COMP 4442 Midterm, Winter 2024

Michael_Ghattas_Exam_MT

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

# Load required libraries
library(leaps)        # For regression and model selection
library(ggplot2)      # For visualization
library(ggpubr)       # For Q-Q plots
library(dplyr)        # For data manipulation
```

```
##
## 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(car)          # For hypothesis testing
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
# Load datasets for the exam
q1_data <- read.csv("Q1data.csv")
q2_data <- read.csv("Q2data.csv")
q3_data <- read.csv("Q3data.csv")
q4_data <- read.csv("Q4data.csv")
q5_data <- read.csv("Q5data.csv")
q6_data <- read.csv("Q6data.csv")

# Display first few rows of each dataset for initial inspection
head(q1_data)
```

```
##     miles form
## 1 81171.1    a
## 2 81418.9    a
## 3 79781.4    a
## 4 79093.0    a
## 5 81211.8    a
## 6 76364.1    a
```

```
head(q2_data)
```

```
##   lowesterol lipidown post.ldl
## 1         No       No   171.99
## 2         No       No   170.50
## 3         No       No   172.60
## 4         No       No   174.90
## 5         No       No   166.27
## 6         No       No   173.96
```

```
head(q3_data)
```

```
##      loss retail area traffic
## 1 1870.22      0 1681     373
## 2 1531.97      1 1448     149
## 3 1437.93      1 1322     217
## 4 1537.32      1 1356     347
## 5 1627.61      1 1429     384
## 6 1690.62      1 1568     228
```

```
head(q4_data)
```

```
##          y x1 x2 x3 x4 x5 x6 x7 x8
## 1 237.7532  7  6  6 15  8  9  7 10
## 2 325.3688 15 13  1  8 17 11 11 10
## 3 259.7192 11 11 11  0  7 10 10 13
## 4 366.6531  9  7  8 12 16 14 19 15
## 5 296.5794 14 10 11 12  7 14  7 12
## 6 367.2463 12 14 12 12 13 15 12 13
```

```
head(q5_data)
```

```
##          y x1 x2 x3 x4 x5 x6 x7 x8
## 1 237.7532  7  6  6 15  8  9  7 10
## 2 325.3688 15 13  1  8 17 11 11 10
## 3 259.7192 11 11 11  0  7 10 10 13
## 4 366.6531  9  7  8 12 16 14 19 15
## 5 296.5794 14 10 11 12  7 14  7 12
## 6 367.2463 12 14 12 12 13 15 12 13
```

```
head(q6_data)
```

```
##   age      loc hard danger
## 1  79    rural    0      1
## 2  77    urban    0      1
## 3  87    rural    1      1
## 4  40    rural    1      1
## 5  53 suburban    0      0
## 6  81 suburban    0      1
```

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.

```
# Load the data
tires <- read.csv("Q1data.csv", header = TRUE, sep = ",")

# Convert 'form' to a factor
tires$form <- as.factor(tires$form)

# Display structure of the data
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)

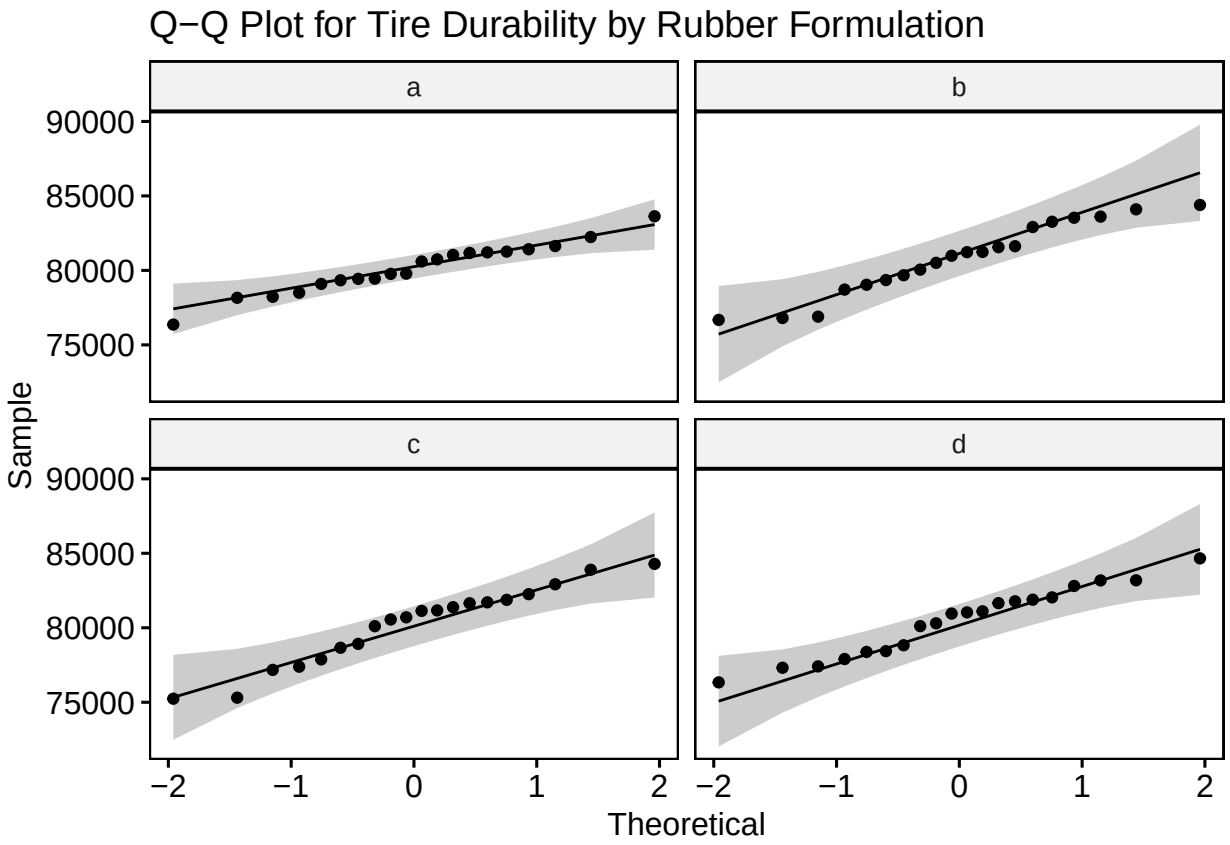
You will assess the assumption of normality in two ways: quantitatively and visually. In this first code chunk, please conduct an appropriate *quantitative* assessment of the normality assumption and display the results.

```r
# Quantitative assessment of normality using the Shapiro-Wilk test
shapiro_results <- by(tires$miles, tires$form, shapiro.test)

# Display the results
shapiro_results
```

```
## tires$form: a
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.97868, p-value = 0.916
##
## ------------------------------------------------------------
## tires$form: b
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.94588, p-value = 0.3088
##
## ------------------------------------------------------------
## tires$form: c
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.9488, p-value = 0.3493
##
## ------------------------------------------------------------
## tires$form: d
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.96382, p-value = 0.6225
```

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

```r
# Visual assessment of normality using Q-Q plots for each formulation
ggqqplot(tires, x = "miles", facet.by = "form", title = "Q-Q Plot for Tire Durability by Rubber Formula
```



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

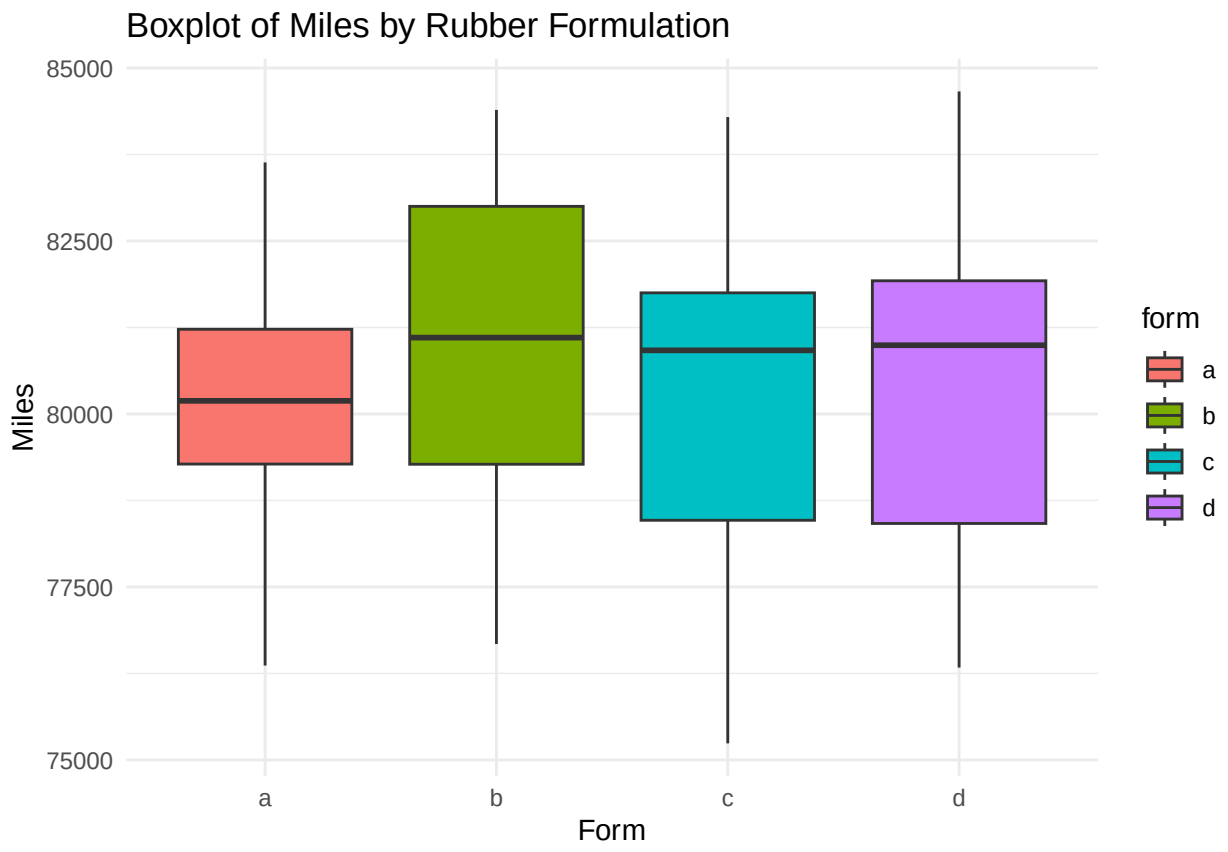
You will assess the assumption of equality of variances in two ways: quantitatively and visually. In this first code chunk, please conduct an appropriate *quantitative* assessment of the equality of variances assumption and display the results.

```
# Quantitative assessment of equality of variances using Levene's Test
levene_test_result <- leveneTest(miles ~ form, data = tires)

# Display the result of Levene's Test
levene_test_result
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.9644  0.414
##       76
```

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

```r
# Visual assessment of equality of variances using a boxplot
ggplot(tires, aes(x = form, y = miles, fill = form)) +
  geom_boxplot() +
  labs(title = "Boxplot of Miles by Rubber Formulation", x = "Form", y = "Miles") +
  theme_minimal()
```



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.

```r
# Fit the ANOVA model
tires.aov <- aov(miles ~ form, data = tires)

# Display the results of the 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 (4 points)

A) What is the null hypothesis being tested by the ANOVA you conducted? Be specific.

Your answer here: The null hypothesis (H0) for the ANOVA is that there is no difference in mean tire durability (measured in miles) across the four different rubber formulations. In other words, the mean tire durability for all formulations is the same.

B) What is the alternative hypothesis being tested by the ANOVA you conducted? Be specific.

Your answer here: The alternative hypothesis (H1) is that at least one rubber formulation has a different mean tire durability compared to the others. In other words, not all means are equal.

C) Based on the results of your analysis, do you *reject* or *fail to reject* this null hypothesis?

Your answer here: Based on the ANOVA results, the p-value is 0.793, which is greater than the standard significance level of 0.05. Therefore, we fail to reject the null hypothesis. This means that there is no statistically significant evidence to suggest that the mean tire durability differs across the four rubber formulations.

D) Which of the following statements is best supported by the result of this analysis? Statement 1: There is evidence to suggest that at least one tire formulation has different mean durability than the other formulation Statement 2: There is evidence to suggest that all tire formulations have different mean durability Statement 3: There is no evidence to suggest that there are differences in mean durability across tire formulations

Your answer here: There is no evidence to suggest that there are differences in mean durability across tire formulations, based on the high p-value (0.793) obtained from the ANOVA. (Statement 3)

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.

```
# Load the data
drugs <- read.csv("Q2data.csv", header = TRUE, sep = ",")

# Convert 'lowesterol' and 'lipidown' to factors
drugs$lowesterol <- as.factor(drugs$lowesterol)
drugs$lipidown <- as.factor(drugs$lipidown)

# Display the structure of the dataset
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.

```
# Fit the two-way ANOVA model with interaction
drugs.aov <- aov(post.ldl ~ lowesterol * lipidown, data = drugs)

# Display the results of the ANOVA
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.

A) Is the main effect of Lowesterol significant?

Your answer here (yes/no): No.

B) Is the main effect of Lipidown significant?

Your answer here (yes/no): Yes.

C) 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 district 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 district. They have four pieces of information 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-oriented 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.

```r
# Load the data
mall <- read.csv("Q3data.csv", header = TRUE, sep = ",")

# Convert 'retail' to a factor
mall$retail <- as.factor(mall$retail)

# Display the structure of the dataset
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 and display the results. Do not include interaction terms or polynomial terms, and do not apply any transformations. Outcome: loss. Predictors: retail, area, and traffic.

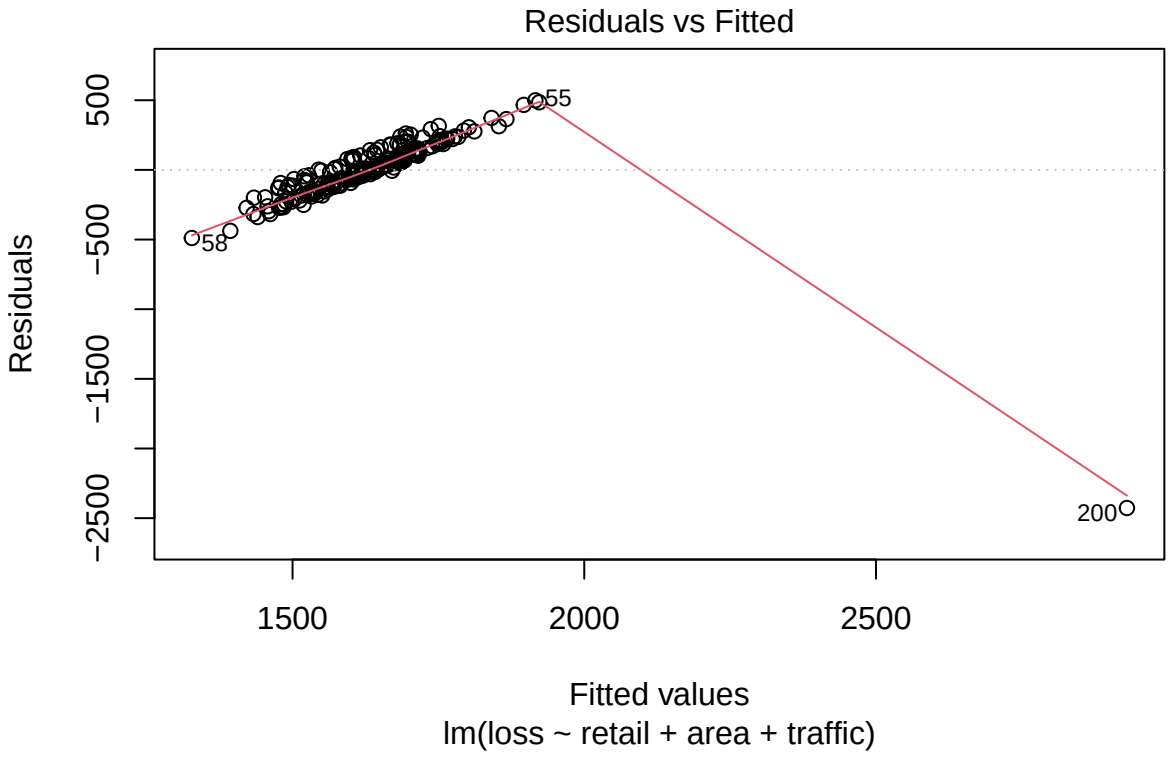
```r
# Fit the multiple regression model
mall.reg <- lm(loss ~ retail + area + traffic, data = mall)

# Display the results of the regression
summary(mall.reg)
```

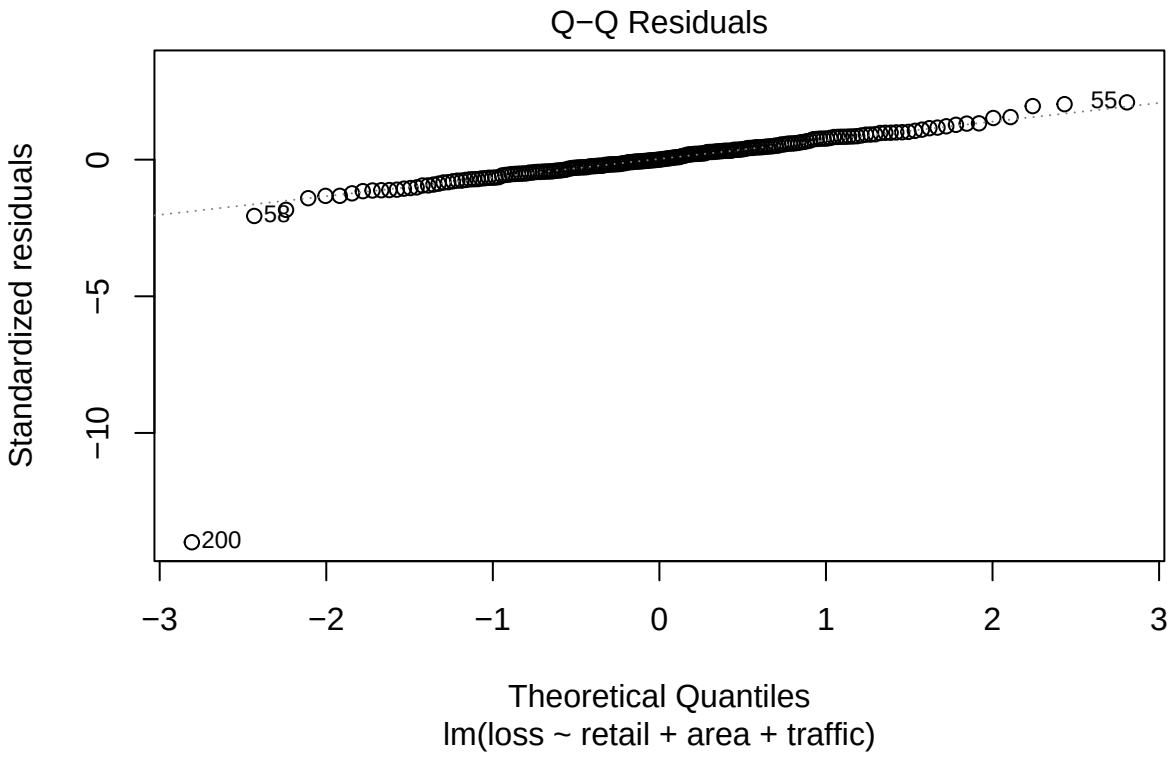
```
##
## Call:
## lm(formula = loss ~ retail + area + traffic, data = mall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2427.2  -103.7     0.8   117.4   499.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 966.02901   95.02144  10.166  < 2e-16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

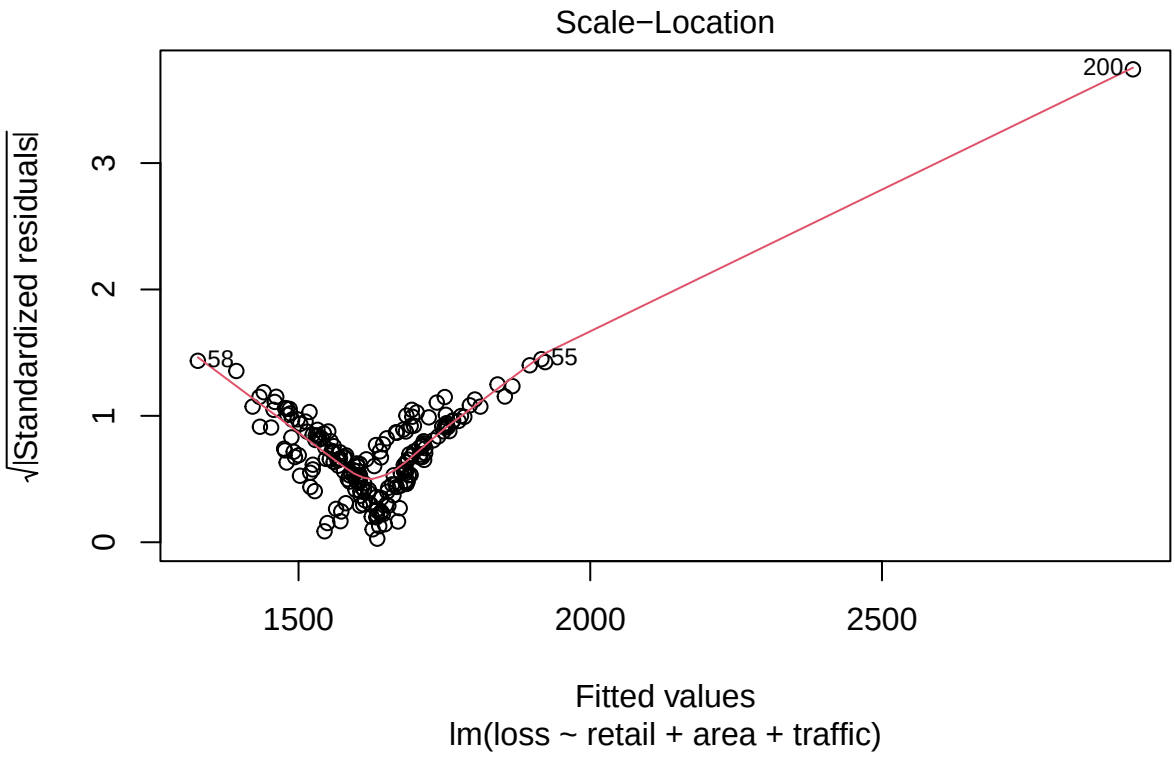
Q3, Part 2: Checking diagnostic plots (4 points)

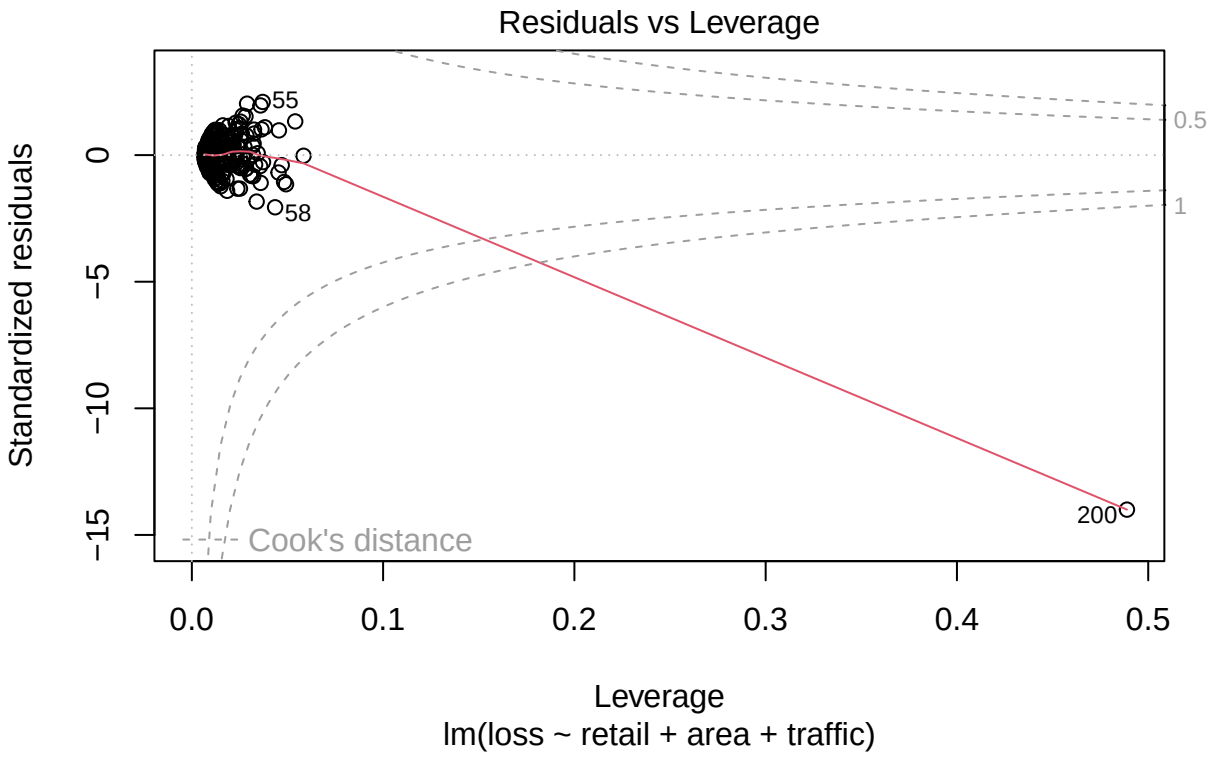
Please display the diagnostic plots for the model you fit in the previous part of this question and answer the question below:

```
# Plot diagnostic plots for the regression model
plot(mall.reg)
```




Q−Q Residuals
Standardized residuals
0
−5
−10
55
58
200
−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)



A) What is the most obvious problem that all of the diagnostic plots for this model share? Be specific.

Your answer here: The most obvious problem across all the diagnostic plots is the presence of outliers and influential points. Specifically, data points 200, 58, and 55 show up prominently across several diagnostic plots (Residuals vs Fitted, Scale-Location, Residuals vs Leverage), indicating that they are potentially influential observations. Additionally, the Scale-Location plot suggests heteroscedasticity (non-constant variance), as the spread of residuals increases with fitted values, which violates one of the key assumptions of linear regression.

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

Your answer here: A good solution to the problem of influential points is to investigate and potentially remove, transform these outliers, and consider whether they should be removed or treated differently.

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.

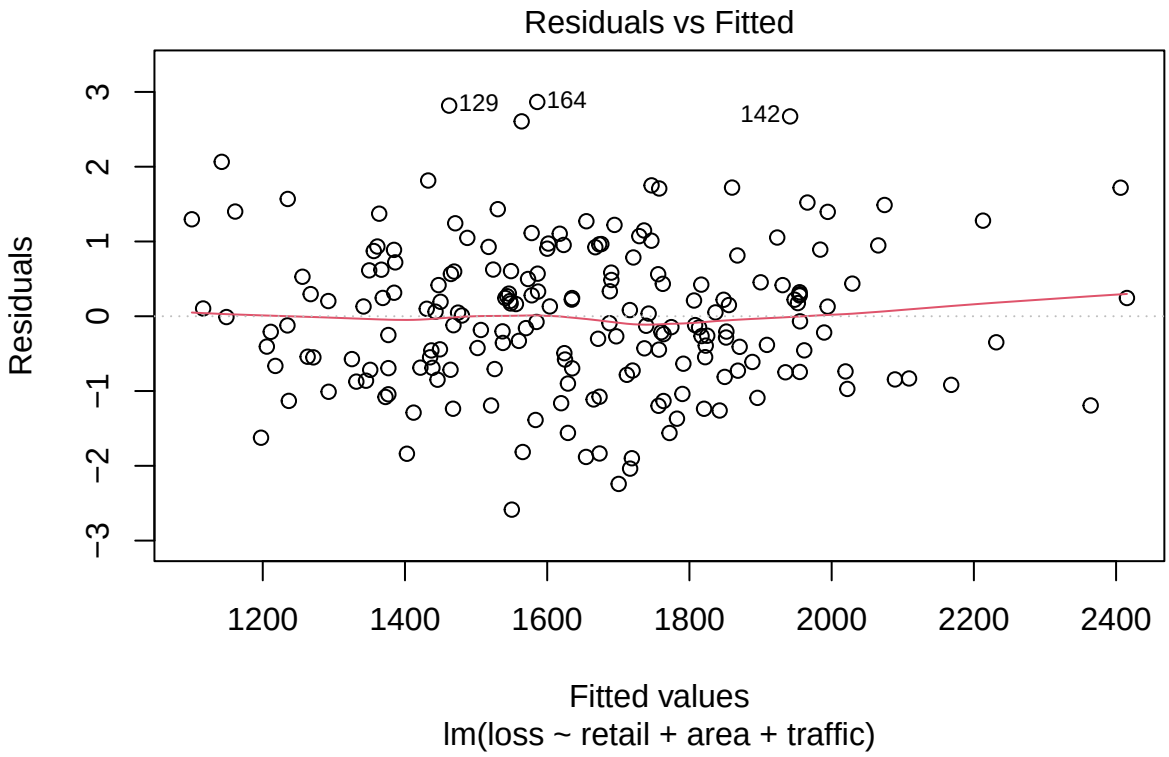
```
# Identify and remove the influential outliers (points 200, 58, 182)
outliers <- c(200, 58, 182)
mall_clean <- mall[-outliers, ]
```

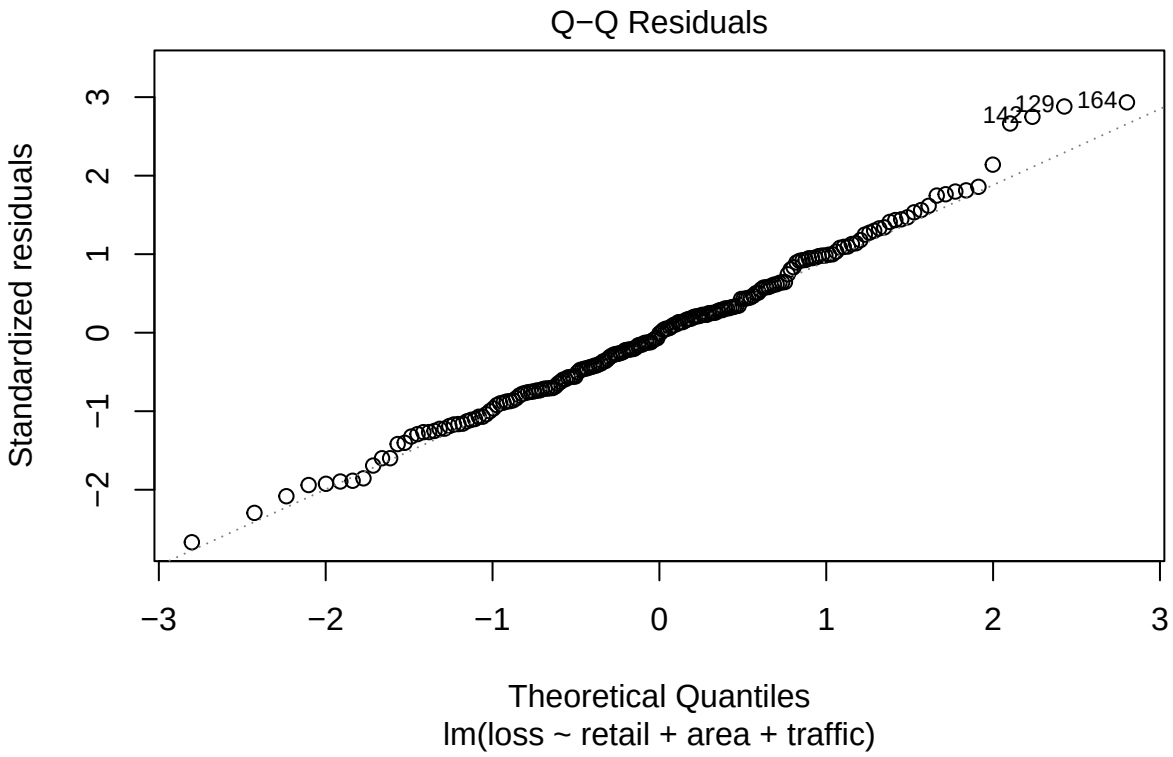
```r
# Re-fit the regression model with the log-transformed dependent variable (after removing outliers)
mall.reg.change <- lm(loss ~ retail + area + traffic, data = mall_clean)

# Display the results of the updated regression model
summary(mall.reg.change)
```

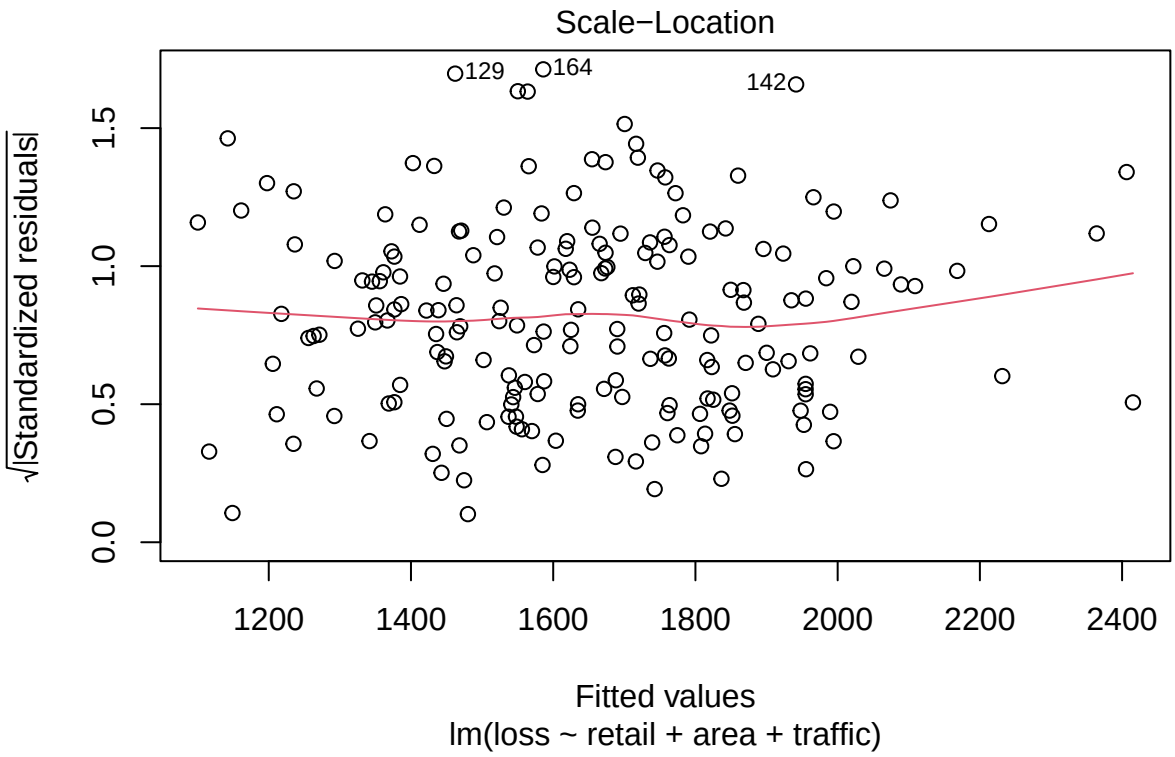
```
##
## Call:
## lm(formula = loss ~ retail + area + traffic, data = mall_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58596 -0.69377 -0.01069  0.58337  2.86567
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 3.5395202  0.4927698    7.183 1.44e-11 ***
## retail1     4.9437873  0.1649149   29.978  < 2e-16 ***
## area        0.9996377  0.0002782 3592.832  < 2e-16 ***
## traffic     0.5005326  0.0007162  698.890  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9822 on 193 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.439e+06 on 3 and 193 DF,  p-value: < 2.2e-16
```

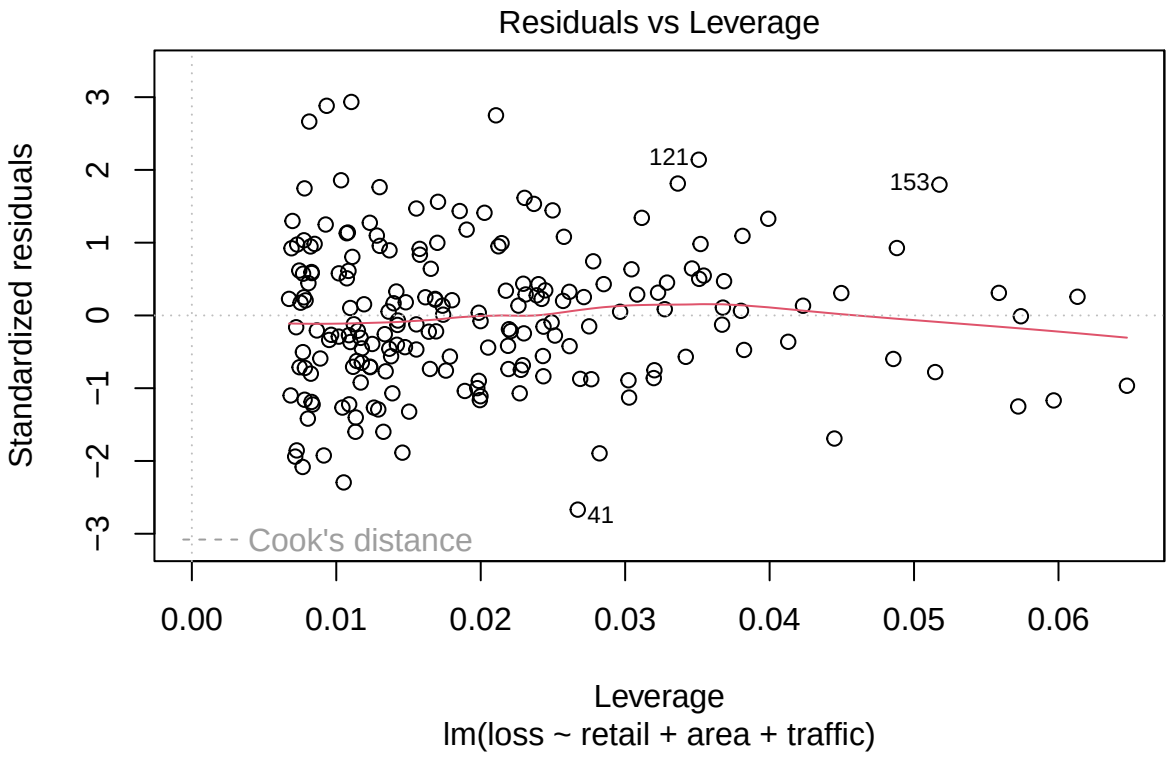
Next, display the updated diagnostic plots and answer the question below

```r
# Display the updated diagnostic plots for the re-fitted model
plot(mall.reg.change)
```


Residuals vs Fitted
129
164
142
Residuals
3
2
1
0
−1
−2
−3
1200
1400
1600
1800
2000
2200
2400
Fitted values
lm(loss ~ retail + area + traffic)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(loss ~ retail + area + traffic)
143
129
164


Scale−Location
129
164
142
√|Standardized residuals|
0.0
0.5
1.0
1.5
1200
1400
1600
1800
2000
2200
2400
Fitted values
lm(loss ~ retail + area + traffic)



C) Did your solution to the problem you identified in Q3, Part 2 greatly improve the diagnostic plots of the model? (Hint: If your answer isn't 'yes', try a different solution.)

Your answer here: Yes, removing the influential outliers (points 200, 58, and 182) greatly improved the diagnostic plots. The spread of residuals now appears much more evenly distributed, with no clear patterns or large deviations from linearity. This indicates that the model now better meets the assumption of homoscedasticity. The residuals in the Q-Q plot follow the theoretical normal distribution line more closely, indicating that the normality assumption is better met after removing the outliers. The spread of residuals is now much more consistent, indicating that the heteroscedasticity issue has been largely addressed. No highly influential points are evident, and the Cook's distance is no longer showing any extreme values, indicating that the model is not being unduly influenced by any single observation. Thus, the solution has significantly improved the model's fit and its adherence to the assumptions of linear regression.

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

D) Select the statement that is a correct interpretation of the intercept: Statement 1: The predicted loss for a retail-oriented store with an area of zero square feet and average weekly traffic of zero people is 3.60 USD. Statement 2: The predicted loss for any store is 3.60 USD. Statement 3: The predicted loss for a store with zero revenue from retail sales, an area of zero square feet, and average weekly traffic of zero people is 3.60 USD. Statement 4: The predicted loss for a service-oriented store with an area of zero square feet and average weekly traffic of zero people is 3.60 USD. Statement 5: The predicted loss for a store with the mean revenue from retail sales, the mean area, and the mean average weekly traffic is 3.60 USD.

Your answer here: The predicted loss for a service-oriented store with an area of zero square feet and average weekly traffic of zero people is 3.60 USD. Statement 4, because the intercept represents the predicted loss for a service-oriented store (retail = 0), with zero area and zero traffic. Meaning, the intercept is 3.539, meaning that when all predictor variables (retail, area, and traffic) are equal to zero, the predicted loss is 3.54.

E) Select the statement that is a correct interpretation of the coefficient associated with retail: Statement 1: Holding area and average weekly traffic constant, the predicted loss for a service-oriented store is 4.95 USD higher than for a retail-oriented store. Statement 2: The predicted loss for a retail-oriented store is 4.95 USD. Statement 3: Holding area and average weekly traffic constant, the predicted loss for a retail-oriented store is 4.95 USD higher than for a service-oriented store. Statement 4: Holding area and average weekly traffic constant, the predicted loss for a retail-oriented store increases by 4.95 USD for each additional dollar of revenue from retail sales. Statement 5: The predicted loss for a retail-oriented store is 8.55 USD.

Your answer here: Holding area and average weekly traffic constant, the predicted loss for a retail-oriented store is 4.95 USD higher than for a service-oriented store. Statement 3, because this correctly interprets the coefficient as the difference in predicted loss between retail and service-oriented stores, holding other variables constant. i.e. The coefficient for retail1 is 4.944, meaning that holding area and traffic constant, the predicted loss for a retail-oriented store (retail = 1) is 4.95 USD higher than for a service-oriented store (retail = 0).

F) 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? Show your work in the code chunk and type your answer below.

```
# Predicted loss for a non-retail store (retail = 0) with area = 1000 and traffic = 200
intercept <- 3.5395202
beta_retail <- 4.9437873
beta_area <- 0.9996377
beta_traffic <- 0.5005326

retail_value <- 0  # Non-retail store
area_value <- 1000  # 1000 square feet
traffic_value <- 200  # 200 average weekly traffic

# Calculate predicted loss
predicted_loss <- intercept + (beta_retail * retail_value) + (beta_area * area_value) + (beta_traffic *
predicted_loss
```

```
## [1] 1103.284
```

Your answer here: The predicted amount of loss for a non-retail store that has an area of 1000 square feet and average weekly traffic of 200 is $1103.28.

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.

```r
# Load the dataset
many.var <- read.csv("Q4data.csv", header = TRUE, sep = ",")

# Display the structure of the dataset
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.

```r
# Start with an empty model
null_model <- lm(y ~ 1, data = many.var)

# Define the full model with all predictors
full_model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = many.var)

# Perform forward selection
forward.model <- step(null_model, scope = list(lower = null_model, upper = full_model), direction = "fo
```

```
## 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
```

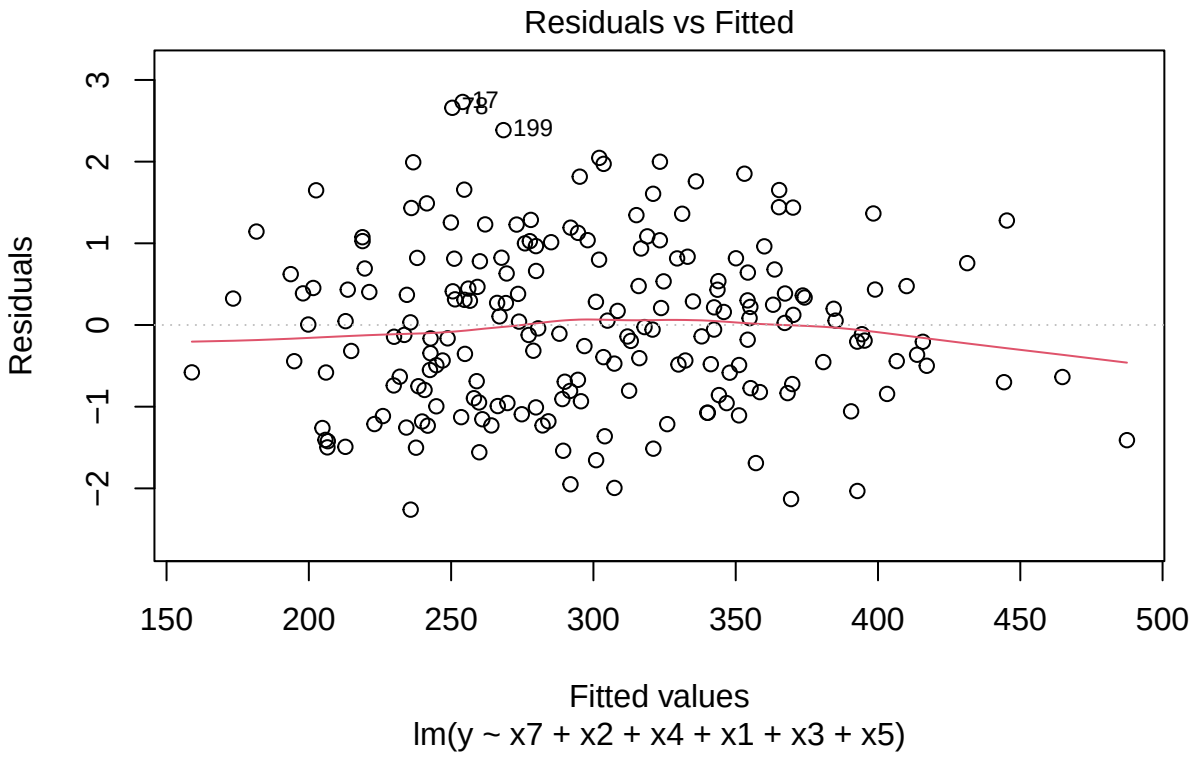
```
## <none>              201.87 15.856
## + x8    1   0.67420 201.19 17.187
## + x6    1   0.42727 201.44 17.432
```

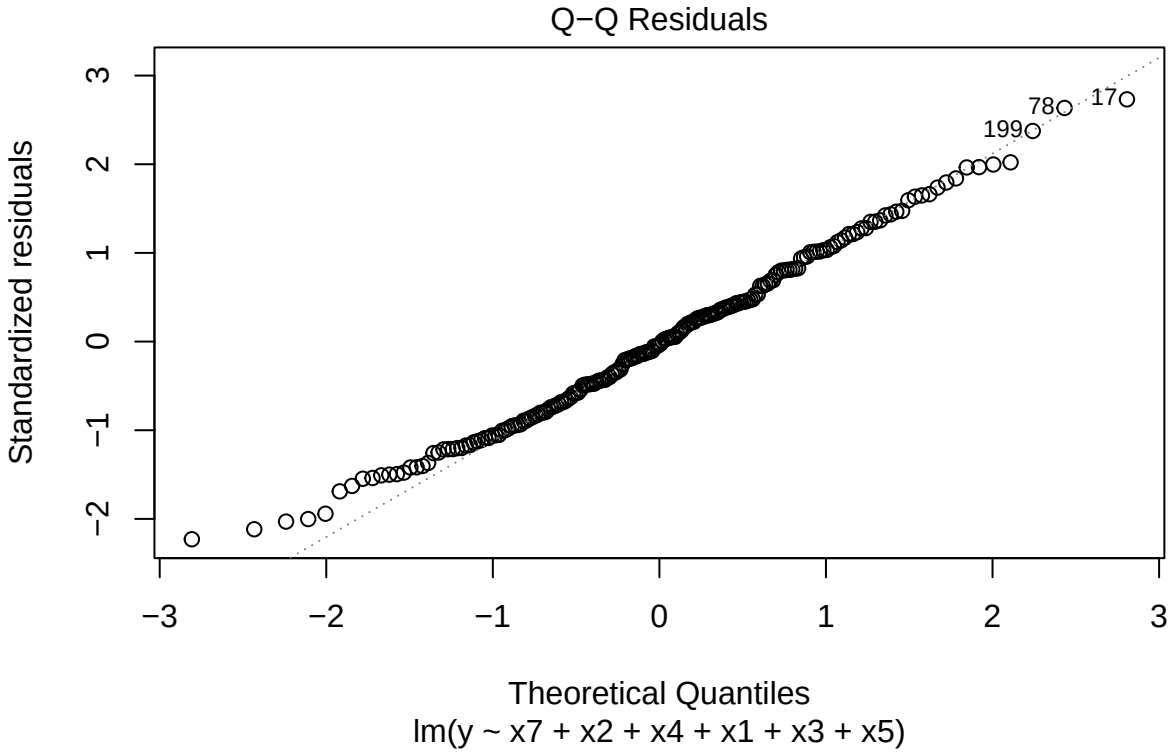
The model selected by forward selection:

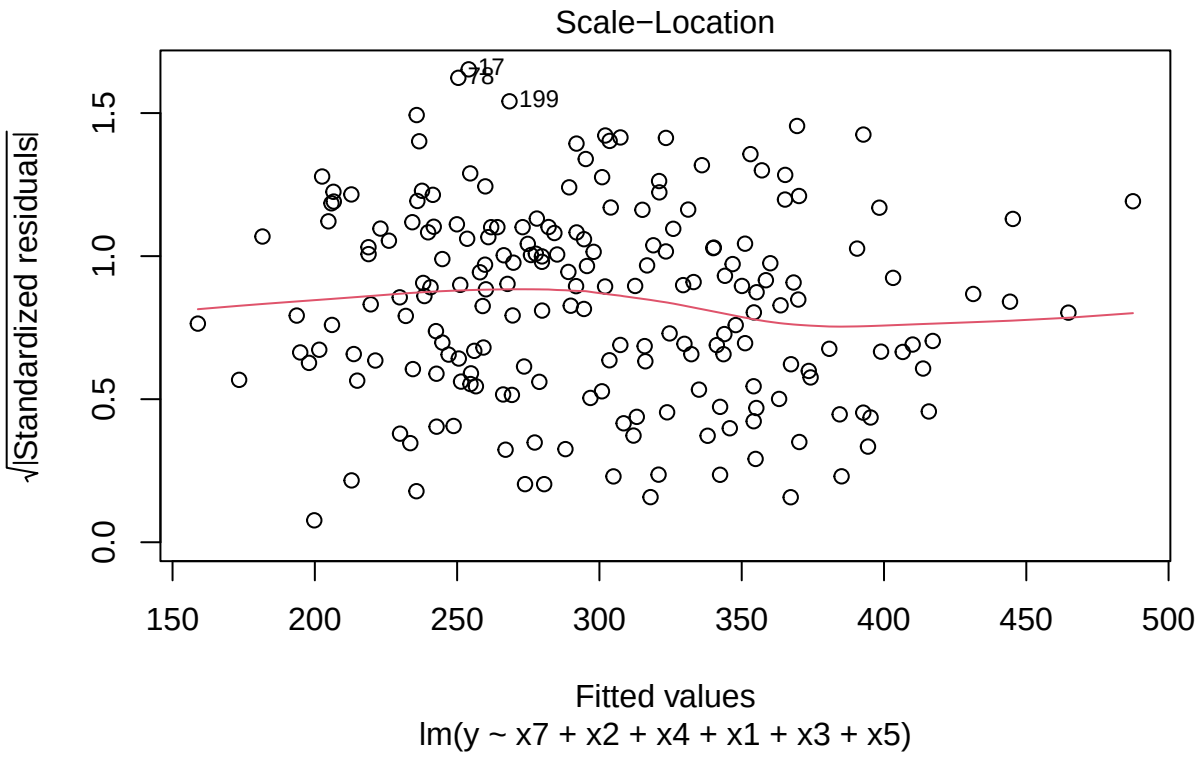
```
# Display the model selected using forward selection
summary(forward.model)
```

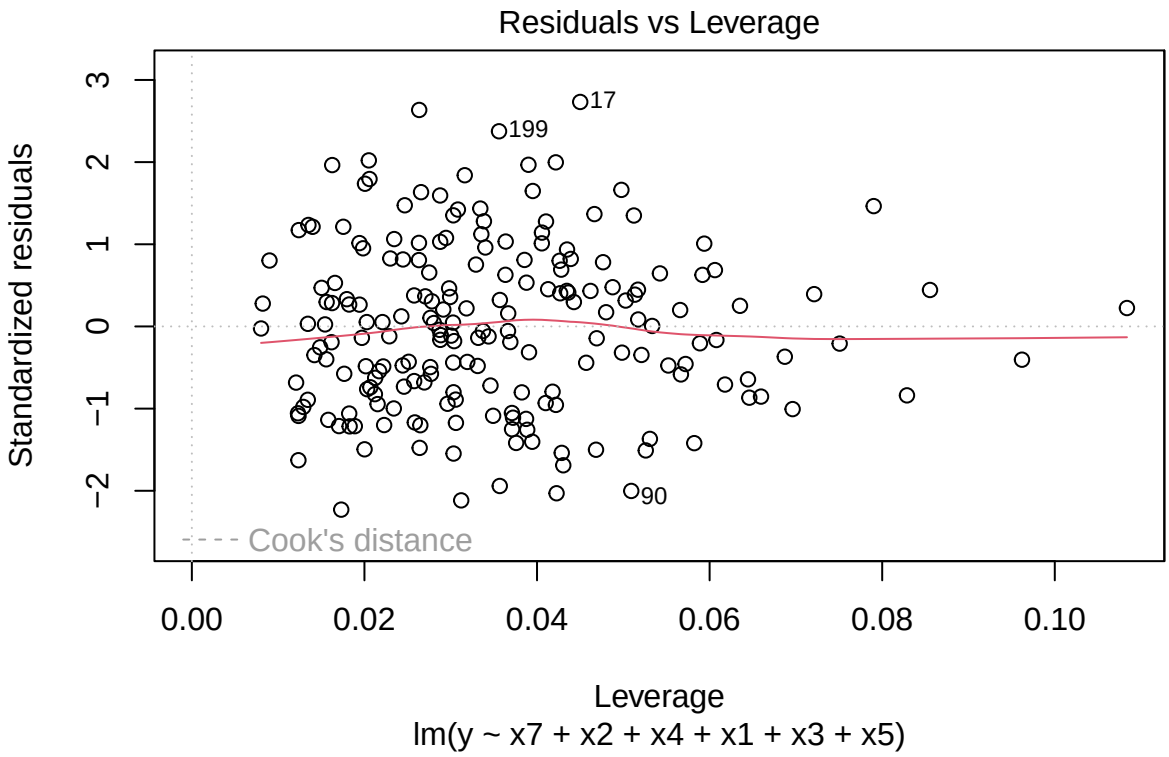
```
##
## Call:
## lm(formula = y ~ x7 + x2 + x4 + x1 + x3 + x5, data = many.var)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.26021 -0.77847 -0.03343  0.68273  2.73044
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.20153    0.39170  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
```

```
plot(forward.model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x7 + x2 + x4 + x1 + x3 + x5)
17
78
199
−2
−1
0
1
2
3
150
200
250
300
350
400
450
500


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(y ~ x7 + x2 + x4 + x1 + x3 + x5)
17
78
199


Scale−Location
17
78
199
√|Standardized residuals|
0.0
0.5
1.0
1.5
150
200
250
300
350
400
450
500
Fitted values
lm(y ~ x7 + x2 + x4 + x1 + x3 + x5)



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.

```r
# Start with the full model containing all predictors
full_model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = many.var)

# Perform backward selection
backward.model <- step(full_model, 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
```

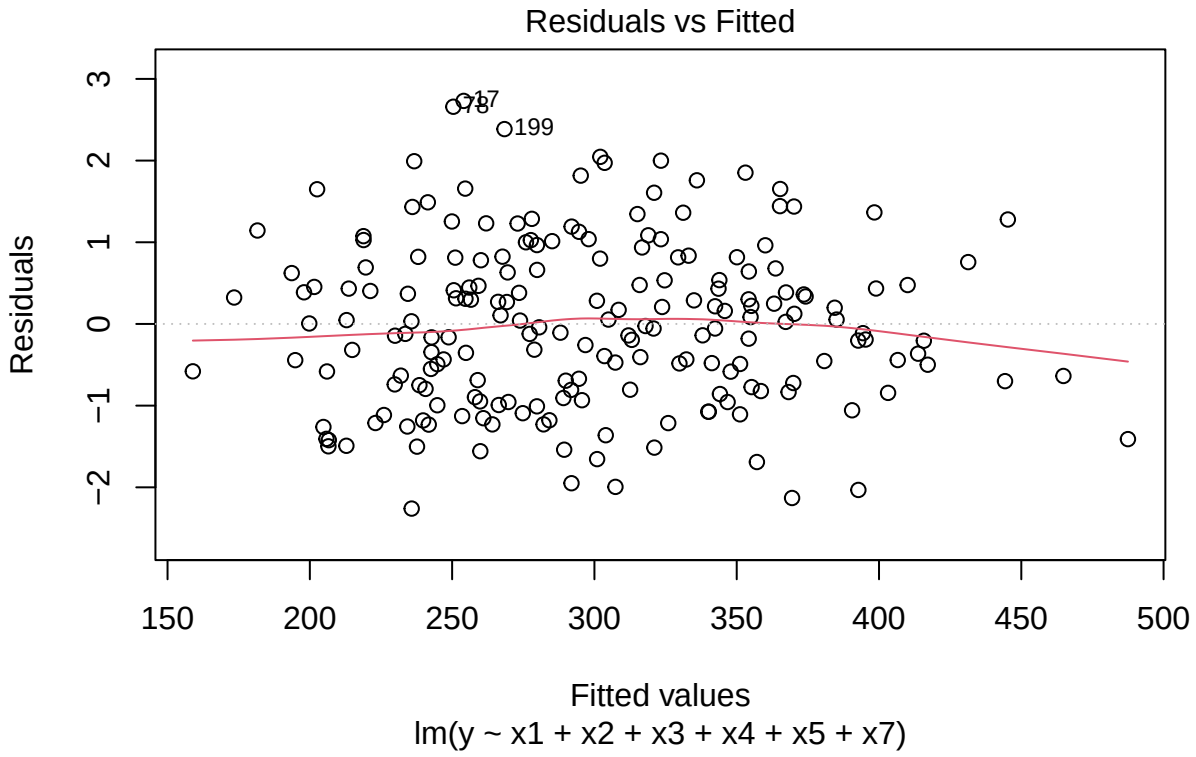
The model selected by backward selection:

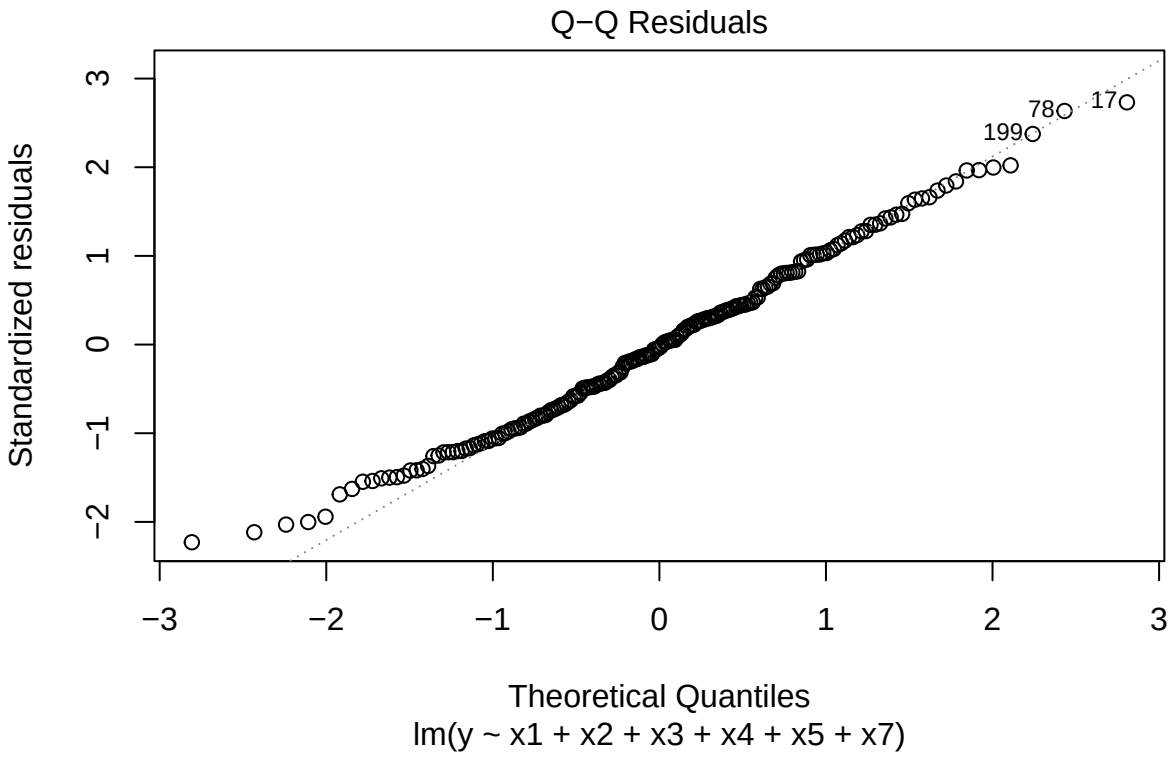
```
# Display the selected model from backward selection
summary(backward.model)
```

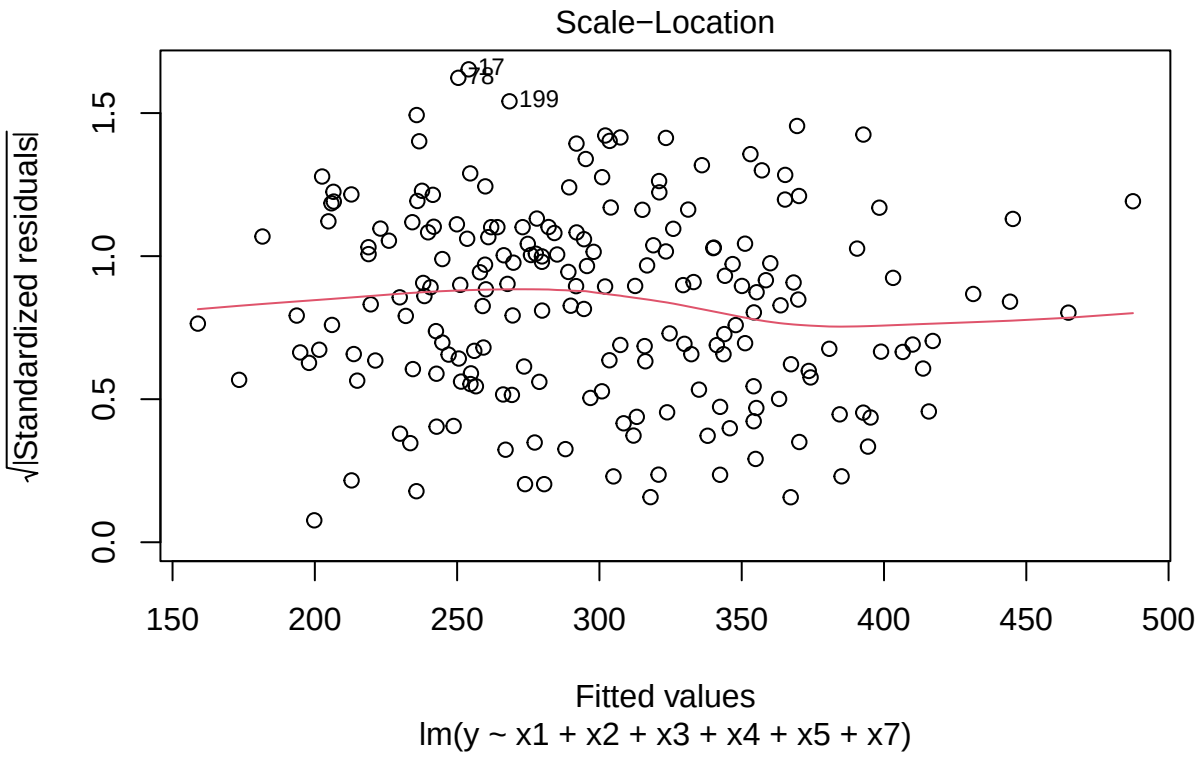
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x7, data = many.var)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.26021 -0.77847 -0.03343  0.68273  2.73044
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.20153    0.39170  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
```

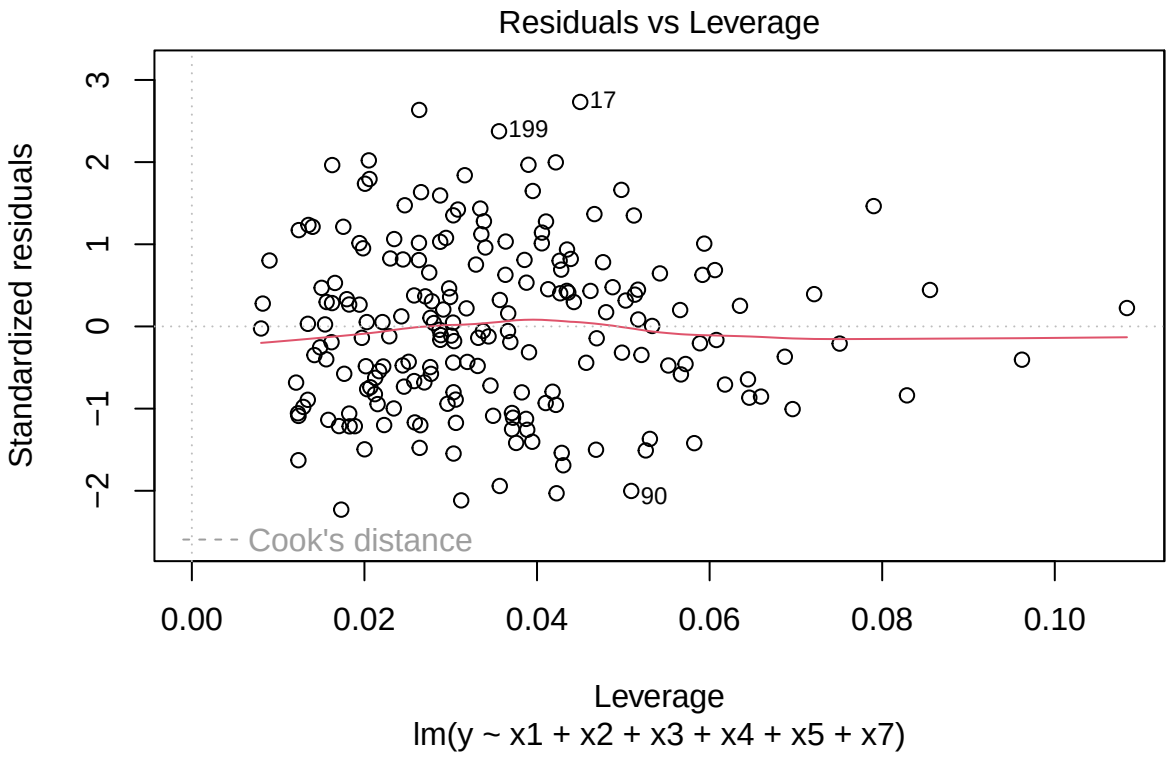
```
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 1.252e+05 on 6 and 193 DF,  p-value: < 2.2e-16
```

```
plot(backward.model)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(y ~ x1 + x2 + x3 + x4 + x5 + x7)
17
78
199


Scale−Location
17
78
199
√|Standardized residuals|
0.0
0.5
1.0
1.5
150
200
250
300
350
400
450
500
Fitted values
lm(y ~ x1 + x2 + x3 + x4 + x5 + x7)



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. Recall from the best subsets demo that you need to include all of the following for full credit: 1) Correctly set up the inputs into the regsubsets() function 2) Correctly run the regsubsets() function 3) Display either the star table or T/F table. 4) Display either a list of model BICs or a scatterplot of the model BICs. 5) Display the output of the model selected using best subsets selection.

```
# Set up inputs into regsubsets and fitting models using all the predictors x1 to x8
best_subset_fit <- regsubsets(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = many.var, nvmax = 8)
```

```
# Run the regsubsets function to evaluate all possible models
best_subset_fit <- regsubsets(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = many.var, nvmax = 8)
```

```
# Display the summary of the regsubsets output i.e. True/False table for which predictors are included
best_subset_summary <- summary(best_subset_fit)

# Display the T/F table i.e. the star table
best_subset_summary$outmat
```

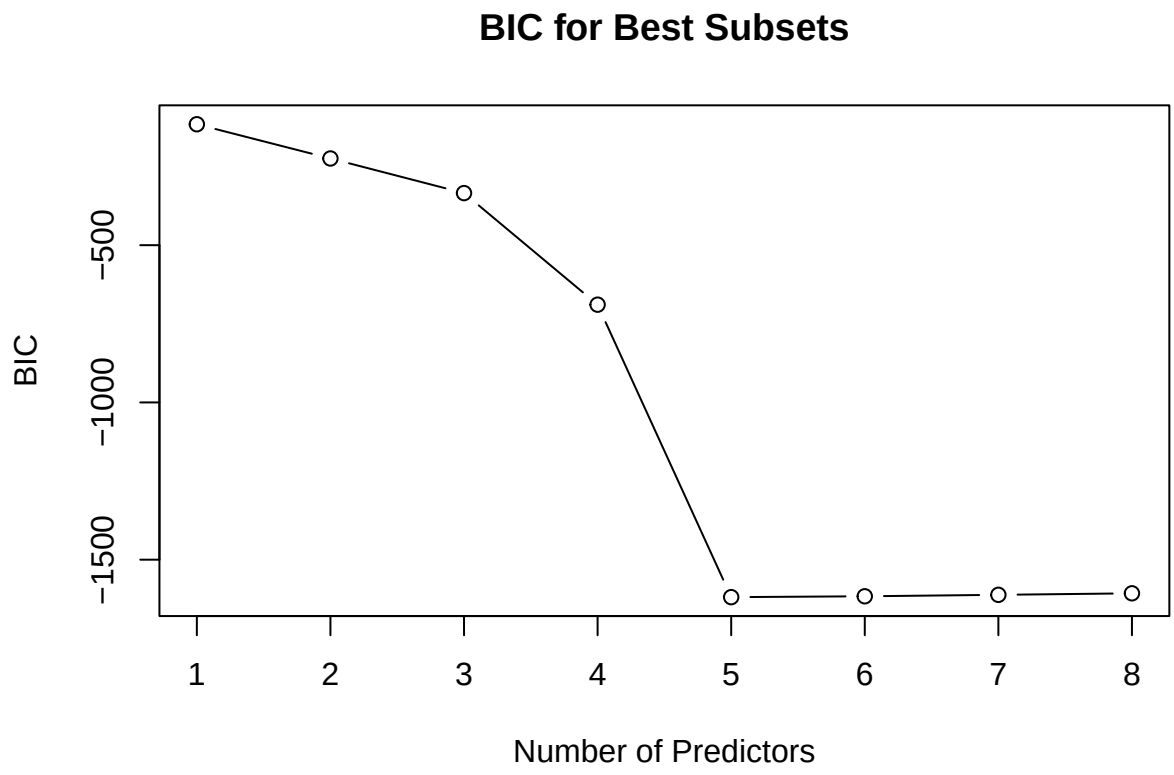
```
##          x1  x2  x3  x4  x5  x6  x7  x8
## 1  ( 1 ) " " " " " " " " " " " " "*" " "
```

```
## 2  ( 1 ) " " "*" " " " " " " " " " " "*" " "
## 3  ( 1 ) " " "*" " " "*" " " " " " " "*" " "
## 4  ( 1 ) "*" "*" " " "*" " " " " " " "*" " "
## 5  ( 1 ) "*" "*" "*" "*" " " " " " " "*" " "
## 6  ( 1 ) "*" "*" "*" "*" "*" " " " " "*" " "
## 7  ( 1 ) "*" "*" "*" "*" "*" " " " " "*" "*"
## 8  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

```r
# Display the BIC values for each model
best_subset_summary$bic
```

```
## [1]  -115.5419  -224.3521  -334.4118  -689.1580 -1619.0824 -1616.3898 -1611.7605
## [8] -1606.9375
```

```r
# Create a scatterplot of BICs
plot(best_subset_summary$bic, type = "b", main = "BIC for Best Subsets", xlab = "Number of Predictors",
```




```r
# Identify the best model and extract the coefficients of the best model selected by BIC
best_bic_model <- which.min(best_subset_summary$bic)

# Display the coefficients of the selected model
coef(best_subset_fit, best_bic_model)
```

```
## (Intercept)          x1          x2          x3          x4          x7
##    4.368135    5.018899    7.031190    2.007329    4.974010   10.028653
```

These are the selected variables based on minimizing the BIC. The BIC plot clearly shows that the model with 5 predictors minimizes the BIC value, indicating the best model according to Bayesian Information Criterion. These are the best subsets selection identified the following model with the predictors. Thus, this model offers the best balance between model fit and complexity, as judged by the BIC.

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.

```
# Load the dataset
Q5.var <- read.csv("Q5data.csv", header = TRUE, sep = ",")

# Display the structure of the dataset
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.

```
# Fit the models
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)

# Display the summaries of the models
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
```

A) 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 all nested within Model 5. A model is nested if it can be derived by removing one or more terms (predictors or interactions) from the full model. Since Model 5 includes all the terms and interaction terms involving x1, x2, and x3, it is the full model. Each of the other models is a simplified version, lacking some terms or interactions that are included in Model 5.

B) 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 all nested within Model 4. Model 4 includes the main effects of x1, x2, and x3, while Models 1, 2, and 3 are reduced models containing only subsets of these predictors. For instance, Model 1 includes only x1, Model 2 includes x1 and x2, and Model 3 includes x1 and x3.

C) 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 within Model 3. Model 1 includes only x1, while Model 3 includes both x1 and x3. Model 2 is not nested within Model 3 because it includes x2, which is absent from Model 3.

D) 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.

```r
# A new model that is nested within Model 5 and contains Model 2 (x1 and x2)
model.new <- lm(y ~ x1 + x2 + x1:x2:x3, data = Q5.var)

# Display the summary of the new model
summary(model.new)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x1:x2:x3, data = Q5.var)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -131.583  -36.485    1.702   33.159  135.634
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.091e+02  1.213e+01  17.235  < 2e-16 ***
## x1          3.340e+00  9.546e-01   3.499 0.000579 ***
## x2          4.664e+00  9.158e-01   5.093 8.26e-07 ***
## x1:x2:x3    1.459e-02  6.038e-03   2.416 0.016592 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.19 on 196 degrees of freedom
## Multiple R-squared:  0.3467, Adjusted R-squared:  0.3367
## F-statistic: 34.67 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.

```
# Perform the nested model test between Model 2 and model.new
m2.vs.new <- anova(model.2, model.new)

# Display the result of the nested model comparison
m2.vs.new
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 + x2 + x1:x2:x3
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1    197 528946
## 2    196 513644  1     15302 5.839 0.01659 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Perform the nested model test between model.new and Model 5
new.vs.m5 <- anova(model.new, model.5)

# Display the result of the nested model comparison
new.vs.m5
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x1: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    196 513644
## 2    192 488971  4     24673 2.422 0.04974 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Your answer here: The test between Model 2 (which includes x1 and x2) and the new model (which includes x1, x2, and the interaction term x1:x2:x3) resulted in a significant p-value ($p = 0.01659$), indicating that the inclusion of the interaction term x1:x2:x3 significantly improves the model's fit. Since the interaction term provides a significant improvement, I would choose the new model over Model 2.

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

Your answer here: The test between the new model (which includes x1, x2, and the interaction term x1:x2:x3) and Model 5 (which includes all the main effects and interaction terms involving x1, x2, and x3) resulted in a p-value of 0.04974. This indicates that Model 5 provides a statistically significant improvement over the new model. Therefore, I would choose Model 5 because it significantly improves the fit compared to the new model.

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 (danger=0) or potentially dangerous if it had levels equal to or greater than 15 parts per billion (danger=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 (hard=0) or high (hard=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

```
# Load the dataset
lead <- read.csv("Q6data.csv", header = TRUE, sep = ",")

# Display the structure of the dataset
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. Do not apply any transformations or include interaction terms or polynomial terms. Be sure to display the results of the analysis.

```
# Convert 'loc' and 'hard' to factors
lead$loc <- as.factor(lead$loc)
lead$hard <- as.factor(lead$hard)

# Fit a logistic regression model with the binary outcome "danger" and predictors age, loc, and hard
danger.model <- glm(danger ~ age + loc + hard, data = lead, family = binomial)

# Display the summary of the logistic regression model
summary(danger.model)
```

```
##
## Call:
## glm(formula = danger ~ age + loc + hard, family = binomial, data = lead)
##
## 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

A) 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: The predictor age is significantly different from zero, with a p-value of 0.0044. The other predictors (locsuburban, locurban, and hard1) are not significantly different from zero.

B) Of the predictor/s you listed in sub-question A (the sub-question right before this one), 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 predictor/s (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): The predictor age has a positive coefficient (0.05024), which suggests that as the age of the house increases, the likelihood of having dangerous levels of lead increases.

Dangerous level of lead *less likely* as values of predictor/s increase/s (your answer here): N/A. None of the predictors indicate a decreased likelihood of dangerous lead levels. All significant predictors show an increased likelihood as their values increase. Also, while the negative intercept indicates that the baseline category has a less-than-even chance of dangerous lead levels, it's not significantly different from zero, meaning that we cannot confidently interpret it.

Note: If there are no predictors to list in one of these choices, write N/A.

End.