Homework 6

The purpose of this homework is to practice using simple linear regression models to analyze data. Please fill in the appropriate code and write answers to all questions in the answer sections, then submit a compiled pdf with your answers through Gradescope by 11pm on Sunday October 29th.

As always, if you need help with the homework, please attend the TA office hours which are listed on Canvas and/or ask questions on Ed Discussions. Also, if you have completed the homework, please help others out by answering questions on Ed Discussions, which will count toward your class participation grade.

Note: for all plots on this homework, please use base R graphics

Part 1: Simple linear regression

As you likely know, former president Donald Trump claims that the 2020 election was stolen from him due to election fraud. While the courts and most experts are confident that the 2020 election was fair, perhaps other US elections did indeed fail to truly reflect the will of the people. Let's explore this question using data from the 2000 US presidential election which was one of the closest US presidential elections.

In 2000, the US presidential election was between a Yale alumnus, George W. Bush who was the Republican candidate, and a Harvard alumnus Al Gore who was the Democratic candidate. There were also a number of "third-party" candidates such as Princeton alumnus Ralph Nader who was the Green Party candidate, and Georgetown alumnus Pat Buchanan who was the Reform Party candidate.

The code chunk below contains data from the 2000 election for the state of Florida in a data frame called `florida_data`. Each observational unit in this data frame contains information from one of the 67 counties in Florida including demographic information on each county as well as the votes received by each candidate in each county.

In the exercises below, you will use linear regression to look at the relationship between the votes that the Republican candidate *George W. Bush* received and the votes that the Reform candidate *Patrick Buchanan* received.

```
load('florida_votes_2000.rda')

dim(florida_data)
```

```
## [1] 67  6
```

Part 1.1 (6 points):

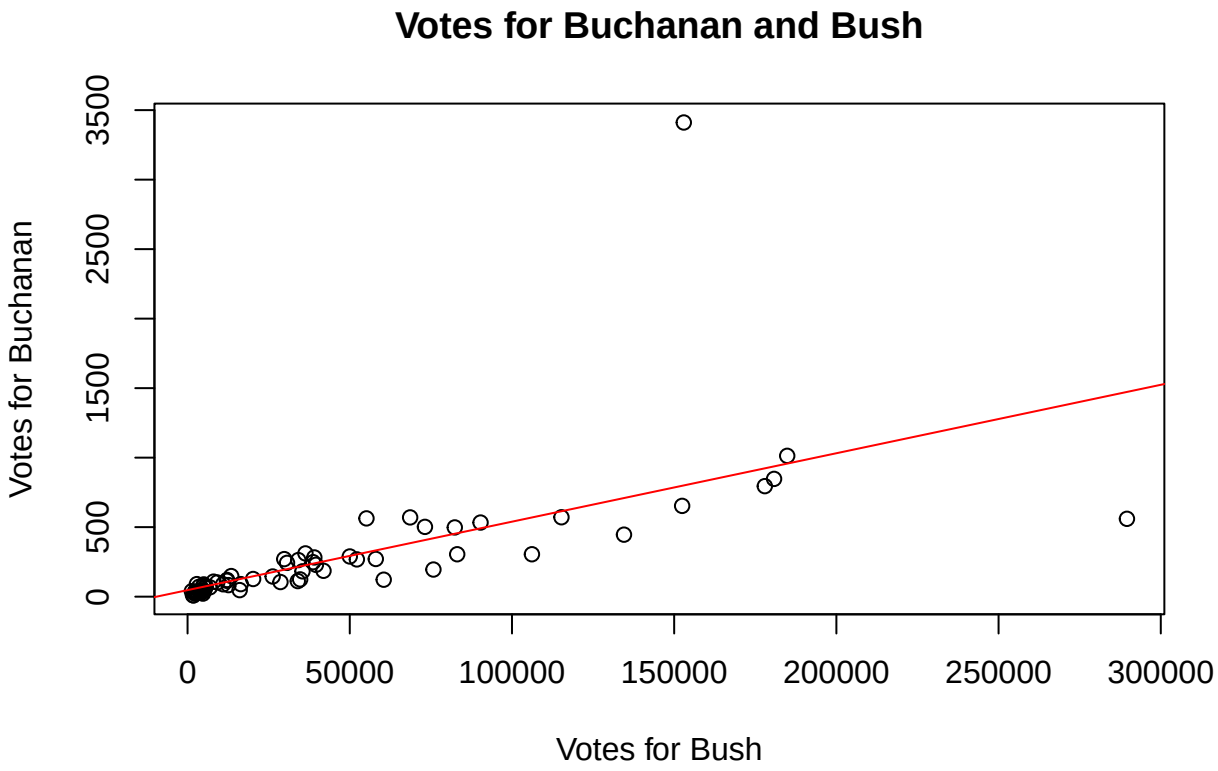
Start the analysis by creating a scatter plot of the number of votes that Pat Buchanan received as a function of the votes that George Bush received. Please use *base R graphics* for this plot and all subsequent plots on this homework.

Once you have created this plot, fit a linear model that can predict the number of votes Buchanan should receive given the number of votes that Bush received, and add the regression line to this plot in red. In the answer section below describe notable features of this graph (i.e., describe trends and unusual points in the plot).

```
plot(florida_data$Bush, florida_data$Buchanan,
     xlab = "Votes for Bush", ylab = "Votes for Buchanan",
     main = "Votes for Buchanan and Bush")

lm_fit <- lm(Buchanan ~ Bush, data = florida_data)

abline(lm_fit, col = "red")
```



Answers:

In general there appears to be a linear relationship between the number of votes received for Bush and Buchanan, but there is an outlier in the data.

Part 1.2 (5 points):

Now extract the coefficients from the linear model and print them. In the answer section below, report how many votes Buchanan is expected to get for every 1,000 votes Bush received, and how many votes the

model predicts that Buchanan would have gotten if Bush had received 0 votes. Also, write an equation that predicts the number of votes Buchanan should get as a function of the number of votes Bush received. Make sure to use use $LaTeX$ for the equation and that you use the proper notation discussed in class.

```
coef(lm_fit)
```

```
##  (Intercept)         Bush
## 46.883236695  0.004924084
```

Answers:

For every thousand votes Bush received, Buchanan is predicted to get 4.924084 votes. The model predicts that Buchanan would have received 46.883 votes if Bush had received 0 votes. $\hat{y} = 0.004925084x + 46.883236695$

Part 1.3 (12 points):

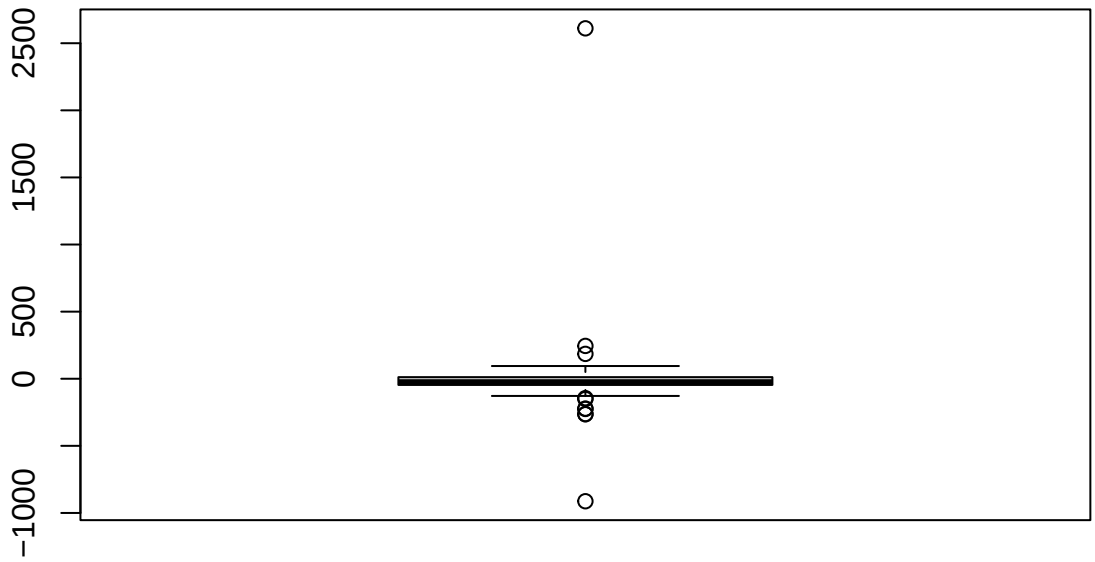
From looking at the plot above, it should be clear that there is one extreme outlier (i.e., one outlier that is much larger than all the others). To see this more clearly, create a box plot of the residuals of the model below. Then report answers to the following questions, **using the appropriate notation discussed in class for each number reported**:

1. What is the county that the outlier corresponds to?
2. How many votes did Buchanan actually received in that county?
3. What is the predicted number of votes that Buchanan should have received for this county based on the regression model fit above for that county?
4. What is the value of the residual for this county?

Please be sure to "show your work" by showing the code you used to come to answers to these questions in the code chunk below. Hint: the `which.max()` function could be useful, and if you save your linear model to an object called `lm_fit`, then extracting values from this object, such as `lm_fit$residuals`, will also be useful.

Finally, use the Internet to come up with a reasonable explanation that could have led to this outlier (embedding images in the markdown document could be useful here as well).

```
boxplot(lm_fit$residuals)
```




```
#florida_data

# outlier county
florida_data$County[which.max(lm_fit$residuals)]
```

```
## [1] "Palm Beach"
```

```
# votes in outlier county
florida_data$Buchanan[which.max(lm_fit$residuals)]
```

```
## [1] 3411
```

```
# residual value for outlier county
max(lm_fit$residuals)
```

```
## [1] 2610.909
```

Answers

1. The outlier corresponds to Palm Beach.
2. Buchanan received 3411 votes.
3. Based on the regression model fit above, Buchanan should have received just under 1000 votes.

4. The residual value is 2610.909

Part 1.4 (8 points):

Suppose that in the outlier county, Buchanan received exactly the number of votes predicted by the regression model, and the residual number of votes Buchanan received in that county were intended to be votes for Al Gore. To examine the consequences of this, start by calculating the total number of votes Bush received and the total number of votes Gore received. Then add the residual number of Buchanan votes from the outlier county to the total number of votes that Gore received. Fill in these values in the R Markdown table below to report these numbers. If the residual votes Buchanan received had indeed been intended for Gore, would this have changed who got the majority number of votes in Florida (and hence who would have won Florida)?

```
(bush_tot <- sum(florida_data$Bush))
```

```
## [1] 2912790
```

```
(gore_tot <- sum(florida_data$Gore))
```

```
## [1] 2912253
```

```
max(lm_fit$residuals) + gore_tot
```

```
## [1] 2914864
```

Answers

Bush votes	Gore + res	Gore votes
2912790	2914864	2912253

Yes, if the residual votes Buchanan received had been intended for Gore, it would have changed who got the majority number of votes in Florida.

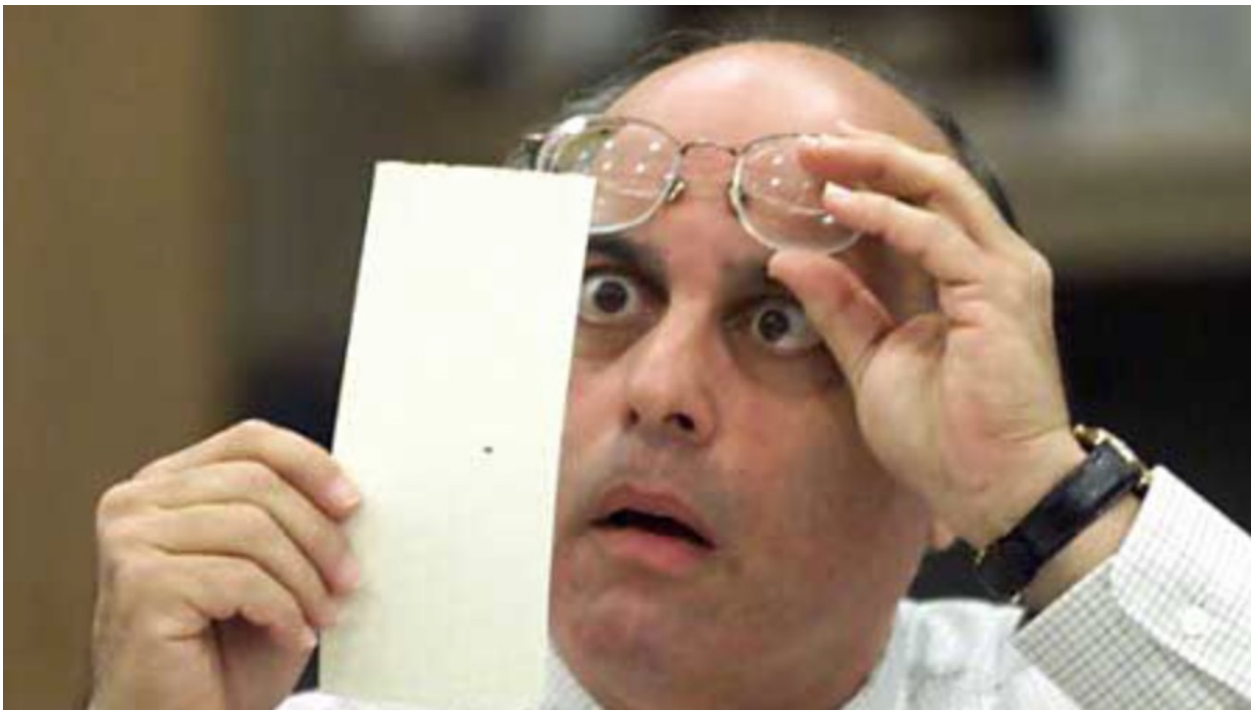
Part 1.5 (4 points):

The United States uses the Electoral College system. In this system, the candidate who gets the majority of the vote in a state wins all the Electoral College votes for that state (at least for most of the states in the US, including Florida). Use the Internet to find the number of Electoral College votes Bush received in 2000. Based on the number of Electoral College votes that Florida had, would there have been a different outcome of the election if Gore had won Florida?

Bonus (0 points), report any possible policy differences that might have been enacted had Gore been elected.

Answers

Bush received 271 electoral college votes in 2000. Florida had 25 electoral college votes. Based on this, the election would have had a different outcome if Gore had won Florida.



Part 2: Statistical inference on regression coefficients

On July 3rd 2015, my 1999 Toyota Corolla broke down on the side of the highway outside of Sturbridge, MA. While I had the car repaired, I knew it was time to sell it and get a new car. I intended to sell my Corolla to the car dealership, the only catch was that I was not sure how much the used Corolla was worth. In the following exercises we will model how much a used Corolla is worth as a function of the number of miles it has been driven.

The data we will look at comes from Edmunds.com which is a website where you can buy new and used cars online. This data set is from the 2015 DataFest competition, which is an undergraduate data science competition that takes place at difference colleges across the United States. The data has been made available to this class for educational purposes, however **please do not share this data outside of the class**.

Part 2.1 (8 points): Let's start by loading the `car_transactions` data set using the code below. Report how many cases and variables the full data set has in the answer section below. Then use the dplyr `select()` and `filter()` functions to create a reduced data frame object called `used_corollas` in which:

1. The only variables that should be in the `used_corollas` data frame are:

a. `model_bought`: the model of the car
b. `new_or_used_bought`: whether a car was new or used when it was purchased

c. `price_bought`: the price the car was purchased for
d. `mileage_bought`: the number of miles the car had when it was purchased

2. The only cases that should be in the `used_corollas` data frame are:

a. Used cars
b. Toyota Corollas
c. Cars that have been driven less than 150,000 miles

3. Finally use the `na.omit()` function on the `used_corollas` data frame to remove cases that have missing values.

Print out the number of rows and columns in this data frame to show you have the correct answer. If you have properly filtered the data, the resulting data set should have 248 cases, so check this is the case before going on to the next set of exercises.

```
# load the data set
load("car_transactions.rda")

used_corollas <- car_transactions %>%
  select(model_bought, new_or_used_bought, price_bought,
         mileage_bought) %>%
  filter(model_bought == "Corolla", new_or_used_bought == "U",
         mileage_bought <= 150000) %>% na.omit

dim(used_corollas)
```

```
## [1] 248   4
```

Answers

Part 2.2 (10 points):

Now that we have the relevant data, let's examine the relationship between a car's price and the number of miles driven. Let's begin analyzing the data by taking the following steps:

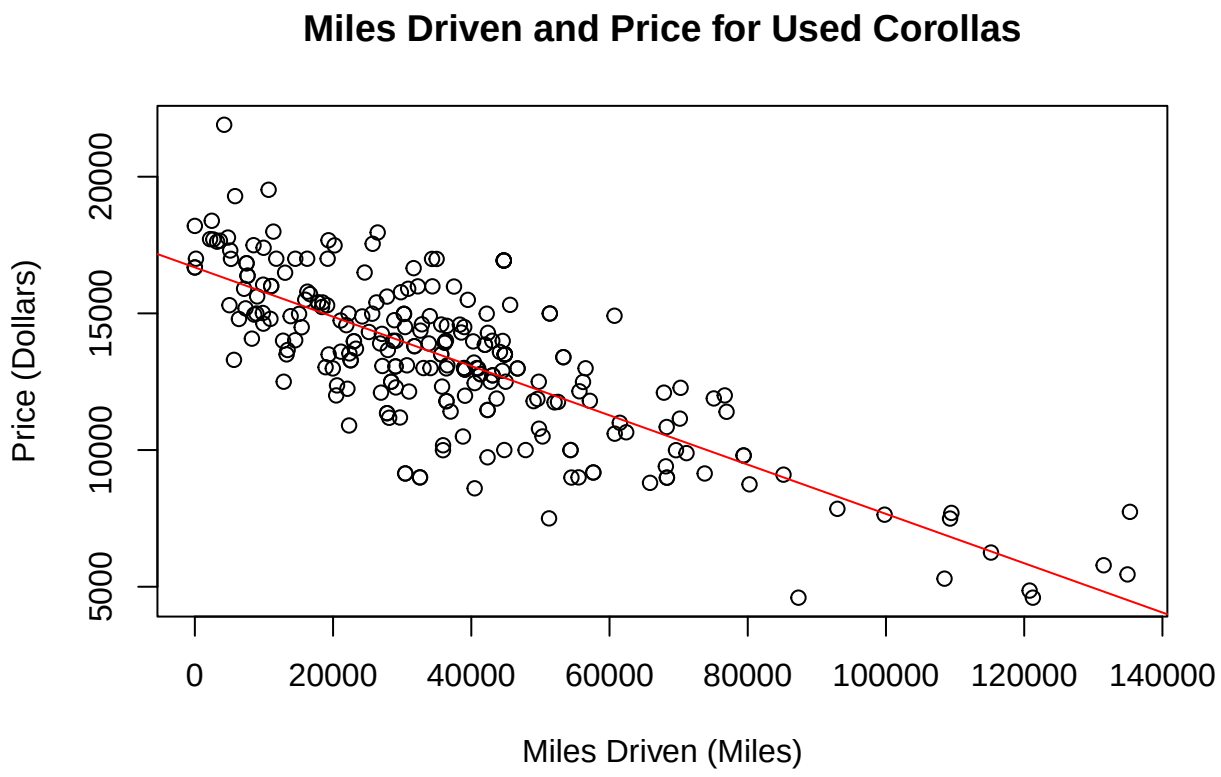
1. Plot the price as a function of the number of miles driven (again, use base R graphics for all plots on this homework).
2. Fit a linear regression model that shows the predicted price as a function of the number of miles driven. Save this model to an object called `lm_fit` which you will use throughout the rest of this homework.
3. Add a red line to our plot showing the regression line fit.
4. Print the regression coefficients found.

Report how much the price of a Corolla decreases for every additional mile it has been driven, and what this regression model suggests a car that has been driven 0 miles would be worth. Also describe whether the the sign and magnitude of these regression coefficient values match what you might expect for car prices. Finally, write out the regression equation. Again, make sure to use use $LaTeX$ for the equation and that you use the proper notation discussed in class.

```r
plot(used_corollas$mileage_bought, used_corollas$price_bought,
     xlab = "Miles Driven (Miles)", ylab = "Price (Dollars)",
     main = "Miles Driven and Price for Used Corollas")

lm_fit <- lm(price_bought ~ mileage_bought, data = used_corollas)

abline(lm_fit, col = "red")
```




```r
coef(lm_fit)
```

```
##    (Intercept) mileage_bought
## 16681.91992781    -0.09018627
```

Answers:

The price of a Corolla decreases by \$0.09 for every additional mile it has been driven, and the regression model suggests that a car that has been driven 0 miles would be worth \$16681.92. The sign and magnitude of these regression coefficient values match what I would expect for car prices, since cars depreciate over time.

Regression equation: $\hat{y} = -0.09018627x + 16681.91992781$

Part 2.3 (6 points): Now use R's `summary()` function to report whether there is statistically significant evidence that the price of a car decreases as a function of the number of miles driven. Also, write out the hypothesis that is being tested using the appropriate symbols/notation discussed in class.

```r
summary(lm_fit)
```

```
##
## Call:
## lm(formula = price_bought ~ mileage_bought, data = used_corollas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4791.8 -1131.9    -0.3  1027.7  5600.7
##
## Coefficients:
##                    Estimate   Std. Error t value            Pr(>|t|)
## (Intercept)    16681.919928   204.459353   81.59 <0.0000000000000002 ***
## mileage_bought    -0.090186     0.004539  -19.87 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1816 on 246 degrees of freedom
## Multiple R-squared:  0.616,  Adjusted R-squared:  0.6145
## F-statistic: 394.7 on 1 and 246 DF,  p-value: < 0.00000000000000022
```

Answer

$H_0 : \beta_0 = 0$

$H_A : \beta_0 \neq 0$

Part 2.4 (6 points): We can create confidence intervals using a t-distribution via the `confint()` function. Report what the confidence interval for the slope of the regression line is. Also, based on the confidence interval, explain why it seems likely that the price of a car is not independent of the number of miles driven.

```r
confint(lm_fit, level = .95)
```

```
##                         2.5 %         97.5 %
## (Intercept)    16279.20570876 17084.63414685
## mileage_bought    -0.09912747    -0.08124508
```

Answer

Part 2.5 (6 points):

This is a "challenge problem" that you should try to figure out without getting help from the TAs.

We can also use the bootstrap to create confidence intervals for the slope of the regression coefficient. To do this you can use the following procedure:

1. Create a bootstrap resampled data frame by sampling with replacement from the `used_corollas` data frame. You can do this using dplyr's `sample_n()` function with the sample size being the number of cases in the `used_corollas` data frame and setting the `replace = TRUE` argument.

2. Fit the regression model using the bootstrap data frame.
3. Extract the regression slope coefficient and save it to a vector object.
4. Repeat this process 1,000 times (this is less than what we normally use because it is computationally expensive to run this bootstrap procedure).
5. Plot the bootstrap distribution and use the "percentile method" to report a 95% confidence interval for the regression slope; i.e., use the `quantile()` function to get the confidence interval for the regression slope.

Report whether the bootstrap confidence interval is similar to the confidence interval using the t-distribution you calculated above.

```
n_cases <- dim(used_corollas)[1]

for (i in 1:1000) {
  corolla_sample <- sample_n(used_corollas, size = n_cases, replace = TRUE)
  boot_fit <- lm(price_bought ~ mileage_bought, data = corolla_sample)
  boot_coef <- coef(boot_fit)
}
```

Answers:

Part 2.6 (8 points): My Toyota had 180,000 miles at the time I wanted to sell it. Based on the regression model fit above, what is the predicted worth of this car? Also, in the answer section below, discuss if this seem like a reasonable estimate.

Answer

Part 3: Regression diagnostics

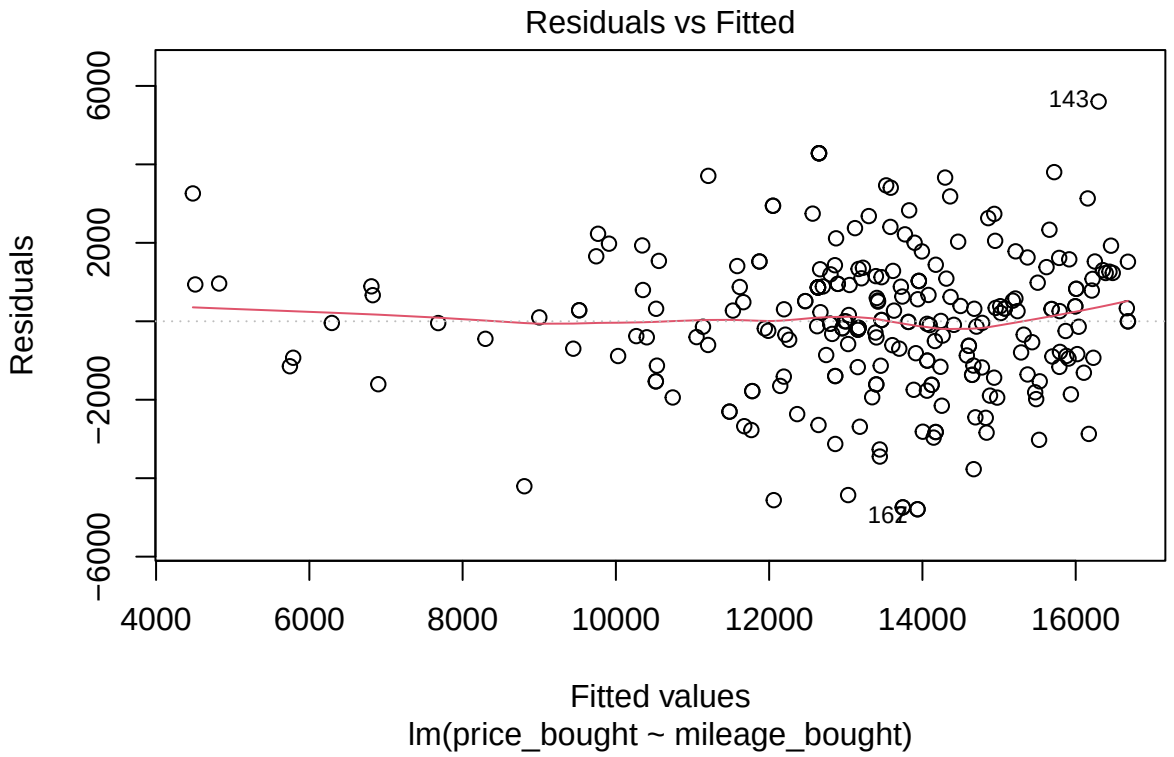
When making inferences about regression coefficients using parametric (t-statistic based) methods, there are a number of assumptions that need to be met to make the mathematical derivations of tests/confidence intervals methods valid. The assumptions are:

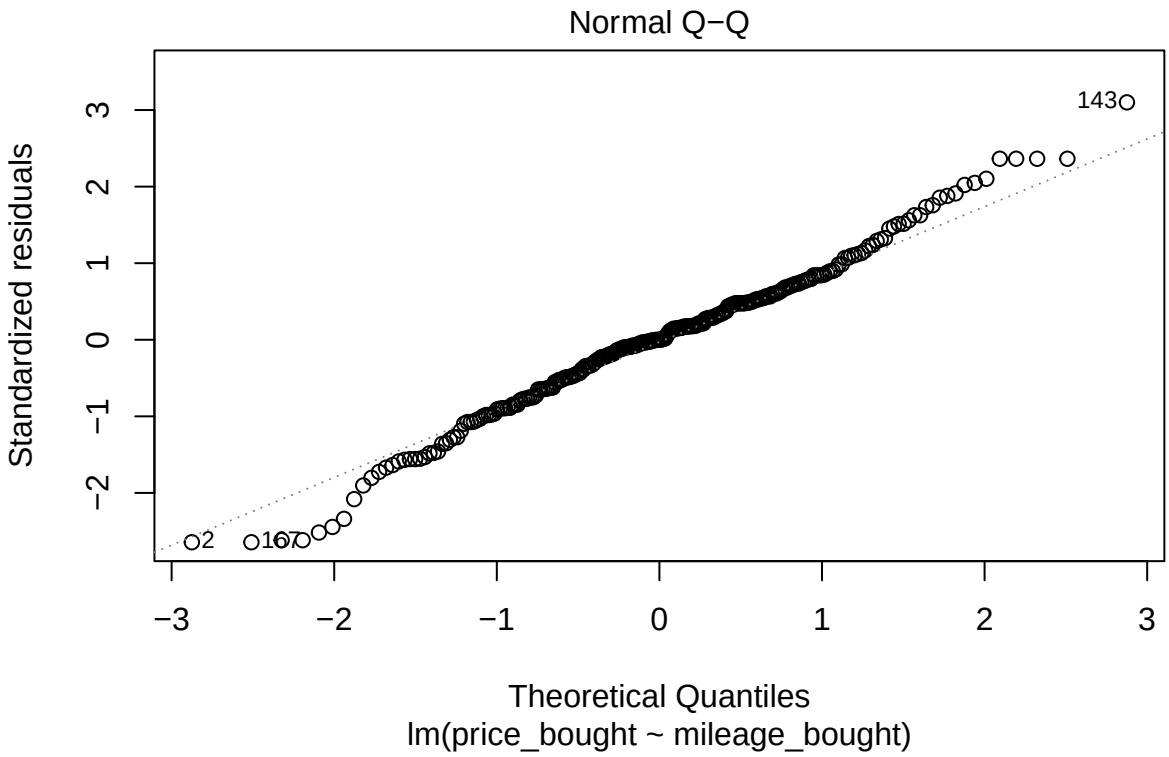
1. **Linearity**: A line can describe the relationship between x and y.
2. **Independence**: Each data point is independent from the other points.
3. **Normality**: The errors are normally distributed.
4. **Equal variance (homoscedasticity)**: The variance of the errors is constant over the whole range of x values.

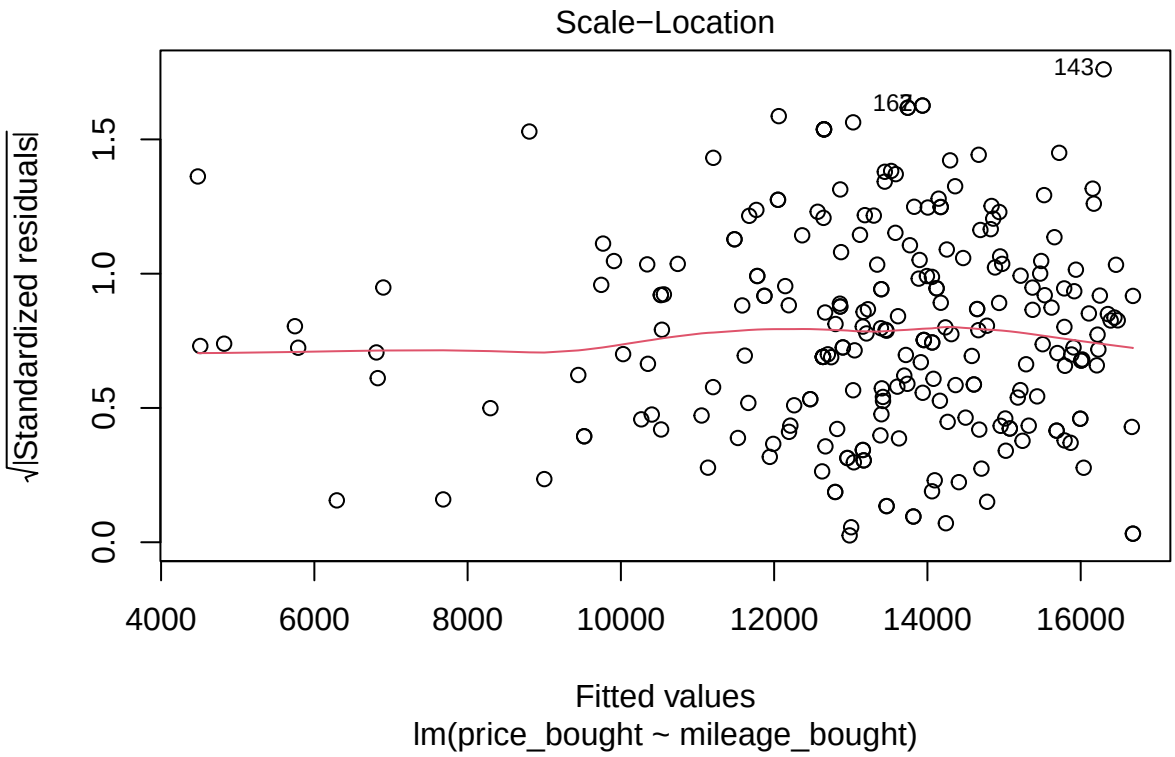
We can check whether these assumptions appear to be met by creating a set of diagnostic plots.

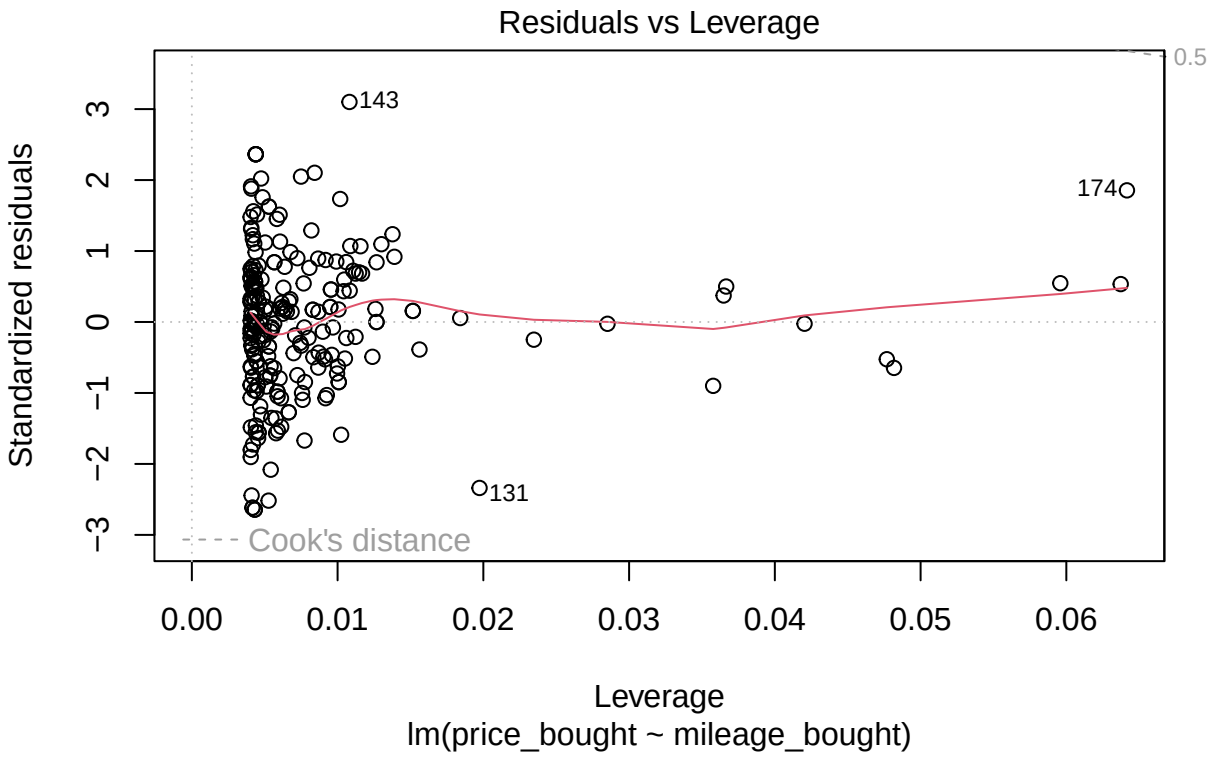
Part 3.1 (5 points): To check for linearity and equal variance of errors (conditions 1 and 4 above), we can create a plot of the residuals as a function of the fitted values. Create such a plot below using information in the `lm_fit` object. Does it appear that the linearity and homoscedasticity conditions are met? Are these results what you would expect from looking at plots above and from the nature of the type of data you are analyzing?

```
plot(lm_fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(price_bought ~ mileage_bought)
143
162
6000
2000
−2000
−6000
4000
6000
8000
10000
12000
14000
16000


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(price_bought ~ mileage_bought)
143
2
167
−3
−2
−1
0
1
2
3


Scale−Location
143
162
√|Standardized residuals|
0.0
0.5
1.0
1.5
4000
6000
8000
10000
12000
14000
16000
Fitted values
lm(price_bought ~ mileage_bought)

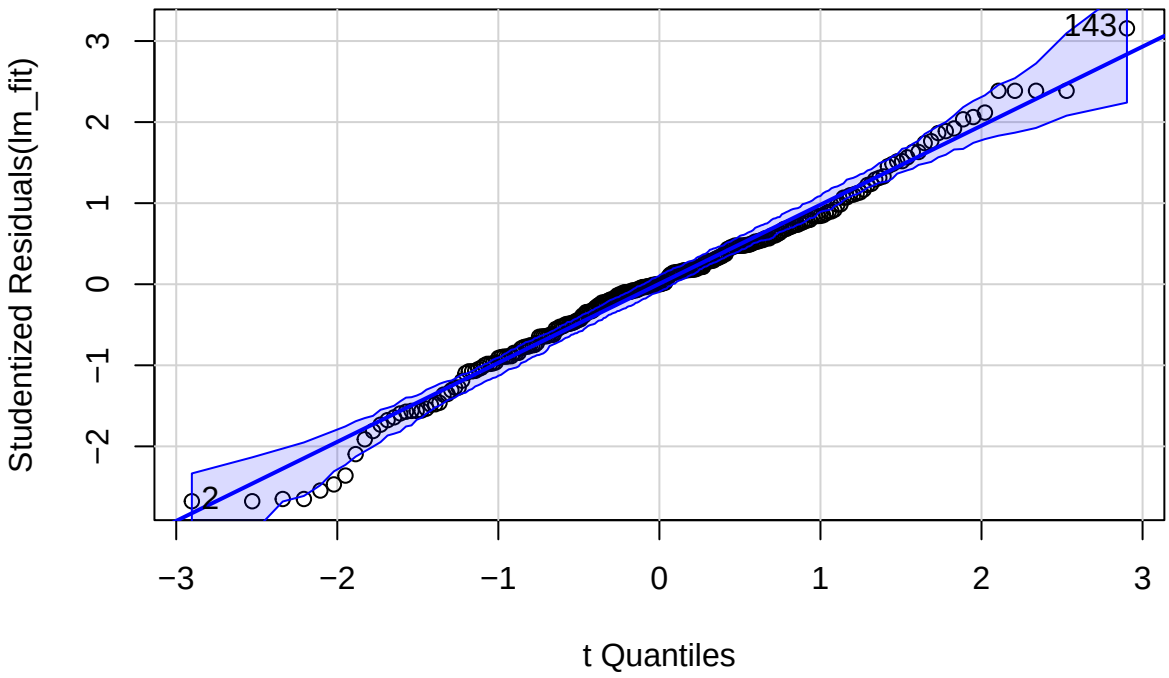


Answers:

It does appear that the linearity and homoscedasticity conditions are met. These are the results that I would expect from looking at plots above and from the nature of the type of data I am analyzing

Part 3.2 (4 points): To check whether the residuals are normally distributed (condition 3 above) we can create a Q-Q plot. The `car` package has a nice function called `qqPlot()` to create these plots. If we pass the `lm_fit` object to the `qqPlot()` function it will create a Q-Q plot of the studentized residuals. Create this plot and report if the residuals seem normally distributed.

```
qqPlot(lm_fit)
```




```
##    2 143
##    2 140
```

Answer:

The residuals do seem normally distributed.

Part 3.3 (5 points): To check if the data points are independent (condition 2 above) requires knowledge of how the data was collected. For example, if the data you have is from a time-series (e.g., recordings of the temperature in New Haven on consecutive days) then there is a high likelihood that the data points might not be independent. On the other hand, if you take a simple random sample from a population where every point is equally likely to be selected, then the data is going to be independent.

Unfortunately I do not know exactly how this data was collected so it is difficult to say if the data is independent here. However, there might be ways to investigate whether it seems plausible that it could be independent. Please describe some ways you might investigate whether the data could be independent (hint: think about the variables in the full `car_transactions` data set). Note: there is no exact 'right answer' here, just describe some possible ideas.

Answer:

I could investigate the dates on which the cars were sold. If the data shows a roughly equal amount of cars sold month-to-month, this could suggest that the data was collected by randomly selecting a fixed number

of samples from each month of a larger data set of car sales, as opposed to documenting how many cars were sold over time.

Reflection (3 points)

Please reflect on how the homework went by going to Canvas, going to the Quizzes link, and clicking on Reflection on homework 6.