INDE546_HW4

Michael Shieh 1826962

1/29/2020

Exercise 1

I want to use the student's average commute time to UW (from question 2) as the dependent variable.

For the variables I intend to include in the model:

1. **Do the participant use Uber/Lyft at least once every week on average (Categorical variable: Yes, No):** Ride sharing usually takes less time than public transportations, which means for students who live in a location with longer commute time they might take Lyft/Uber to save time more regularly.
2. **The age of the participants (Continuous variable: Min: 18, Mean: 24.06, Max: 38):** I think age of the participants would likely affect the housing choice, where different age groups may have different preferences on residental area. This would result in a difference in average commute time.
3. **How often do the participants go on UW campus(Categorical variable: High, Low):** If a student need to go on campus frequently, he may not choose to live in a location where it has long commute time. High Frequency means more than 5 days a week.
4. **Whether the primary means of transportation involves traffics (Categorical variable: Yes, No):** Since transportations that involves traffics (e.g. bus, drive) would more likely run into delays than other means of transportations (e.g. walking, biking, light rail), average commute time could be higher comparing to other means.
5. **Whether the participants share houses with significant other or other family members (Categorical variable: Yes, No):** In this case, those who live with significant other or other family member may choose to live some place that is suitable for both to commute to their jobs or schools. This could result in longer commute time because of the trade-offs between commuting to tow different locations.

Exercise 2

First, I need to prepare the data for the model.

```
survey <- read.csv("Class_Survey_W20.csv", header = TRUE)
#Rename column names of the variables that I will use in this regression model.
survey <- survey %>%
  rename(Age = How.old.are.you., Who = Are.you., Tmode = What.is.your.primary.means.of.transportation.to.and.from.the.U..Washington., CommuteTime = On.average..how.many.minutes.does.it.take.you.to.get.to.the.U..Washington.from.your.home.) %>%
  rename(Transportation = Please.indicate.how.much.you.agree.or.disagree.with.the.following.statements...Seattle.public.transportation.is.reliable.,  Days = In.an.average.week..how.many.days.are.you.on.the.UW.campus., Gender = Are.you..1) %>%
  rename(LyftUber = On.average..how.many.days.in.a.week..out.of.7.days..do.you.use.Uber.Lyft.or.other.ride.hailing.apps., ShareHouse = Do.you.live.with.a.significant.other.or.other.family.members.)
```

```
#Filter the dataset to clear out missing and meaningless data.
#ShareHouse = Yes means living with significant other or other family members. Otherwise, no.
survey <- survey %>% filter(Age >= 0, CommuteTime >= 0, Who == 'Student', ShareHouse %in% c('Yes', 'No')) %>% drop_na(Who, Tmode, Transportation, Days, LyftUber)

#Transform the variables into the way I wnat to put in my regression model.
#Traffics = Yes means primary means of commute is either buses or driving. Otherwise, no.
#Frequency = High means go on UW campus at least 5 days a week. Otherwise, low.
#LyftUber = Yes means taking Lyft/Uber at least once a week on average. Otherwise, no.
survey <- survey %>%
  mutate(Traffics = ifelse((grepl('Bus', survey$Tmode) | grepl('Drive', survey$Tmode)) , "Yes", "No"), AgeSq = Age^2, Frequency = ifelse(grepl('Every', survey$Days), "High", "Low"), LyftUber = ifelse(survey$LyftUber > 0, "Yes", "No"), CommuteTimeSqrt = CommuteTime^0.5, AgeSq = Age^2)
```

I also want to see if the regression model can make good predictions on commute time.

```
#Seperate the data into two half, one as training set, another as testing set.
train.ix <- sample(nrow(survey), floor(nrow(survey)/2))
survey.train <- survey[train.ix,]
survey.test <- survey[-train.ix,]
```

I use the square root of commute time because as stated in class, the value seems to fit into a normal distribution after taking the square root.

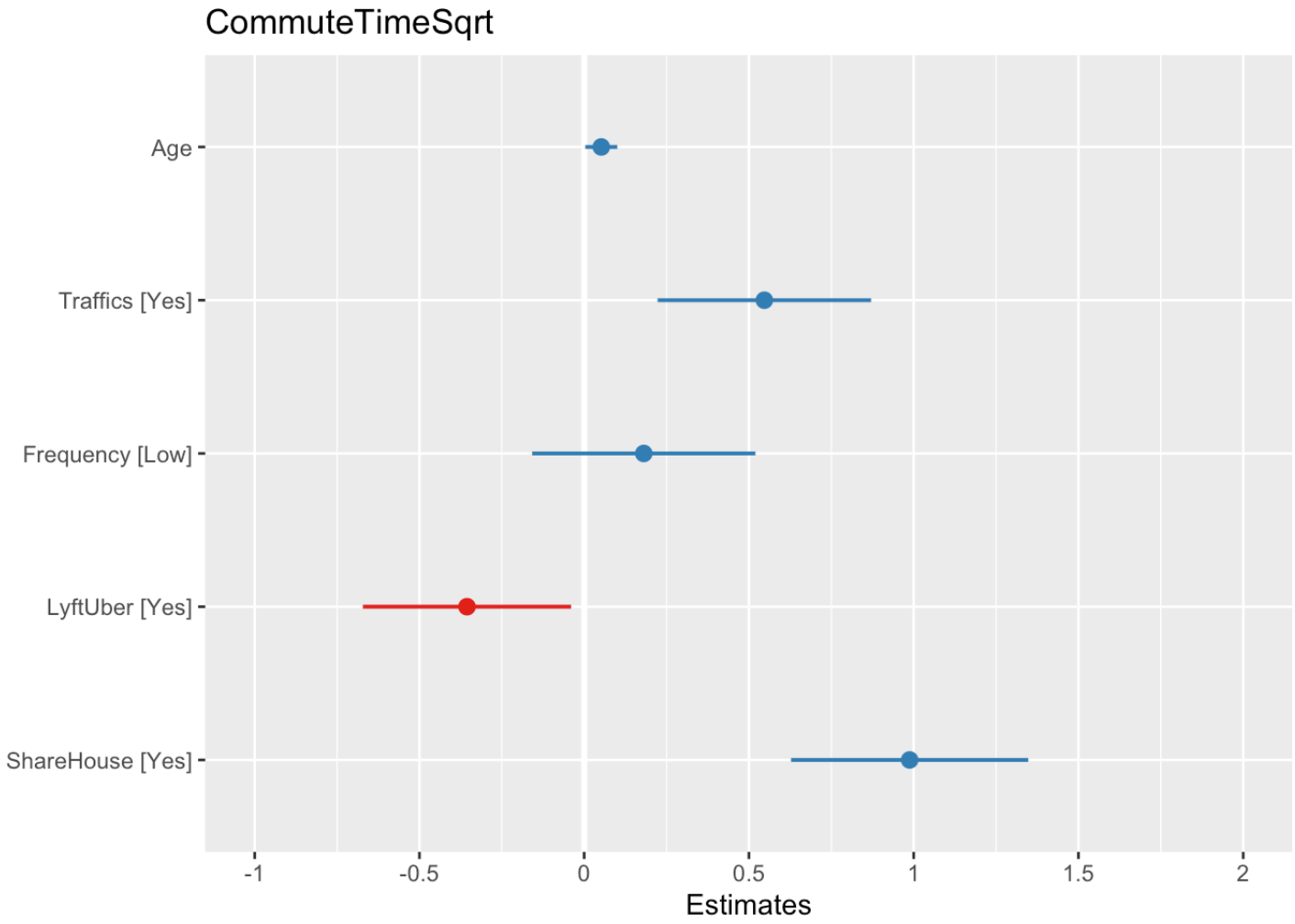
```
#Put variables in Regression Model, use training set to train the model.
LB_HW4 <- lm(CommuteTimeSqrt ~ Age + Traffics + Frequency + LyftUber + ShareHouse, data = survey.train)
summary(LB_HW4)
```

```
##
## Call:
## lm(formula = CommuteTimeSqrt ~ Age + Traffics + Frequency + LyftUber +
##     ShareHouse, data = survey.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3116 -0.6790 -0.0614  0.6683  5.3289
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.68127    0.57972   4.625 6.39e-06 ***
## Age            0.05189    0.02469   2.102  0.03671 *
## TrafficsYes    0.54676    0.16439   3.326  0.00103 **
## FrequencyLow   0.18083    0.17193   1.052  0.29404
## LyftUberYes   -0.35568    0.16029  -2.219  0.02750 *
## ShareHouseYes  0.98772    0.18271   5.406 1.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.178 on 220 degrees of freedom
## Multiple R-squared:  0.2546, Adjusted R-squared:  0.2377
## F-statistic: 15.03 on 5 and 220 DF,  p-value: 1.084e-12
```

Exercise 3

Next, I look at confidence interval for all included variables:

```
plot_model(LB_HW4)
```




```
confint(LB_HW4)
```

```
##                     2.5 %      97.5 %
## (Intercept)    1.538750534  3.82379899
## Age            0.003233026  0.10055559
## TrafficsYes    0.222769330  0.87074329
## FrequencyLow  -0.158002620  0.51967110
## LyftUberYes   -0.671573566 -0.03978942
## ShareHouseYes  0.627629029  1.34781560
```

The confidence interval for Age, TrafficYes, LyftUberYes, and ShareHouseYes do not include zero, so they appear to be sighificant factors for commute time.

The following is the multicollinearity test:

```
vif(LB_HW4)
```

```
##        Age   Traffics  Frequency   LyftUber ShareHouse
##   1.139997   1.099052   1.074576   1.042114   1.048937
```

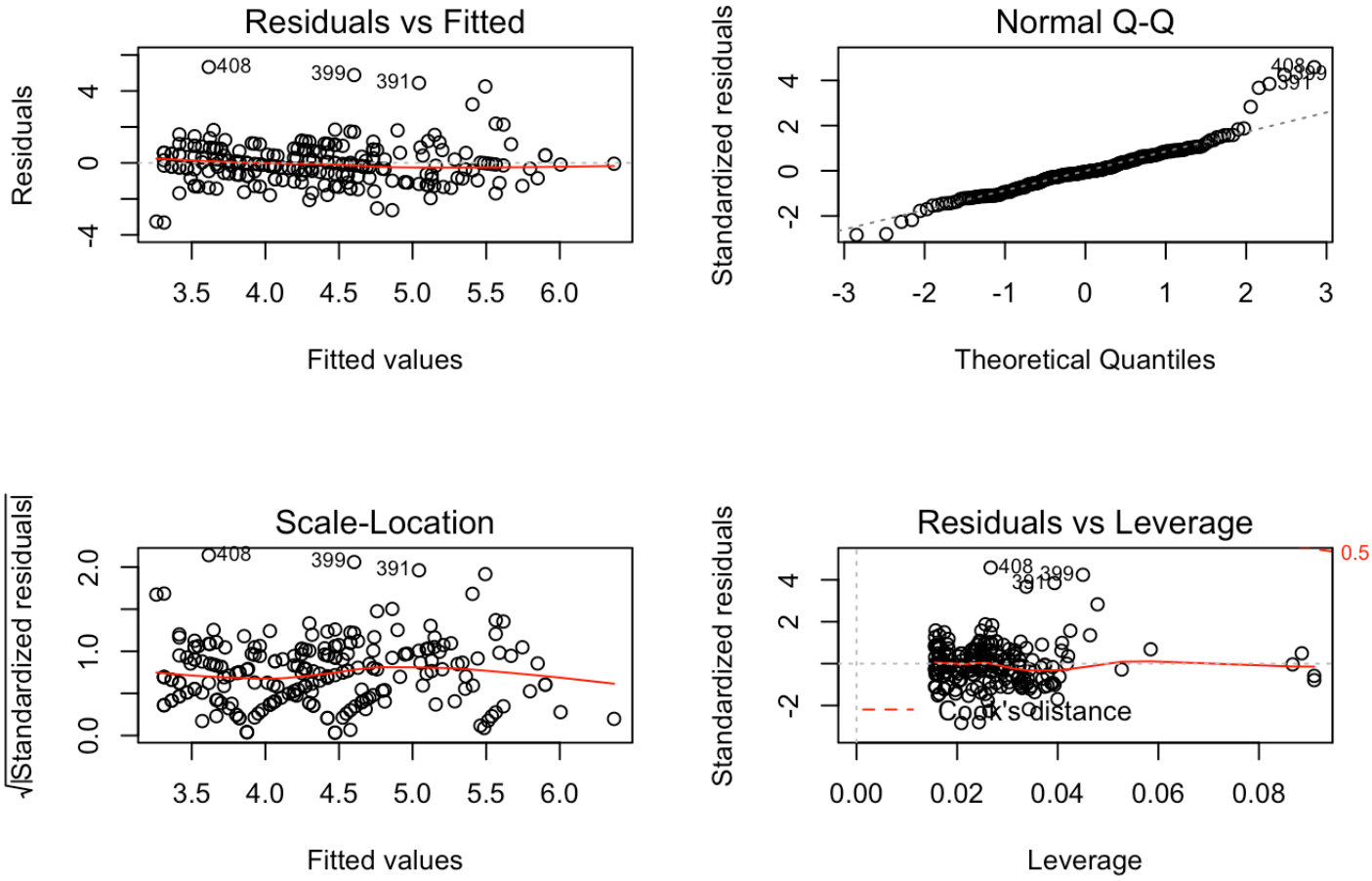
VIF are all close to 1, so no multicollinearity problem appears to be present.

Then, I plot the residuals to see if the normality assumption holds:

```
mean(LB_HW4$residuals)
```

```
## [1] -3.413126e-17
```

```
par(mfrow = c(2, 2))
plot(LB_HW4)
```



The mean of the residuals is approximately zero.

It seems most of the data fits on the linear line from Normal Q-Q plot, and there doen't seem to have a pattern from Residuals vs Fitted plot. However, I did observe there might be some outliers existing in the data, and based on the R-square value (around 0.25) I think that this

model is still being affected by a lot of noises.

Overall, I'd say the model adheres to the OLS assumptions pretty well. Lastly, I will use the testing data to see if the regression model is making a good prediction.

```
predict.lm <- predict(LB_HW4, survey.test)
cor(predict.lm, survey.test$CommuteTime)
```

```
## [1] 0.3755044
```

The correlation shows that the model is not making a good prediction, which should aligns with the siginificant amount of noises existing.

Exercise 4

From the result of the model, I would say that students whose means of transportation experience traffics and/or living with significant other or other family members are likely to have longer commute time. Also, it seems that taking Lyft/Uber regularly can help reduce commute time. Note that, 64% of the responses (137 out of 214) stated that they take Lyft/Uber because of conveniences, I thought this could mean saving time on longer commutes. As it turns out, the average commute time for those is actually 18.138 minutes, where the average commute time for all students is 20.631 minutes. It means they are more likely those who have shorter commutes, and by taking Lyft/Uber they would have even faster commute.

Exercise 5

Since I choose variables based on my own reasoning, there can be other significant variables that I omitted in my model, or the sampling can be biased that I could over-estimate the effect of certain variables, for example, younger generation, which is a majority of the student population, might be more willing to take Uber/Lyft. Therefore, I might be over-estimating this effect because of my samples. There can also be some simultaneity happening between two variables that we can't easily tell.