Exercise 1. Select one of the continuous variables from the survey data. State some hypotheses about what other variables in the data set might be useful for predicting the dependent variable you have chosen. Briefly explain your reasoning (1-2 sentences per predictor). Identify at least 3 or 4 predictors, including at least one numerical and at least one categorical predictor.

Dependant Variable:

timeUW <- Q12 On average how many minutes does it take you to get to the U Washington from your home

Independent Variables:

public_transport_frequency_five <- Q10 Since Sept 2021 when school began, how often have you used the following mode to and from the UW Transit bus light rail (5 or more days a week) - these group of users would be the most likely to take public transport to UW and as a result would lead to time spent to get from UW to home

milesUW <- Q11 How many miles do you live from the UW - the distance from one's home to UW determines how much time they need to travel in order to get to school

vehicle <- Q14 Do you have a vehicle you can use on a regular basis - If one has a vehicle they use on a regular basis, that would significantly decrease the amount of time one needs to get to UW

licence <- Q34 Do you have a US driver's licence - If one has a driver's licence, then they would be able to drive to UW from their homes

E2. Estimate a linear regression model in R that includes the variables you have identified above. Create a journal ready table (as described in class) that shows your parameter estimates, standard error, test statistics, p-value, and confidence interval. Explain your rationale for the particular model specification that you have chosen (i.e. what variables, if any, are squared or logged or otherwise transformed? What variables, if any, are interacted with one another?)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.0033	1.9574	8.69	0.0000
public_transport_frequency_fiveyes	4.6234	1.8662	2.48	0.0140
milesUWlog	10.0869	0.5738	17.58	0.0000
vehicleYes	-5.2536	1.6776	-3.13	0.0020
licenseYes	2.1414	2.2511	0.95	0.3425

Figure 1: Summary table of linear regression model

milesUW was transformed through a natural log function to milesUWlog. This was done as in the univariate analysis of Miles to UW, the histogram plot follows a logarithmic distribution rather than a linear one. The other variables chosen were bivariate categorical data (yes, no) and did not require any transformation

There exists a positive relation between time to UW with all of the variables except for if the users own a vehicle which exhibits a negative correlation. This makes sense because for users who drive a vehicle on a regular basis would require significantly less time to reach UW than one that has to take public transport.

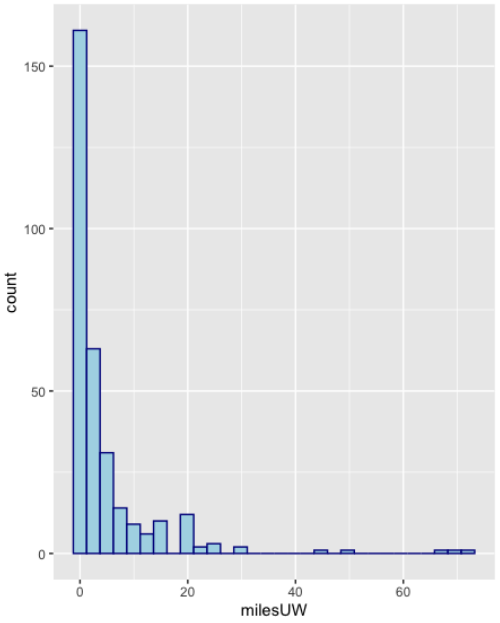
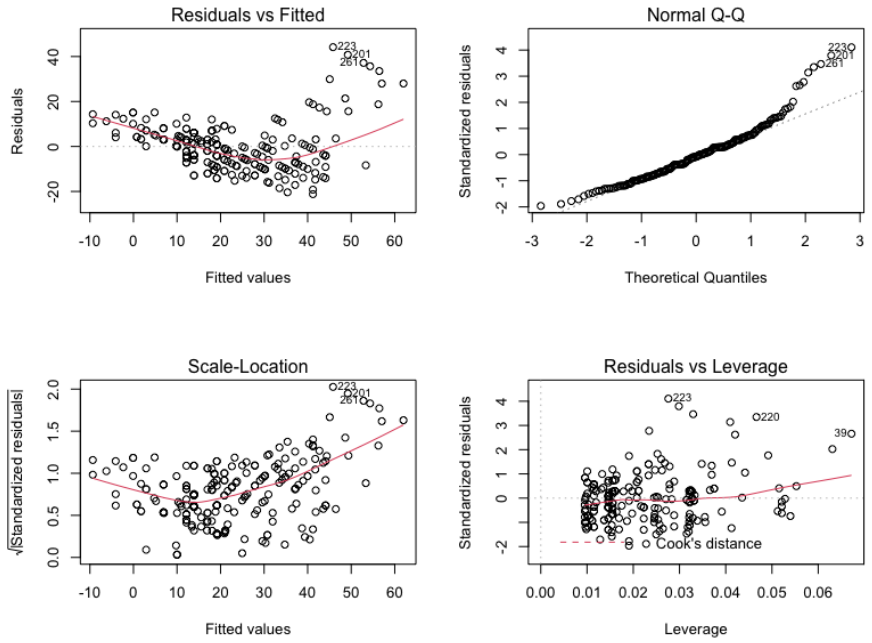


Figure 2: Histogram of miles to UW

<u>E3</u>. How well do you think your model adheres to the OLS assumptions? Support your conclusion by presenting and integrating information from multiple sources, including model estimation results, diagnostic plots, and statistical tests, as appropriate.



In the Normal Q-Q graph, most of the data are along the straight line with some on the right tail end deviating from it. This would suggest the residuals mostly follows the normality assumption

In the residuals vs fitted graph, there are no discernable patterns however the line exhibits a slight U-shaped curve. This would suggest that linear relationship assumption mostly holds.

In the scale-location graph, points are not evenly spread and the line has a slight U-shaped bent. This could suggest that the variables do not quite have the same variance and may not be totally homoscedastic.

In the residual vs leverage graph, all points are within the Cook's distance and there are no extreme values that would influence the regression results.

```
public_transport_frequency_five                 milesUWlog
                       1.069963                   1.227084
                        vehicle                    license
                       1.309425                   1.168153
```

VIF values of the 4 dependent variables are close to 1 which suggests that there is very little multicollinearity between the terms.

Since the 5 assumptions of OLS are not broken, this would suggest that the model adheres to the OLS assumption

E4. Interpret your model. What insights have you gained about each variable as it relates to the outcome? What caveats would you offer?

All the independent variables chosen have Pr(>|t|) < 0.05 except for the variable licenceYes. This would suggest that except for licenseYes, the other independent variables exhibit a statistically significant relationship with the dependent variable timeUW, while licenceYes does not. Therefore, owning a licence may not be an important variable to consider in modelling a regression model to predict time to UW.

E5. Can you think of any sources of endogeneity that you might not be able to directly test for using the data collected from the survey?

Some sources of endogeneity within the source includes time taken to get to UW (Q12) and miles away from UW (Q11), how often one speeds up at the yellow traffic light (Q15) and how many times one has received a moving violation ticket in the past 5 years (Q19)