COMP 4442 Midterm, Fall 2021

Ben Karabinus

```
knitr::opts_chunk$set(echo = TRUE)

library(leaps)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(ggpubr)
library(ggeasy)
library(lawstat)
# Load any additional packages, if any, that you use as part of your answers
here
```

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 in
to memory. You may need to adapt this line to work on your computer
# transform rubber formulation to factor
tires$form <- as.factor(tires$form)
# check the data structure
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.

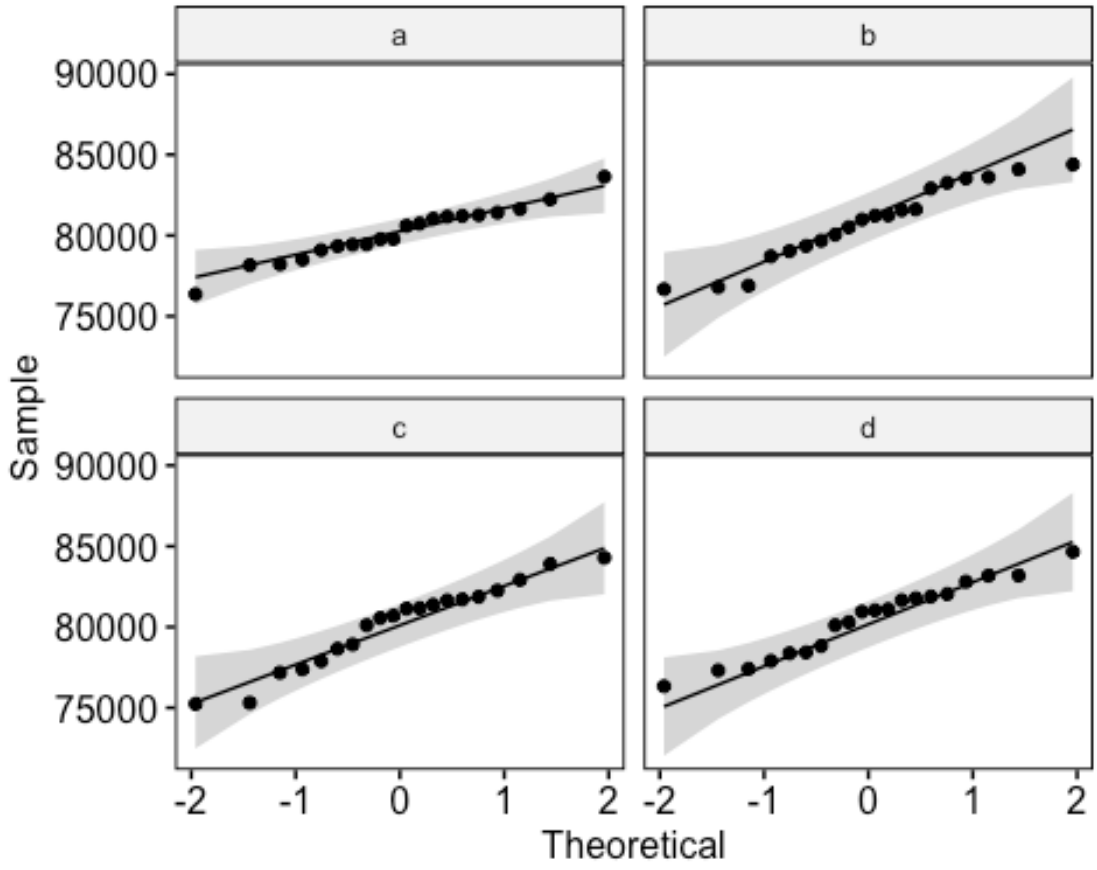
In this first code chunk, please conduct an appropriate *quantitative* assessment of the normality assumption and display the results.

```
# use the Shapiro-Wilk test with dplyr to assess normality and display as tib
ble
sw_tires <- tires %>%
                group_by(form)%>%
                summarize(pval=shapiro.test(miles)$p)
sw_tires
```

```
## # A tibble: 4 × 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.

```
# use ggqqplot to assess the assumption of normality for data by factor level
ggqqplot(tires, x="miles", facet.by = "form")
```



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

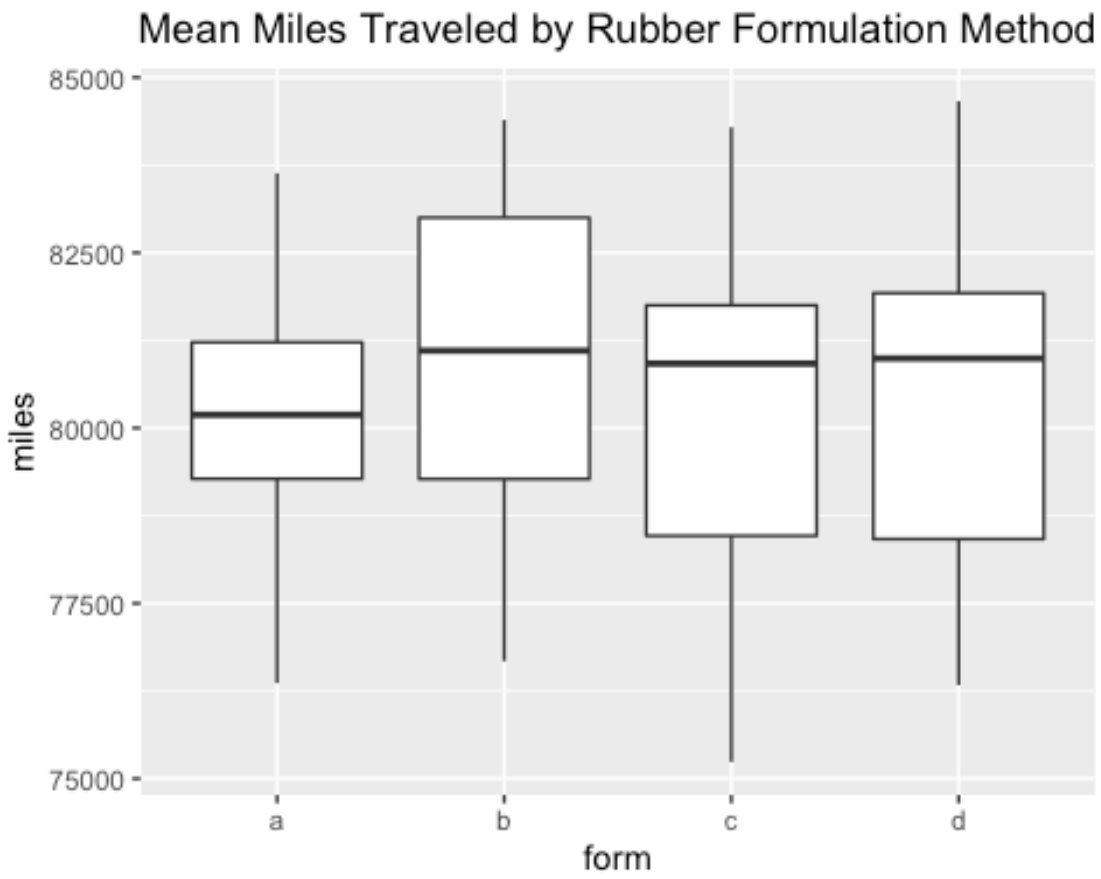
You will assess the assumption of equality of variances in two ways: quantitatively and visually.

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

```
# conduct the Brown-Forsythe test
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.

```
# create ggplot geom_boxplot to assess variance between groups visually
ggplot(tires, aes(x=form, y=miles))+
       geom_boxplot()+
       ggtitle("Mean Miles Traveled by Rubber Formulation Method")+
       ggeasy::easy_center_title()
```



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.

```
# create the tires ANOVA
tires.aov <- aov(miles ~ form, data = tires)
# display a summary of the tires ANOVA
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 (2 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 and alternate hypothesis being tested is as follows:

H_0: The mean number of miles traveled before structural failure of tires is the same across the four different types of rubber formulation methods.

H_a: The mean number of miles traveled before structural failure of tires is different for at least one of four rubber formulation methods.

Based on the results of the above ANOVA we fail to reject the null hypothesis "H_0" and can conclude that there is not a statistically significant difference in mean miles traveled before structural failure of tires created using any of the four rubber formulation methods.

2) 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:

Rubber formulation method does not have a significant effect on the durability of tires.

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 in
to 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.

```
# create two-way ANOVA with interaction term
drugs.aov<-aov(post.ldl ~ lowesterol*lipidown,data=drugs)
# print ANOVA summary
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

2) Is the main effect of Lipidown significant?

Your answer here (yes/no):

Yes

3) 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 int
o 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.

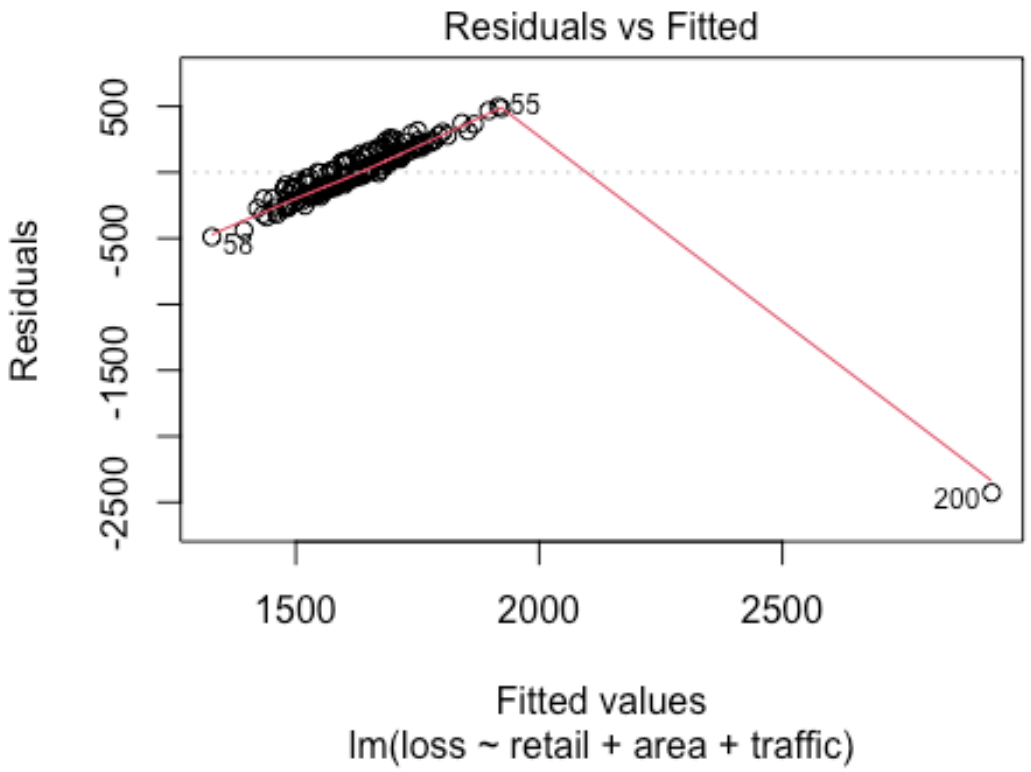
```
# create the "mall.reg" regression model
mall.reg <- lm(loss ~ retail+area+traffic, data=mall)
# print model summary
summary(mall.reg)
```

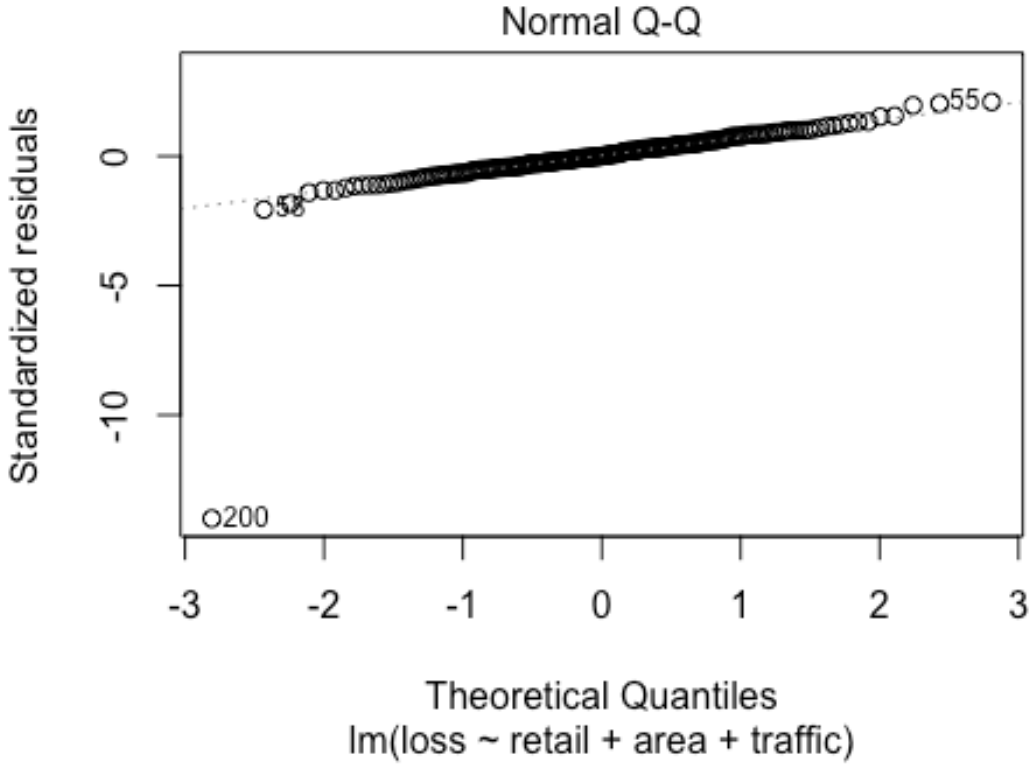
```
##
## Call:
## lm(formula = loss ~ retail + area + traffic, data = mall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2427.2  -103.7     0.8   117.4   499.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 966.02901   95.02144  10.166  < 2e-16 ***
## 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
```

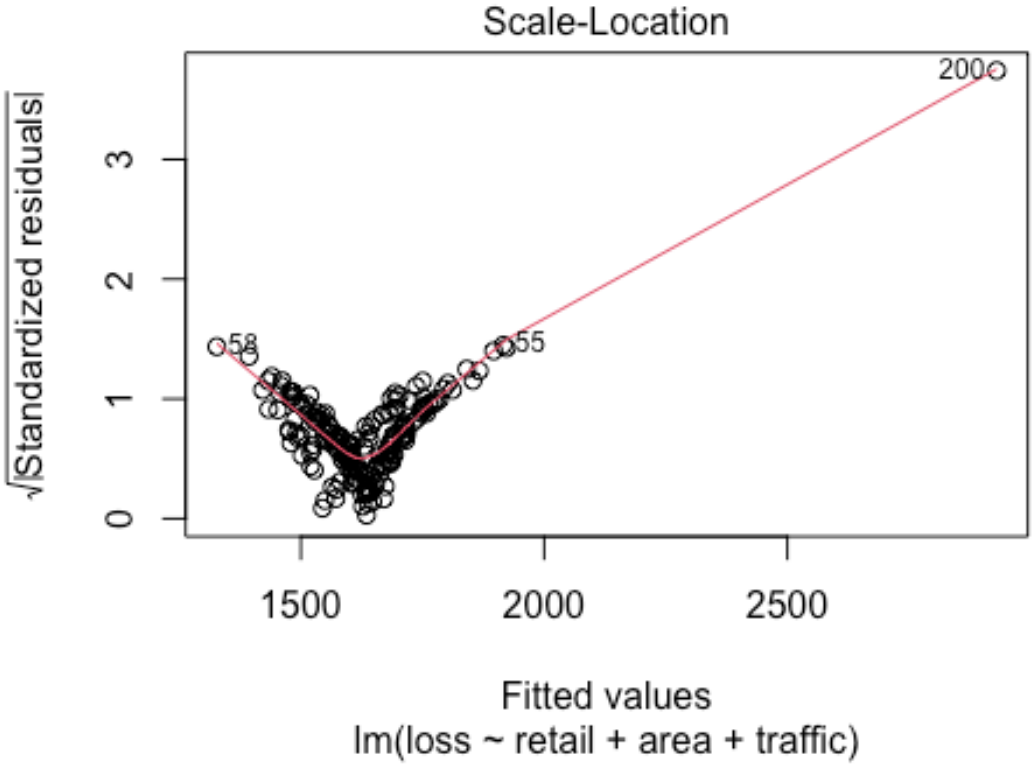
Q3, Part 2: Checking diagnostic plots (4 points)

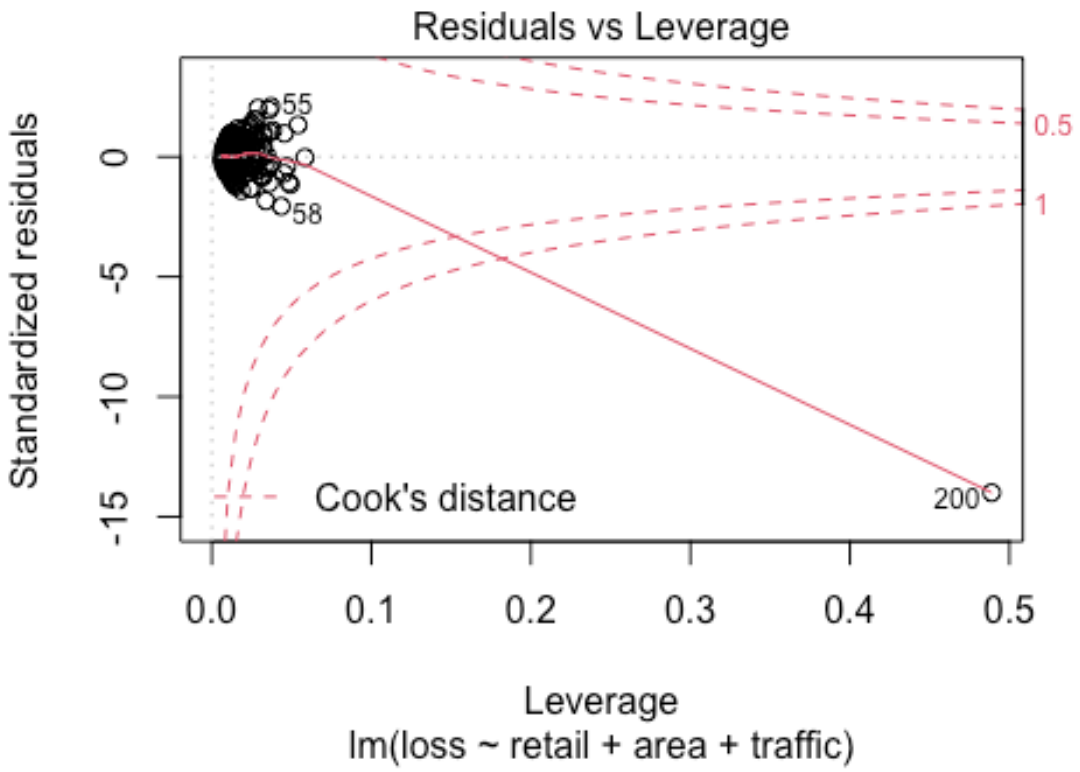
Please display the diagnostic plots for the model you fit in the previous part of this question and answer the question below:

```
# print regression diagnostic plots (mall.reg)
plot(mall.reg)
```


Residuals vs Fitted
Residuals
500
-500
-1500
-2500
55
58
200
1500
2000
2500
Fitted values
lm(loss ~ retail + area + traffic)


Normal Q-Q
Standardized residuals
0
-5
-10
200
55
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(loss ~ retail + area + traffic)


Scale-Location
√|Standardized residuals|
0
1
2
3
200
58
55
1500
2000
2500
Fitted values
lm(loss ~ retail + area + traffic)



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

Your answer here:

Each of the diagnostic plots shows that observation 200 can be considered an outlier and it's inclusion in the model might have an undue effect on parameter estimates determined by the model.

2) What would be a good solution to this specific problem?

Your answer here:

In my opinion the solution to the problem with observation 200 is to remove it from the model entirely. That being said, the "loss" for observation 200 is much lower than that of other records. This should be investigated further to justify removal.

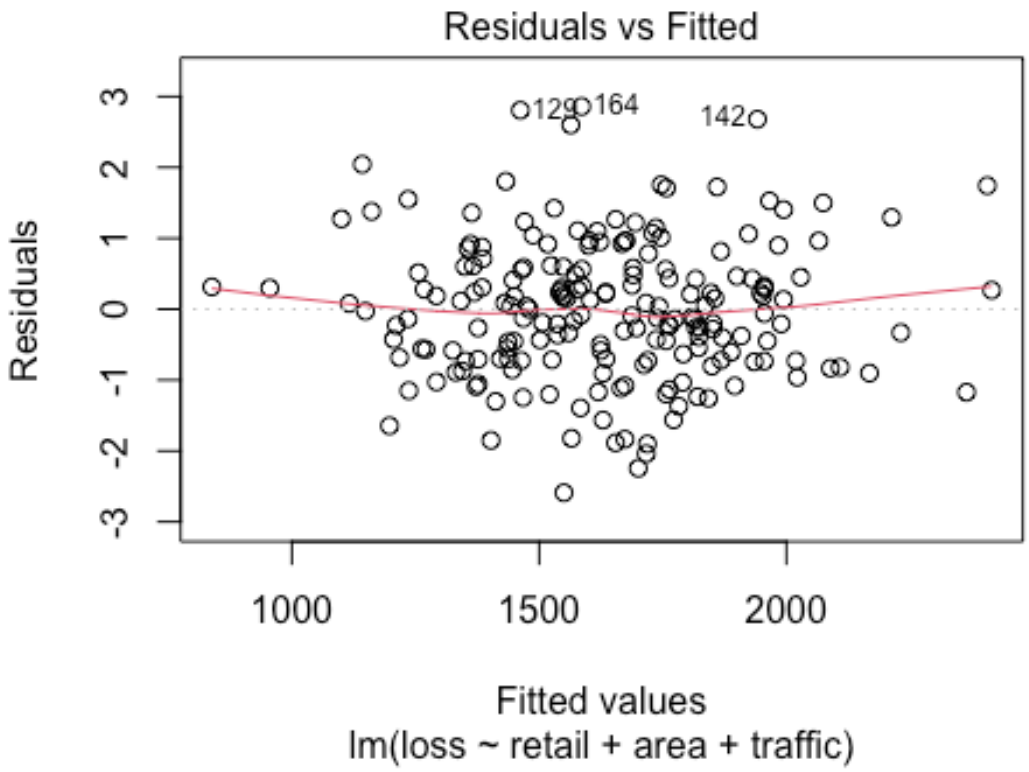
Q3, Part 3: Re-fitting the regression model (4 points)

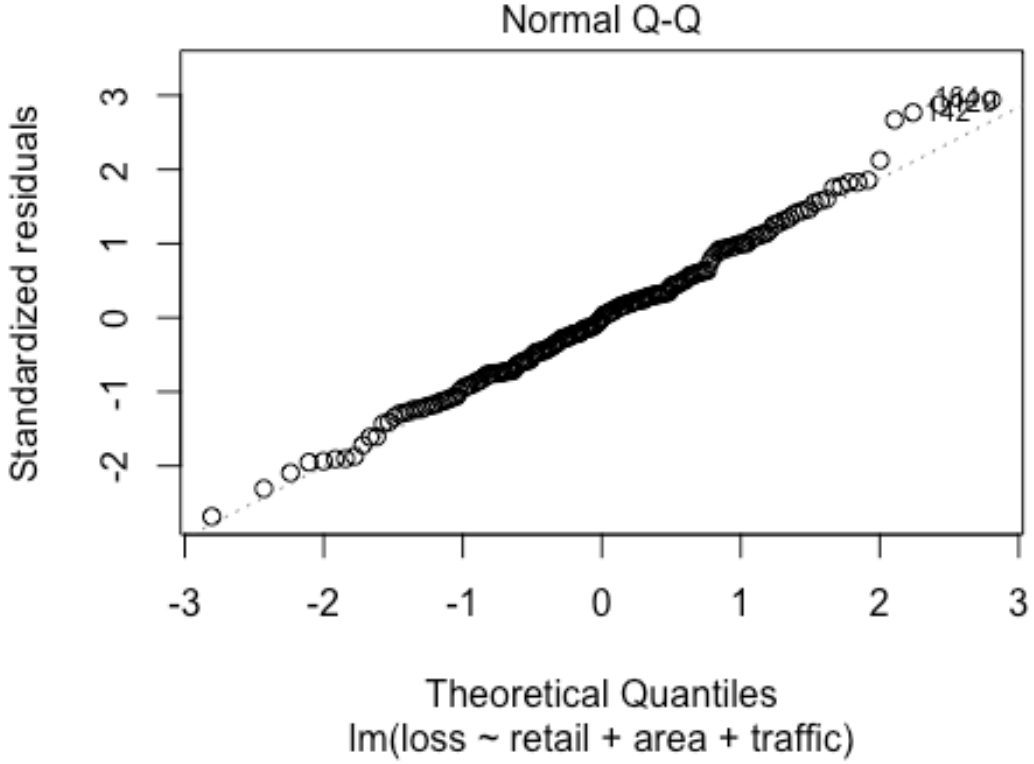
Now, implement the solution you proposed in the last part and re-fit the regression model. Be sure to display the results of your updated analysis.

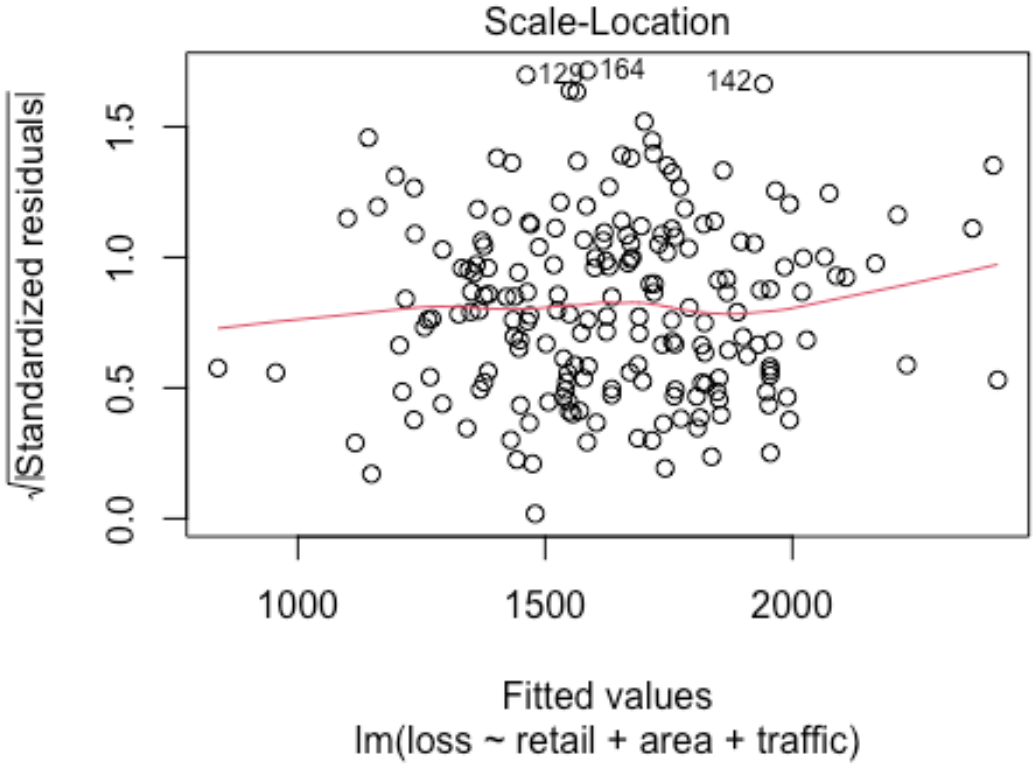
```
# remove observation 200
mall.remove <- mall[-c(200),]
# create the regression model (mall.reg.change)
mall.reg.change <- lm(loss ~ retail+area+traffic, data = mall.remove)
# print the model summary
summary(mall.reg.change)

##
## 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
```

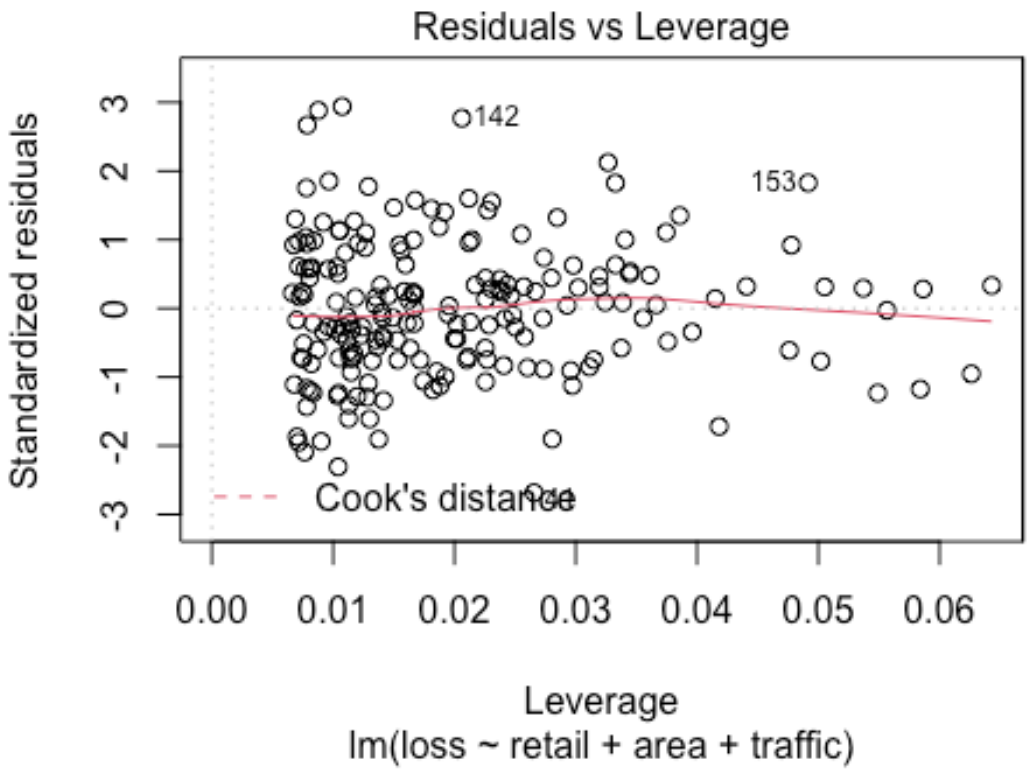
Next, display the updated diagnostic plots and answer the question below

```
# print diagnostic plots (mall.reg.change)
plot(mall.reg.change)
```


Residuals vs Fitted
Residuals
Fitted values
lm(loss ~ retail + area + traffic)
129
164
142
1000
1500
2000
-3
-2
-1
0
1
2
3


Normal Q-Q
Standardized residuals
-2
-1
0
1
2
3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(loss ~ retail + area + traffic)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
129
164
142
1000
1500
2000
Fitted values
lm(loss ~ retail + area + traffic)



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

Your answer here:

Yes, removing observation 200 from the data noticeably improved the diagnostic plots when compared to plots generated prior to removal.

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

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

Your answer here:

The intercept can be interpreted as the predicted loss in dollars holding all other variables constant and assuming the store is a service oriented location.

4) Interpret the coefficient associated with retail:

Your answer here:

Holding all other variables constant predicted loss will increase by ≈ 4.9477571 if the store is a retail oriented location.

5) 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?

```
# alternative method pull coefficients from the model
# intercept <- mall.reg.change$coefficients[1]
# area_coeff <- mall.reg.change$coefficients[3]
# traffic_coeff <- mall.reg.change$coefficients[4]
# intercept
# area_coeff
# traffic_coeff
# predicted_value <- intercept+(1000)*area_coeff+(200)*traffic_coeff
# predicted_value

# create the observation
d_predict <- data.frame(retail = "0", area = 1000, traffic = 200)
# predict sim_tot_fat
predict(mall.reg.change, d_predict)

##        1
## 1103.303
```

Question 4: Automated model selection - 35 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.

```
# check the structure
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 ...
```

```
# create the null model for forward selection (model)
fwd.null <- lm(y ~ 1, data=many.var)
# sanity check
summary(fwd.null)
```

```
##
## Call:
## lm(formula = y ~ 1, data = many.var)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.153  -46.898   -4.929   46.329  187.606
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  298.463      4.445   67.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.86 on 199 degrees of freedom
```

```
# create the scope for forward selection (formula)
fwd.scope <- as.formula("y ~ x1+x2+x3+x4+x5+x6+x7+x8")
# sanity check
fwd.scope
```

```
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8
```

```
# create the forward selection model
forward.model <- step(fwd.null,scope=fwd.scope,
                      direction="forward",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:

```
# print model summary
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.

```
# create the full model for backward selection (model)
bwd.full <- lm(y ~ ., data=many.var)
# sanity check
summary(bwd.full)

##
## Call:
## lm(formula = y ~ ., data = many.var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2392 -0.7997 -0.0197  0.6882  2.6925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.22834    0.45419   9.310   <2e-16 ***
## x1           5.01665    0.01502 334.070   <2e-16 ***
## x2           7.03130    0.01548 454.279   <2e-16 ***
## x3           2.00540    0.01415 141.734   <2e-16 ***
## x4           4.97023    0.01529 325.076   <2e-16 ***
## x5           0.02367    0.01459   1.622    0.106
## x6           0.00953    0.01414   0.674    0.501
## x7          10.03055    0.01530 655.668   <2e-16 ***
## x8          -0.01282    0.01544  -0.830    0.407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.025 on 191 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 9.35e+04 on 8 and 191 DF,  p-value: < 2.2e-16

# create the scope for backward selection (formula)
bwd.scope <- as.formula("y ~ 1")
# sanity check
bwd.scope

## y ~ 1

# creat the backward model
backward.model <- step(bwd.full, scope = bwd.scope,
                       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:

```
# print model summary
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.

```
# create formula for the full model
fmla.full <- as.formula("y ~ x1+x2+x3+x4+x5+x6+x7+x8")
# sanity check
fmla.full

## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8

# create the vector of outcomes
y <- many.var$y
# sanity check
y
```

```
##   [1] 237.7532 325.3688 259.7192 366.6531 296.5794 367.2463 269.5185 214.1
142
##   [9] 311.8367 242.0116 366.9312 256.4772 279.2977 316.4570 312.9366 350.0
568
##  [17] 256.7945 278.7901 229.0988 331.8610 323.9957 342.6213 354.9206 315.7
294
##  [25] 354.9112 246.5775 332.5510 355.4016 268.8212 302.7927 280.5202 242.3
934
##  [33] 221.8377 254.4994 282.9937 263.1927 306.8723 399.7356 258.2973 384.6
293
##  [41] 265.4019 367.3217 204.4284 363.3916 364.3393 212.9588 234.8279 354.4
873
##  [49] 415.5530 311.7465 303.1225 219.9096 229.7972 322.5811 236.1592 158.3
102
##  [57] 320.7251 410.5435 258.3513 443.5742 374.5447 273.8994 205.2411 251.9
140
##  [65] 373.9324 354.0302 319.5406 350.6934 324.7245 273.9033 333.9309 446.5
588
##  [73] 342.3425 205.4421 345.8474 270.1416 324.3911 253.0811 242.9576 295.7
095
##  [81] 220.3299 288.1907 274.2433 316.3723 182.7386 317.6845 432.1039 367.3
954
##  [89] 219.9211 305.3948 380.3220 302.6194 293.1787 233.5237 361.0146 248.6
012
##  [97] 345.9748 202.0069 287.9522 280.7683 211.3514 289.2367 214.5762 199.8
125
## [105] 243.8090 268.5226 355.3597 256.2614 369.1508 273.7384 343.2534 259.8
164
## [113] 299.0103 347.3069 251.1741 389.4882 335.2718 413.3644 240.5480 390.7
160
## [121] 262.9099 203.5076 242.6234 486.0698 256.8990 267.1241 173.7259 305.5
601
## [129] 392.5135 258.8508 301.1659 235.7310 395.0896 354.4546 293.8904 338.9
394
## [137] 308.6823 194.2570 252.3992 402.3449 385.1445 399.3966 280.5495 337.7
726
## [145] 289.9922 350.9001 367.7066 277.1307 329.3940 280.9027 344.0317 286.1
604
## [153] 239.8642 250.9939 238.8925 344.4408 224.9223 371.5712 297.0116 290.9
881
## [161] 244.3198 260.8923 340.7789 238.6242 278.5739 317.9305 354.9622 276.9
451
## [169] 251.6963 304.9849 416.6315 233.4297 294.6901 299.2895 204.2074 339.1
259
## [177] 238.6709 320.0278 406.1445 231.3340 357.6738 194.4230 237.4652 325.2
027
## [185] 464.1628 394.2550 254.9995 287.8387 221.6592 205.0307 233.0087 198.3
368
## [193] 337.8896 330.2279 266.4376 370.3698 278.6888 304.1059 270.7773 257.1
242
```

```r
# create the model matrix
x <- model.matrix(fmla.full, data = many.var)
# sanity check
x
```

```
##     (Intercept) x1 x2 x3 x4 x5 x6 x7 x8
## 1             1  7  6  6 15  8  9  7 10
## 2             1 15 13  1  8 17 11 11 10
## 3             1 11 11 11  0  7 10 10 13
## 4             1  9  7  8 12 16 14 19 15
## 5             1 14 10 11 12  7 14  7 12
## 6             1 12 14 12 12 13 15 12 13
## 7             1  9 12 10  3 11 10 10  9
## 8             1 -3 10 12 12 13  4  7 11
## 9             1 15  6 15 12 13 18 10 15
## 10            1  5  9  5 14  7  6  7 13
## 11            1  3 18  7  9  9 21 16  1
## 12            1 17 14  4  4  3 17  4  8
## 13            1  3  6  8 10 16 16 15  9
## 14            1  1 20 10  7  4  3 11  9
## 15            1 15 10 14  9 12  8  9 -3
## 16            1 14 11 12  9  8 10 13 16
## 17            1  4  2 15  5 14 18 16 14
## 18            1  2 14 11 13  9 13  8 14
## 19            1  6 10 15 11 15 12  4 15
## 20            1 19 12  4  6  9 15 11  5
## 21            1 14 11 16 14  5 14  7  8
## 22            1 -2  8 21 18 -3 20 16  4
## 23            1 15  8 11  7  1 12 16 18
## 24            1  8 11  7  6  6  6 15 -1
## 25            1  5 23 14  9  8 10  9 -1
## 26            1 10  9 17  5  9 15  7 10
## 27            1  6 15 18 11 15  8 10  7
## 28            1 15  9 22 14 15  2 10 16
## 29            1 10 -1 11 14 17 12 13 14
## 30            1 11  9  2 15 19  9 10 -1
## 31            1 12  6 14 15 18 17  7 10
## 32            1 16  8 11 16 18 13  0  9
## 33            1  7 11  8  2  7  9  8  2
## 34            1  5  7 13  8  8 11 11 16
## 35            1  9 12 10 10 19  9  8 18
## 36            1  6 10  6  7  8  9 11  2
## 37            1 13  1 15  2 13  5 19 13
## 38            1 22  8 11 11 15  5 15  6
## 39            1  6  3 14  7 13  8 14  1
## 40            1 24 10 12  5  7 13 14  3
## 41            1  6 12 -1 22 10 14  4  5
## 42            1 12 16 11 10 18 15 12 11
## 43            1  1 12 16  8 12 14  4 12
## 44            1 16  5  9 17 23 12 14  7
```

```
## 45            1  8 20  4  0  7 -1 17 10
## 46            1  6 11 13  9 16 14  3 24
## 47            1  8  4 11 12  2  5  8 11
## 48            1 14 14  8  9  7  2 12  8
## 49            1 15 19  1  0  4  3 20  8
## 50            1  9  5  9 14 -1 13 14  5
## 51            1 18  2 -5 15  0  8 13  6
## 52            1 10  7  5  7 11 13  7 10
## 53            1 10  7  8  0 -1 10 11  7
## 54            1  7 12  1 13 11 10 13 14
## 55            1 15 10 -1 14 10 18  2  9
## 56            1  4 13  9  1  6  8  2 19
## 57            1  9 16  7 13  1 10  8  6
## 58            1  4 10 10 15  9  8 22  7
## 59            1  3  7  8  5  4  8 15  8
## 60            1 14 21  6 18 13  5 12 10
## 61            1 12 15  7 10  7 -4 14 10
## 62            1  8  6 11 15 19  7  9  8
## 63            1 15  2  9  7  5 13  6  5
## 64            1 18 11  7  5 10 11  4 13
## 65            1 16 15  9  5 13 10 14  6
## 66            1 10 18  9  7  3  6 12 15
## 67            1 12 10  3 12 13 11 12 13
## 68            1 11 10 18  9 10  1 14  9
## 69            1 16  4  9 11  8  8 14 10
## 70            1  7  9 13 19 11  1  5 12
## 71            1  9 11  8  6  2  8 16  4
## 72            1 15 15  5 14  8 13 18  5
## 73            1 16  5 11  6 16 10 17 12
## 74            1  8  6  7  3 12 12  9  6
## 75            1  8 10  6 16  7  7 14  9
## 76            1  5 14 11 18  7  5  3 14
## 77            1 15 18 11  5  9 18  7  8
## 78            1  9  5  8 18  9 11  6  6
## 79            1  5  8 13 14  4 18  6 11
## 80            1 16  6  9 18  9 13  6  1
## 81            1  9  5  0  7  1 15 10 12
## 82            1 10 11 11  5  6  6 11  3
## 83            1 10  5  9  5  7 12 14  5
## 84            1 10 11  7 10  3 15 12  7
## 85            1  4  0  6  9 11 14 10  6
## 86            1 17  8 18 17 11 11  5 22
## 87            1 13 18 10 13 10 10 15 14
## 88            1 13 13 11  7  4 10 15 15
## 89            1  5 10  7  5  5  3  8 11
## 90            1  5  8 18  1 11 10 18 14
## 91            1 11  3 20 16  5 13 18 17
## 92            1 11 19  8  7 -1 13  6 10
## 93            1  7 10  6  8  7  7 13 12
## 94            1  9  9  9  7  4 10  7 11
```

```
## 95            1 15 10 15 14 12  9 11 12
## 96            1 15 -4 16  7  7  7 13 11
## 97            1  7  5 18 13  5 12 17 15
## 98            1 11  7  4 15 14 13  1 23
## 99            1  6 16  3  7  6 16 10 10
## 100           1  8 16 14 11  4  7  4 13
## 101           1  9  9 15  6  8 14  4 17
## 102           1  7  7  8 11 12 -1 13  8
## 103           1  0 15 15  9 14 11  3 13
## 104           1 -2 14 11  5  5  5  6 11
## 105           1 12  7  8  9  8 12  7 10
## 106           1 11  7  2 11  4 13 10 12
## 107           1  7  6  9 11 13 18 20 16
## 108           1 15 10 15 17 12 13 -1 16
## 109           1 15  9 16 19 13 20 10  4
## 110           1  0  6  9 10  9 17 16 15
## 111           1  6  2 10 13 19 10 21 14
## 112           1  7 13 10 10 13 11  6 20
## 113           1  5  4 15  8  9 20 17 13
## 114           1 11  8 16 14  8 14 13 16
## 115           1  7  8 12  6  7  9 10 11
## 116           1 17 18  7  6 10 15 13  5
## 117           1 15  6  9 11  8  9 14 16
## 118           1  6 10 17 25 13  6 15  5
## 119           1 12  7  4 10  9  9  7  7
## 120           1  7 11  8 20  6  6 16 18
## 121           1  7 12 10  4  9 17 10 11
## 122           1  3  1 14  6 10  8 12  3
## 123           1 11  4 15 13 17  7  6 11
## 124           1 19 13  8 12  8  2 22 12
## 125           1  0  3 18 13  9  0 13 11
## 126           1  8  8  3 10 16 12 11 11
## 127           1 10  2 10 13 10 15  2  7
## 128           1  1  7 15 15  4  3 14  8
## 129           1 10 15  4 23  8  8 11  5
## 130           1 14  6  4  7  2  6 10 16
## 131           1  9 10 13 13 11 13  9  8
## 132           1 10  6  7 11  9 19  7  7
## 133           1  7 19 11 14 14 19 13 15
## 134           1 11  9  6 10 13  7 17 12
## 135           1 11  8  7 15  2 17  9  4
## 136           1 10 13 12 14 12  2 10  1
## 137           1 17  8 11  4 20  2 12  6
## 138           1  2 10 17 15 19 12  0 10
## 139           1 11  3 19 13  5 15  7  7
## 140           1  2 21 18 15 11  9 13 11
## 141           1 10 10 15 14 11  1 16  4
## 142           1 14  7 -5 17  9 10 20 11
## 143           1 10  5 18 17 11  3  7 11
## 144           1 14 12 11  9  3 11 11 15
```

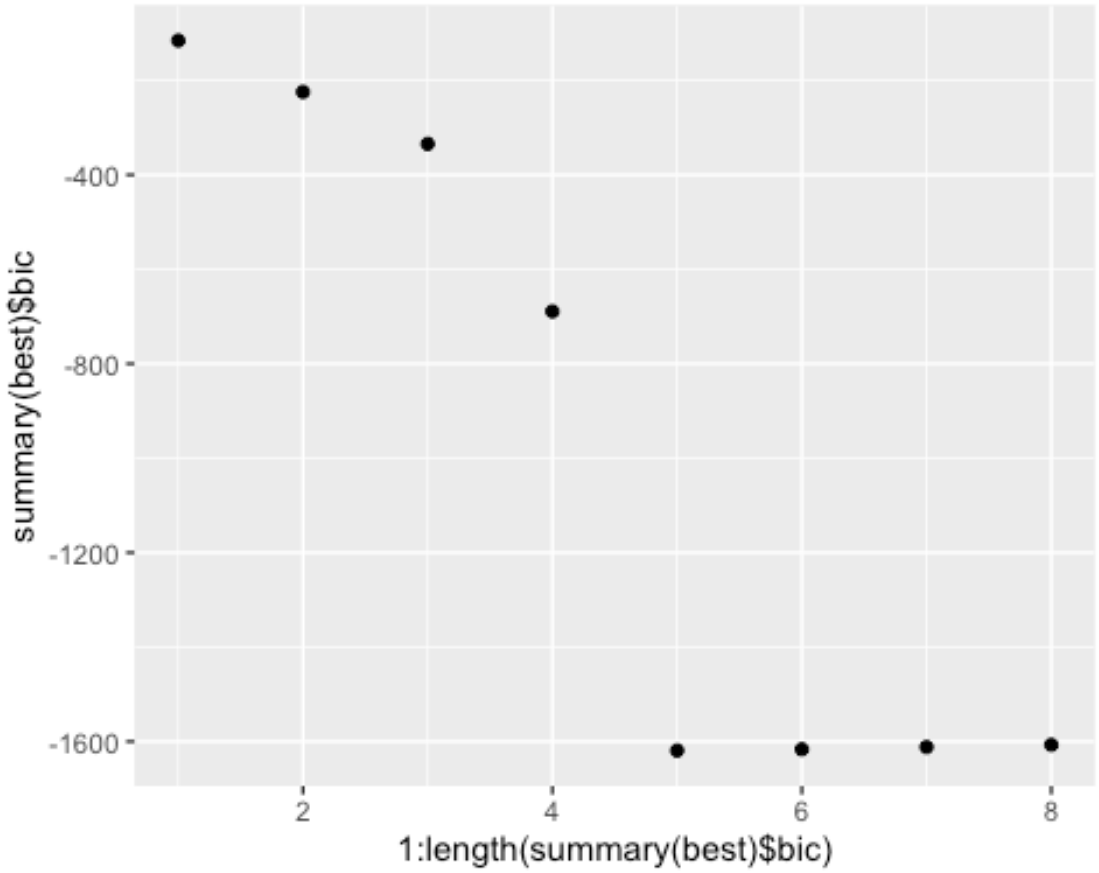
```
## 145           1 10 13 13 16 19 19  4  9
## 146           1 12 10 10 11  9 14 14  7
## 147           1  9  7 19  4  6 18 21 16
## 148           1  2 14 12  4 12  3 12  6
## 149           1 14  5 20 14 11 17 11  6
## 150           1 13 13  8  5  7 11  8  8
## 151           1  4  7 10 22 12 16 14  3
## 152           1 11 15 10 12 20 11  4  4
## 153           1 16  2  6  8  2 18  9  1
## 154           1 12 -3 11 11  6 17 13  9
## 155           1  8 13  6 10 20 16  4  7
## 156           1  8  9 13 12  5 14 15 12
## 157           1  4  8  5  7 17 17 10 10
## 158           1 12 12 28  9  5 15 12  5
## 159           1 12 13 12  9 16 18  7 11
## 160           1 15 10 16 12  6 23  5 15
## 161           1  5 14 11 11 10  1  4  8
## 162           1 18  8 17  7 15 18  4  9
## 163           1 14 11  2  7  9 10 15 10
## 164           1 13  3  2  9 10 -1 10 11
## 165           1 18  5 17  9  8 12  7 12
## 166           1 11  9 10 13 12 12 11  8
## 167           1  9 12  8  9 -5 21 16 13
## 168           1 -1  8 15 14 21 11 12  7
## 169           1  1  9  2 17  4 15  9 15
## 170           1 11  8  2  9  5 15 14  7
## 171           1 10 15 11 15  8 19 16 16
## 172           1 12  3  9 12  6  9  7 16
## 173           1  4 12  1 17  7  8 10  6
## 174           1  8  8 10  8  5  6 14 12
## 175           1  8  9  5 11  6  4  3 11
## 176           1  9 13 17  9 11 16 12 13
## 177           1 10  4  7 10  7  9  9  8
## 178           1 11  5 17  8  4  3 15  3
## 179           1 16 16 17  9 17 12 13  3
## 180           1 11  7  4  5  9  9  9 17
## 181           1  5 11 -2  5 14 16 23  5
## 182           1 11  4 11  7 15  8  5  6
## 183           1 14 10 13  7 14 18  3 12
## 184           1  5  7 18 20 14  6 11 12
## 185           1 19 15 17  7 15  6 19 12
## 186           1 12 13 14  6  5 12 18 19
## 187           1  5  7 13  8  1  8 11  6
## 188           1  7  7 15  0 11 11 17 15
## 189           1  5  4 17 18  7 22  4 17
## 190           1  0 14 12 14  9 10  1  7
## 191           1  6  6  9 20  8  8  4 10
## 192           1  9 -2  1  2 11  7 15 17
## 193           1 13 18  6 14 10  9  6  9
## 194           1 13  9 13  6 17  0 14 12
```

```
## 195           1  7 10 13  4 11 13 11 10
## 196           1 14 10 10 13 17  7 14 12
## 197           1 12 12 12 15  7  7  3 16
## 198           1 13  9  7  5  3 15 13 10
## 199           1  7 13  6  1 13  6 12 11
## 200           1  5  4 10  8 16 10 14 11
## attr(,"assign")
## [1] 0 1 2 3 4 5 6 7 8
```

```
# get model parameters using regsubsets (exhaustive)
# remove intercept (redundant)
best <- regsubsets(x=x[,2:ncol(x)], y=y, method="exhaustive", nvmax=8, nbest=
1)
# create the star table
star <- summary(best)
# get boolean vectors of best model for each size
subsetsBoolean<-summary(best)$which
# create visual of best BIC's for models
bic_plt <- qplot(1:length(summary(best)$bic), summary(best)$bic)

# print star table and plot
star
```

```
## Subset selection object
## 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 ) "*" "*" "*" "*" "*" "*" "*" "*"
```

```
bic_plt
```



The model selected by best subsets selection:

```
# min function searches BICs returns model number of lowest BIC
best.subset.bic<-which(summary(best)$bic==min(summary(best)$bic))[1] # Min fu
nction searches BICs and shows model number
# sanity check
best.subset.bic
```

```
## [1] 5
```

```
# pull dimension (variable) names
varnames <- attr(subsetsBoolean, "dimnames")[[2]]
# get  the best variables
best.varnames <- varnames[subsetsBoolean[best.subset.bic,]]
# print variables
best.varnames
```

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

```
# create the model
best.subsets.model <- lm(y ~ x1+x2+x3+x4+x7, data=many.var)
# print model summary
summary(best.subsets.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
```

Q4, Part 4: Comparing models - 5 points

Now that you have conducted three methods of automated model selection on the same data set, please compare the models that were selected by each method. If the methods included different predictors in their final models, list those predictors that were different.

Please write your answers below:

1) Were there any differences between the forward and backward models?

Your answer here (yes/no):

If yes, list the predictors that did not appear in both models:

No

2) Were there any differences between the forward and best subsets models?

Your answer here (yes/no):

If yes, list the predictors that did not appear in both models:

Yes, x5 only appeared in the forward selection model.

- x5

3) Were there any differences between the backward and best subsets models?

Your answer here (yes/no):

If yes, list the predictors that did not appear in both models:

Yes, x5 only appeared in the backward selection model.

- x5

Question 5: Nested model selection - 15 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 i
nto 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:

Models 1, 2, 3 and 4 are nested relative to model 5. Models 1, 2, 3 and 4 can be specified by setting parameters of model 5 to zero.

2) 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:

Models 1, 2 and 3 are nested relative to model 4. Models 1, 2 and 3 can be specified by setting parameters of model 4 to zero.

3) 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. Model 1 can be specified by setting parameters of model 3 to zero.

4) 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.

```
# create the model
model.new <- lm(y~x1+x2+x3+x1:x2+x1:x3+x2:x3, data=Q5.var)
# print model summary
summary(model.new)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3, data = Q5.var)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.639  -35.007    0.533   31.164  138.521
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 177.8768    32.8473   5.415 1.81e-07 ***
## x1            3.1001     2.4367   1.272   0.2048
## x2            4.5158     2.5724   1.756   0.0808 .
## x3            4.2968     2.2056   1.948   0.0528 .
## x1:x2         0.3086     0.1621   1.904   0.0584 .
## x1:x3        -0.1070     0.1401  -0.764   0.4459
## x2:x3        -0.1445     0.1546  -0.934   0.3513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.54 on 193 degrees of freedom
## Multiple R-squared:  0.373,  Adjusted R-squared:  0.3535
## F-statistic: 19.14 on 6 and 193 DF,  p-value: < 2.2e-16
```

Q5, Part 2: Nested model testing - 5 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.

```
# conduct nested model test
m2.vs.new <- anova(model.2, model.new)
# print results
m2.vs.new
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    197 528946
## 2    193 492946  4     36000 3.5237 0.008412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# conduct the nested model test
new.vs.m5 <- anova(model.new, model.5)
# print the results
new.vs.m5
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3
## 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    193 492946
## 2    192 488971  1    3974.5 1.5606 0.2131
```

6) 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. With a $p - value \approx 0.008$ we can reject the null hypothesis and accept the alternate, that model.new accounts for significantly more variability in the outcome than model.2.

7) 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 the new model. We fail to reject the null hypothesis. Model.5 does not account for significantly more variability in the outcome than model.new.

Note:

In practice use of model.5 and it's additional predictors might depend upon domain knowledge, the goal of the analysis, etc.

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 int
o 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.

```
# change hard to categorical variable
lead$hard <- as.factor(lead$hard)
# check the data structure
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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 1 2 1 1 1 ...
##  $ danger: int  1 1 1 1 0 1 1 1 0 1 ...
```

```
# fit the model
danger.model <- glm(danger ~ age+loc+hard,
                    data = lead,family="binomial")
```

```
#print the model summary
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
## hard1        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 were significantly different from zero? There is at least one.

Your answer here:

The results of the analysis show that the following predictor coefficients were significant:

*age

2) Of the statistically significant predictor/s you identified in the first 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 presences of a dangerous level of lead is *less likely* as the value of the predictor increases?

More likely as values of predictor/s increase/s (your answer here):

Of the statistically significant predictor coefficients identified by the analysis, “age” (the age of the home) is the only predictor coefficient that indicates an increase in the probability of dangerous lead levels as the predictor increases.

Less likely as values of predictor/s increase/s (your answer here):

No statistically significant predictors indicated a decrease in likelihood of dangerous lead levels as the predictor increases.