p-value: 1.635e-07

summary(model.2)

##   
## Call:  
## lm(formula = y ~ x1 + x2, data = Q5.var)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -133.797 -36.549 -0.182 36.812 136.086   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 196.4584 11.0837 17.725 < 2e-16 \*\*\*  
## x1 4.7910 0.7511 6.379 1.25e-09 \*\*\*  
## x2 5.8835 0.7735 7.606 1.13e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 51.82 on 197 degrees of freedom  
## Multiple R-squared: 0.3272, Adjusted R-squared: 0.3204   
## F-statistic: 47.91 on 2 and 197 DF, p-value: < 2.2e-16

summary(model.3)

##   
## Call:  
## lm(formula = y ~ x1 + x3, data = Q5.var)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -125.945 -44.400 -5.019 37.200 146.540   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 236.9560 12.6905 18.672 < 2e-16 \*\*\*  
## x1 4.7638 0.8492 5.609 6.81e-08 \*\*\*  
## x3 1.5077 0.7903 1.908 0.0579 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 58.4 on 197 degrees of freedom  
## Multiple R-squared: 0.1454, Adjusted R-squared: 0.1368   
## F-statistic: 16.76 on 2 and 197 DF, p-value: 1.894e-07

summary(model.4)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3, data = Q5.var)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -131.016 -35.653 -1.669 31.783 137.423   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 173.7291 13.6542 12.723 < 2e-16 \*\*\*  
## x1 4.9706 0.7416 6.703 2.13e-10 \*\*\*  
## x2 6.0391 0.7628 7.917 1.77e-13 \*\*\*  
## x3 1.9118 0.6915 2.765 0.00624 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 50.96 on 196 degrees of freedom  
## Multiple R-squared: 0.3525, Adjusted R-squared: 0.3426   
## F-statistic: 35.56 on 3 and 196 DF, p-value: < 2.2e-16

summary(model.5)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3,   
## data = Q5.var)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -126.938 -33.651 -1.145 30.878 138.484   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 135.24083 47.33528 2.857 0.00475 \*\*  
## x1 6.84148 3.85878 1.773 0.07782 .   
## x2 9.08692 4.47064 2.033 0.04347 \*   
## x3 7.94309 3.65649 2.172 0.03106 \*   
## x1:x2 -0.10183 0.36629 -0.278 0.78131   
## x1:x3 -0.43930 0.30055 -1.462 0.14546   
## x2:x3 -0.54001 0.35227 -1.533 0.12693   
## x1:x2:x3 0.03700 0.02962 1.249 0.21309   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 50.47 on 192 degrees of freedom  
## Multiple R-squared: 0.3781, Adjusted R-squared: 0.3554   
## F-statistic: 16.67 on 7 and 192 DF, p-value: < 2.2e-16

1. If Model 5 (model.5) is considered to be the “full model”, which of the remaining models - Models 1, 2, 3, and 4 - are nested relative to it?

Your answer here: model.1, model.2, model.3, and model.4 are all nested relative to model.5

1. If Model 4 (model.4) is considered to be the “full model”, which of the remaining models - Models 1, 2, and 3 - are nested relative to it?

Your answer here: model.1, model.2, model.3 are all nested relative to model.4

1. If Model 3 (model.3) is considered to be the “full model”, which of the remaining models - Models 1 and 2 - are nested relative to it?

Your answer here: model.1 is nested relative to model.3

1. In the code chunk below, specify a new model that is nested relative to Model 5 AND in which Model 2 is nested. That is, specify a model that fits the nested model relationship depicted below:

Model 5 (7 predictor coefficients) <– (Your model, 3-6 predictor coefficients) <– Model 2 (2 predictor coefficients)

Please note that you cannot chose any of the models already fitted in this question. You must specify a model that hasn’t yet been fitted.

model.new <- lm(y~x1+x2+x1:x2, data=Q5.var)  
   
summary(model.new)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x1:x2, data = Q5.var)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -127.907 -34.727 0.507 33.707 136.073   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 229.9463 18.8254 12.215 <2e-16 \*\*\*  
## x1 1.3416 1.7421 0.770 0.4422   
## x2 2.5199 1.7165 1.468 0.1437   
## x1:x2 0.3490 0.1594 2.190 0.0297 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 51.32 on 196 degrees of freedom  
## Multiple R-squared: 0.3433, Adjusted R-squared: 0.3332   
## F-statistic: 34.15 on 3 and 196 DF, p-value: < 2.2e-16

# Q5, Part 2: Nested model testing - 10 points

For this part, you will conduct two nested model tests. In the first test, you will test Model 2 and the new model you specified. In the second test, you will test the new model you specified and Model 5. After you’ve done this, answer the two questions below.

# Your code for the nested model test between model.2 and model.new  
  
m2.vs.new <- anova(model.2, model.new, test='F')  
   
m2.vs.new # Displays result

## Analysis of Variance Table  
##   
## Model 1: y ~ x1 + x2  
## Model 2: y ~ x1 + x2 + x1:x2  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 197 528946   
## 2 196 516314 1 12632 4.7952 0.02972 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Your code for the nested model test between model.new and model.5  
  
new.vs.m5 <- anova(model.new, model.5, test='F')  
   
new.vs.m5 # Displays result

## Analysis of Variance Table  
##   
## Model 1: y ~ x1 + x2 + x1:x2  
## Model 2: y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 196 516314   
## 2 192 488971 4 27343 2.6841 0.03278 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. Based on the result of the test between Model 2 and your new model, which model would you choose?

Your answer here: I would choose the new model. Adding the interaction term has a significant result p = 0.02972)

1. Based on the result of the test between your new model and Model 5, which model would you choose?

Your answer here: I would choose Model 5. The more complex model leads to a significantly improved fit over the new model (p= 0.03278).

## Question 6: Basic logistic regression - 10 points total

A state public health agency wants to investigate the presence of dangerous amounts of lead in drinking water across households within the state. Investigators collected tap water samples from 150 single-family homes and obtained information about each house. Based on advice from an environmental agency, the investigators classified a tap water sample as being safe if it had levels below 15 parts per billion (0) or potentially dangerous if it had levels equal to or greater than 15 parts per billion (1). In addition, they tested the “hardness” (i.e, presence of dissolved calcium, magnesium, and other minerals) of the water sample, which they categorized as being low (0) or high (1). They also noted the age of the house in years and the location type of the house (urban, suburban, or rural). The data from this hypothetical study is contained in the Q6data.csv file.

Run the code chunk below to load the data into memory before beginning your work on this question

lead <- read.csv("Q6data.csv", header=TRUE, sep=",") # Loads the CSV file into memory. You may need to adapt this line to work on your computer  
  
str(lead)

## 'data.frame': 150 obs. of 4 variables:  
## $ age : int 79 77 87 40 53 81 87 40 15 72 ...  
## $ loc : chr "rural" "urban" "rural" "rural" ...  
## $ hard : int 0 0 1 1 0 0 1 0 0 0 ...  
## $ danger: int 1 1 1 1 0 1 1 1 0 1 ...

# Q6, Part 1: Fitting a logistic model - 5 points

Fit a logistic regression model using “danger” (categorical) as the outcome and “age” (continuous), “loc” (categorical), and “hard” (categorical) as predictors. Be sure to display the results of the analysis.

danger.model <- glm(formula= danger~ age + loc + hard, data= lead, family= binomial)  
  
# Don't forget to use the summary() function to display your results!  
  
summary(danger.model)

##   
## Call:  
## glm(formula = danger ~ age + loc + hard, family = binomial, data = lead)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5820 0.1015 0.1905 0.3662 1.0961   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.10729 0.74522 -0.144 0.8855   
## age 0.05024 0.01764 2.848 0.0044 \*\*  
## locsuburban 0.02653 0.73739 0.036 0.9713   
## locurban 1.35343 1.15982 1.167 0.2432   
## hard 1.94739 1.09188 1.784 0.0745 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 78.653 on 149 degrees of freedom  
## Residual deviance: 58.288 on 145 degrees of freedom  
## AIC: 68.288  
##   
## Number of Fisher Scoring iterations: 7

# Q6, Part 2: Interpreting a logistic model - 5 points

1. Based on the results of your analyses, which predictor coefficients (i.e, not including the intercept) were significantly different from zero? There is at least one.

Your answer here: age

1. Of the statistically significant predictor/s you identified in the previous sub-question, which predictor/s (if any) indicate that the presence of a dangerous level of lead is *more likely* as the value of the predictor increases? Which predictors (if any) indicate that that the presence of a dangerous level of lead is *less likely* as the value of the predictor increases?

Dangerous level of lead more likely as values of predictor/s increase/s (your answer here): age

Dangerous level of lead Less likely as values of predictor/s increase/s (your answer here): none