Is Toronto Losing Money?

An investigation into plausible predictors for subway delays [Assignment 2]

Sahil Patel - 1006747905

Introduction

Subway delays cost cities money, and it adds up. In New York alone, subway delays have been estimated to cost people almost 389 million US dollars annually [5]. Simple 5 minute delays can cause upwards of 466 thousand US dollars everyday projected by the New York government. These estimates are derived from multiply the average New York wage by the aggregate amount of time lost by the commuters. For an individual person, 5 minutes could seem like nothing, however given that almost 5.5 million people [7] take the subway on an average weekday in New York, the time adds up. If even just one train has 5 minutes of delay, the impact could be tens of thousands of dollars.

That being said, New York is a completely different city than Toronto, however it is not unreasonable to suspect that subway delays in Toronto could have a huge economic impact. It is estimated that on any given weekday, almost 1.55 million people use the Toronto subway system to commute around Toronto [8]. As a result, any delay in the subway system could have a huge impact as commuters try and get to work. Given that subway delay in Toronto almost certainly has some negative impact on the economy, it would be important to try and determine if there are any ways to make the subway system more robust. For example, if more delays occur as a result of some event, it would be beneficial to the city, and the commuters, to revise the system to prevent future delays. These revisions, by lowering the amount of delay, would help reduce the economic impact of subway delays and save commuters precious time.

So how do we decide where the subway system might need revisions? Given the complexity of the subway system, its believed the delays are inevitable, however any underlying patters in the subway delay, might indicate that there are factors affecting the delay that could be controlled. For example, if there is significantly more delay time on a Monday, that might indicate that something is happening on Mondays, causing the increased delays. From that information, further investigation could be done to determine the cause of the increased amount of delay so that it can be resolved. Another possible source of increased error could be the time of day. For example, as the time approaches rush hour, we could expect to see longer delays as more people try and use the subway system. If this increase is significant, it might mean that the subway system should be altered to handle the increase of people at rush hour.

As a whole it is hypothesized that as the time approaches rush hour, on weekdays, the amount of subway delay will increase linearly. What this means is that as the current time approaches rush hour, we expect an increase in subway delay. This is hypothesized as a result of the increased population using the subway system during this time, thus correlating to more events that could cause subway delays. To investigate this hypothesis, subway delays during the month of September in 2021 are used to map any relationship between these variables and the amount of subway delay in the morning leading up to rush hour.

Data

Data Collection Process

The data for this analysis was collected by the Toronto Transportation Commission [9] and published on open.toronto.ca[10]. No methods for how the data was collected were published, and as a result we are prevented from identifying any biases that might have been present in the collection process. As for limitations, these are also hard to identify given the lack of information on the data collection process, however, as a whole, a limitation includes the overall validity of the result of this analysis. Given that no collection process was published, it is plausible that the methods used were invalid and could void any meaningful results of this analysis.

Data Summary

Given the data wasn't terribly messy, not much cleaning was required. A new variable describing the number of hours past midnight was introduced that would take the hours and the minutes divided by 60, and add them together. Given that the time in the data originally was in military time (24 hour clock), this new variable would represent the number of hours past midnight. Next, any observations where the recorded vehicle number was 0 were removed as our investigation only cares about currently actively used vehicles. Penultimately, any observations within the new hours past midnight variable after 12 (meaning noon) were removed. This is because our investigation only cares about noticing trends during the morning and as a result, all data after 12 can be removed. Lastly, the cleaned data set was created by only extracting our variables of interest, which include our new generated variable (the number of hours past midnight), the day (eg. Monday), and the subway delay time in minutes.

As mentioned briefly above, this analysis only required 3 variables. The response variable, the item trying to be predicted is, is the amount of subway delay experienced at any given subway station. Subway delay was recorded in minutes by the TTC supposedly every instance subway delay occurred and numerical summaries for the subway delay can be seen in the appendix. The next variable, one of the predictors, was the day of the recorded delay. This variable stores the day of the week a delay occurred. For example, for a given delay, the day of the week could be "Monday" or "Saturday". The last predictor and variable of importance was the number of hours past midnight. This continuous variable stores the time, quantified as the number of hours past midnight. An example of this variable could be 4.50; this represents that a given subway delay happened at 4:30 AM. Visualizations and numerical summaries for these variables can be seen in the appendix.

As shown above in **Table 1**, it is quite evident that there exists a somewhat large variation for both the number of hours past midnight (time), and the amount of subway delay in minutes. For the time a delay occurred, normal appearing data is present such that a majority of the data lies between around 6:30 and 10:00 in **Figure 1**. This could suggest that as the time approaches rush hour, more delays are expected to occur. Furthermore, this could suggest that the time could be related to the delay. Viewing **Figure 2** we see that the overall amount of delay, the value trying to be predicted, seems to have a number of outliers that represent long amounts of delay time. Lastly, inespecting **Table 2** lends us to notice two things. The first being that it appears that some days have many more instances of delays. For example, there were a recorded 57 delays on Thursday, but only 33 delays on Saturday. This combined with the fact that there also are noticeable differences in the mean amount of delay on given days lends us to believe that the day impacts how much delay that is expected. Thus, as a whole it might be plausible to explain the amount of delay experienced in the subway as a result of the time (in hours after midnight) and the day.

All analysis for this report was programmed using `R version 4.1.1`.

Methods

In order to attempt to predict the amount of delay in minutes given the time and day, a linear regression will be used. Then to verify whether or not the model as a whole has any predictive value, an ANOVA F test will be used. The linear model will take the form:

$$y = \beta_0 + \beta_1 x_1 + \Sigma_{i=2}^{8}\beta_i 1_{x_2=(i-1)^{th}\text{day of the week}} + \epsilon$$

In this model y (the delay in minutes) is explained by the predictors and coefficients on those predictors. β_0 denotes the intercept of the regression line. Specifically, this means that β_0 represents a baseline amount of delay (in minutes) that is expected given that it is midnight on any given day. Furthermore, let x_1 to represent the time (represented as the number of hours past midnight), then β_1 will represent the change in our delay as the day progresses by one hour. Penultimately, if we let x_2 denote the day, then $\Sigma_{i=2}^{8}\beta_i 1_{x_2=(i-1)^{th}\text{day of the week}}$ (where 1 represents an indicator function) represents the amount of delay change we expect on that given day. For example, suppose $x_2 =$ Monday. Given that Monday is the second day of the week after Sunday, then we would get that $\Sigma_{i=2}^{8}\beta_i 1_{x_2=(i-1)^{th}\text{day of the week}} = \beta_3$. Thus on Mondays, we would expect β_3 minutes more of delay. Lastly ϵ represents any natural error in the data.

In order to verify that the linear model supposed above might be plausible, it is necessary to check a number of assumptions. First is whether the relationship between the variables might be linear in nature. As seen in **Figure 3**, it seems like there is a slight increase in delay as the number of hours past midnight increases thus supporting the plausibility of a linear relationship. However, this might be attributed to the possible increase in the variance in the amount of delay as the number of hours past midnight increases. Thus, homoskedasticity, or the constant error variance, is assumed but it will be necessary to verify this assumption in the future. Then it will be assumed that there are uncorrelated error terms that are distributed normally, thus satisfying the assumptions to proceed with a linear regression.

In order to verify the calculated values of the $\hat{\beta}_i$s, an ANOVA F test will be conducted. As a whole, the results of this test state whether the calculated linear regression actually does a good job explaining the trend in the sampled data. As a whole, the null hypothesis of this test is that there doesn't exist a linear relationship between any of the predictors (the time and day) and the response (the subway delay time). Complementing this, the alternative hypothesis states that at least one of the predictors explains the variation in the response by being linearly related to it. The F-statistic is calculated using the formula:

$$F_{statistic} = \frac{SS_{reg}/p}{RSS/(n-p-1)} \sim F(p, n-p-1)$$

Where RSS is the residual sum of squares, SS_{reg} is equal to the total sum of squares minus RSS, n is the number of observations in the sample, and p are the number of predictors. Then from the F-statistic, a p-value is calculated. If the p-value is below the set $\alpha = 0.05$, then we reject the null hypothesis that our predictors are not linearly related to our response variable. If there is enough evidence to reject the null hypothesis, then model selection will be done by calculating simple t statistics for each predictor and using the p-values of these calculated t statistics to extract significant coeffecients (and thus predictors) for the amount of subway delay. However, if there is not enough evidence to reject the null hypothesis, then model selection cannot proceed as none of the predictors are linearly related to the amount of subway delay.

Results

Within the linear model, 2 predictors are used to predict the amount of subway delay (which is conveyed in minutes). The first predictor is the number of hours past midnight. Essentially, this is equivalent to the time but in terms of hours, thus 4:30 becomes 4.5 and so on. This is a continuous variable because time is continuous. The next predictor is the day of the week. Given that there are only 7 days in a week, this is a categorical variable that can be any day of the week.

Table 2. Calculated beta values for the hypothesized linear regression

Beta	Value
$\hat{\beta}_0$	6.342
$\hat{\beta}_1$	0.074
$\hat{\beta}_2$	-1.268
$\hat{\beta}_3$	-2.277
$\hat{\beta}_4$	-3.652
$\hat{\beta}_5$	-3.092
$\hat{\beta}_6$	-2.822
$\hat{\beta}_7$	0
$\hat{\beta}_8$	-4.609

Above in **Table 2** the values for the hypothesized betas have been calculated. Similar to the linear model earlier, $\hat{\beta_1}$ represents effect of the current time on the suspected amount of subway delay. For example, on the same day one can expect 0.074 minutes (approximately 4 seconds) longer of subway delay if they wait one hour before taking the subway. $\hat{\beta_0}$ represents the intercept of this linear regression. More specifically, excluding the effect of the day, at midnight the expected subway delay is 6.342 minutes. As for $\hat{\beta_2}, ..., \hat{\beta_8}$, these represent the change in the expected subway delay, given the time is held constant on the certain days. As a blanket statement, if its the i^{th} day of the week (like Sunday being $i = 1$), then the change in the expected subway delay, assuming time is held constant, is $\hat{\beta_{i+1}}$. For example, Friday ($i = 7$) has no change in the expected subway delay, but if the day is changed from Friday to Saturday ($i = 7$) then we expect a change in delay of $\hat{\beta_{i+1}} = \hat{\beta_8} = -4.609$ minutes assuming the time does not change on each day we are trying to find the expected delay. **Figure 4** models this hypothesized relationship below.

As seen in **Figure 4**, it is possible to notice the model slightly adhering to the distribution of the data validating the linearity assumption, however its clearly noticeable that the model almost always over estimates the amount of subway delay for every day. As a result, its is necessary to check the other assumptions made previously that allow for a linear regression.

As one can notice in **Figure 5**, the residual plot does appear to be clumped in some places and contains a few outliers, however for a majority of the data, the residual plot appears to contain no pattern thus supporting the assumptions for a linear model. Continuing to the Q-Q plot, the slight curve indicates slightly skewed data making it hard to justify the assumption that the error terms are normally distributed. However, given that only a few points deviate extremely from the line, as a whole we can assume the assumption holds and note the existence of outliers. Given the mostly horizontal line on the scale-location graph, the assumption for constant variance holds. As for the leverage graph, there do appear to be points that could be removed due to their low leverage and high residual value, but the chart itself doesn't speak to the assumptions made about using a linear regression. Thus as a whole the assumptions appear to hold for the use of a linear regression, thus, now whether there actually exists a linear relationship, a hypothesis test needs to be conducted on the coefficients.

In order to verify that any predictor in this regression is actually statistically significant, an ANOVA F test is conducted. The F statistic was 1.7150038 with 7 and 312 degrees of freedom. Thus the calculated p-value is 0.1047899. Given that this p-value is not less that the predetermined $\alpha = 0.05$, there is no statistically significant relationship between the estimated amount of delay and the time or day. As a result, a linear regression fails to be of use in this situation anyways, implying that there is no need for any further model selection

All analysis for this report was programmed using `R version 4.1.1`. I used the `lm()` function in base `R` to derive the estimates of a frquentist logistic regression in this section [4].

Conclusions

Subway delays not only slightly inconvenience the rider but also have drastic economic implications when looking at the systems within large systems. As seen in New York, subway delays could cause thousands of dollars daily, and even though Toronto isn't New York, some level of equivalence can be made between given the wide spread use of the subway system in Toronto for workers to commute to their job. As a result, subway delays might also have a significant economic impact, and by both identifying and addressing the factors that cause some of the delay will both relieve the economic impact as well as benefit the commuters. To an attempt to identify any underlying events that cause subway delay, the time and day of the week were used in an attempt to predict the amount of delay. It was hypothesized that as the time approaches rush hour on a weekday, the expected amount of subway delay would increase. This was hypothesized as during this time, more people would be using the subway and more delay causing events could occur.

After fitting a regression, there wasn't enough evidence to state that there were any significant relationship between the time, day of the week, and the expected about of subway delay. What this means, is that regardless of the time of day (including rush hour), and the day of the week it is, all of the experienced delay was rather random. This hints at the fact that there were no underlying events that were causing an increased amount of subway delay rooted in the day of the week it was (like a work day) and what time it was. As a result, it is believed that there is no need to improve the subway system with regards to the how it (as a whole) deals with the influx of people during rush hour during any given day.

Weaknesses

Given that the TTC didn't publish their data collection methodology, any bias that would've arisen in the data collection process is hidden. As a result, all of the results ascertained have to be taken under the assumption that the TTC created a system without any bias. Furthermore, the results were only derived based on data from 2021 in September. As reported, TTC ridership during September of 2021 is still far below [6] previously recorded ridership pre-COVID-19. As a result, the data used represents a period of time where TTC usage isn't representative of a society not suffering from a pandemic, therefore the applicability of the results are limited to the duration of COVID-19 restrictions. After restrictions are lifted and there is a noticeable increase in subway ridership, then there could be a dramatic change from the results shown here. Thus, as a whole the main weaknesses of this study lie in its blind faith in the lack of bias in data collection, and the the fact that it won't be representative of the subway systems in 1-2 years.

Next Steps

Given that this analysis soley looked at how the time of day and what day it is impacted subway delay, further studies could be done to determine if other factors predict the amount of subway delay. For example, if the location is a statistically significant explainer for the amount of subway delay, then more research should be done surrounding that specific location to alleviate subway delay. Furthermore, analysis could be done over older subway delay data. This data could provide an insight into pre-pandemic subway delays, addressing one of the main weaknesses of this analysis. As a whole, there are numerous plausible next steps as the number of factors that could possibly contribute to subway delays is almost limitless and further investigation into them could lend to useful results in an attempt to lower overall delays.

Discussion

As a whole, it is incredibly important to try and find ways to reduce subway delays. Given their economic impact and overall annoyance to average commuters, finding factors that could explain subway delay could help the necessary parties find ways to remove those factors adding to subway delays. Even though this analysis concluded that the time and day of week was of no importance to subway delays, it removes two factors from a pool of possibly hundreds that could infact be adding to the subway. Even though the results derived here are primarily applicable to a pandemic level of riders (dramatically less than there used to be), these levels are expected to continue for a while longer, and thus even finding reasons for delays now are beneficial to reducing the economic impact that subway delays deal.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: October 12, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: October 12, 2021)
4. Peter Dalgaard. (2008) *Introductory Statistics with R, 2nd edition.*
5. Office of the New York City Comptroller. (2017, October). The Economic Cost of Subway Delays. City of New York. Retrieved October 20, 2021, from https://comptroller.nyc.gov/wp-content/uploads/documents/The_Economic_Cost_of_Subway_Delays.pdf.
6. Fox, C. (2021, September 15). TTC ridership could linger below pre-pandemic levels for at least another two years: Report. CTV News Toronto. Retrieved October 20, 2021, from https://toronto.ctvnews.ca/ttc-ridership-could-linger-below-pre-pandemic-levels-for-at-least-another-two-years-report-1.5586754.
7. Introduction to subway ridership. mta.info. (2018). Retrieved October 20, 2021, from http://web.mta.info/nyct/facts/ridership/.
8. Dickens, M., American Public Transportation Association (2019). Retrieved October 20, 2021, from https://www.apta.com/wp-content/uploads/2019-Q1-Ridership-APTA-1.pdf.
9. Toronto Transportation Commission. TTC.ca. (n.d.). Retrieved October 20, 2021, from https://www.ttc.ca/.
10. Toronto Transportation Commission. (2021, October 21). TTC Subway Delay Data. Retrieved October 19, 2021, from https://open.toronto.ca/dataset/ttc-subway-delay-data/.

Appendix

Table 2: Numerical statistics for the delay in minutes the time in hours

Value	Delay	Time
Mean	4.446875	7.020990
Median	3.000000	7.966667
Variance	52.003439	13.632342

Table 3: Number of data points for each day

Day	Count	Mean Delay
Friday	43	6.883721
Monday	41	4.634146
Saturday	33	2.212121
Sunday	58	5.603448
Thursday	57	4.017544
Tuesday	39	3.230769
Wednesday	49	3.755102

Figure 1. Distribution of the hours past midnight with delay occurances

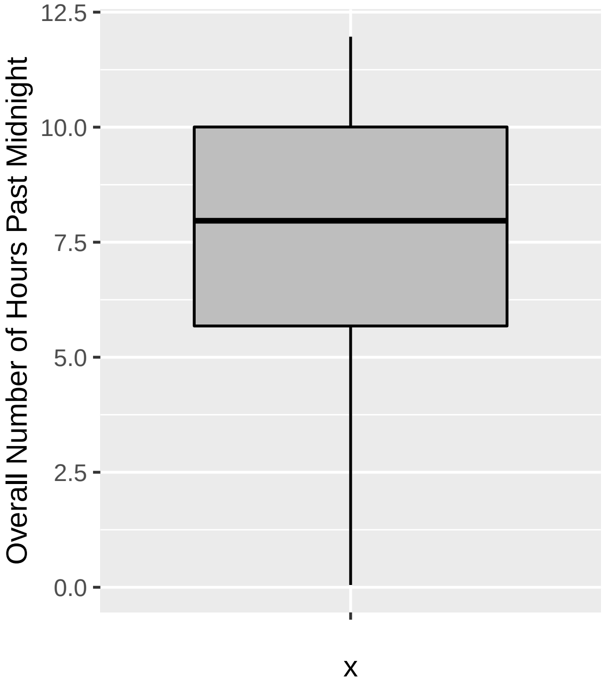


Figure 2. Distribution of the minutes of delay

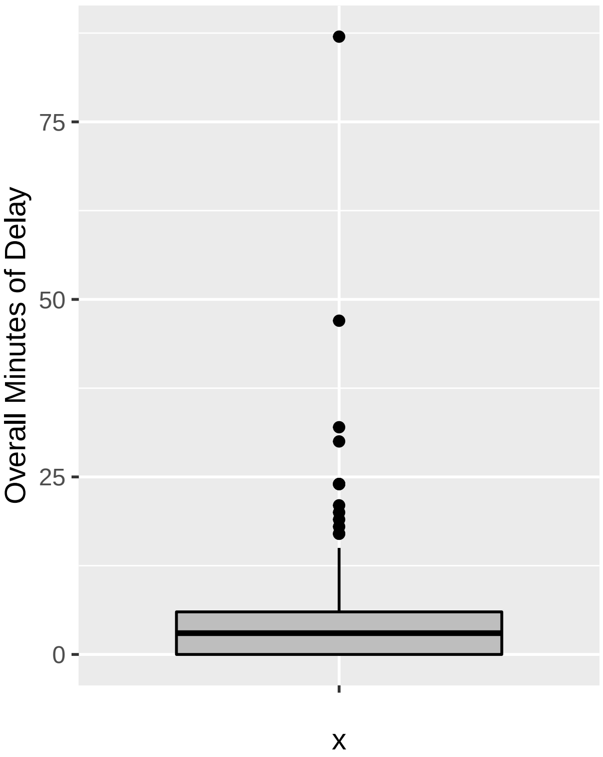


Figure 3. Relationship between the hours past midnight and subway delay

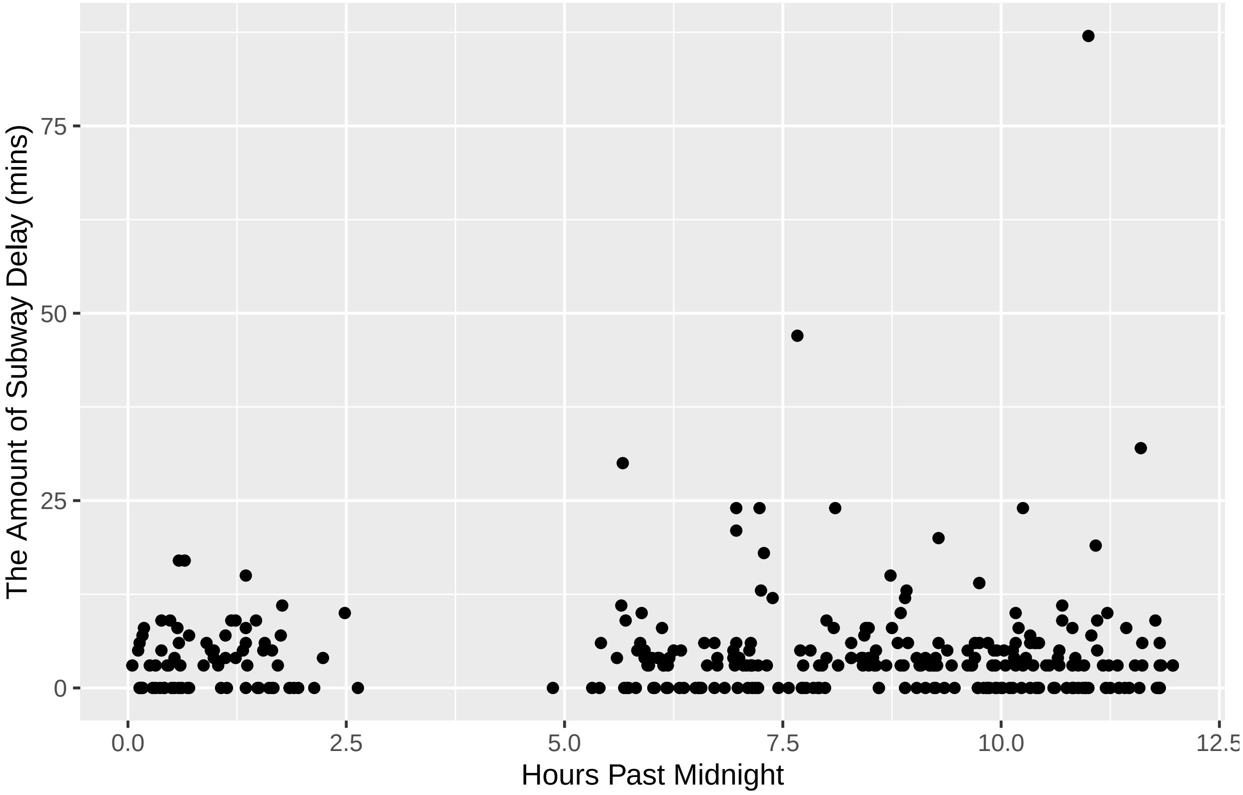
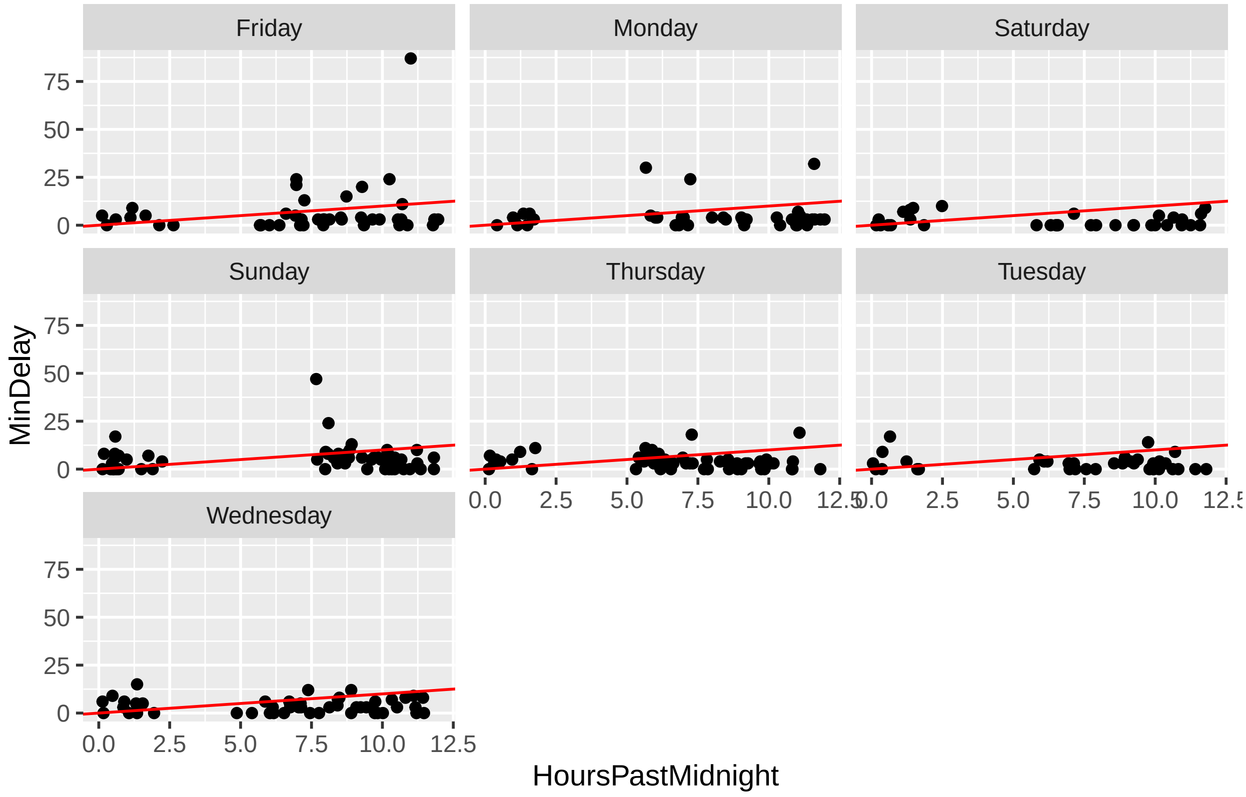


Figure 4. Calculated linear regression overlayed the scatterplots



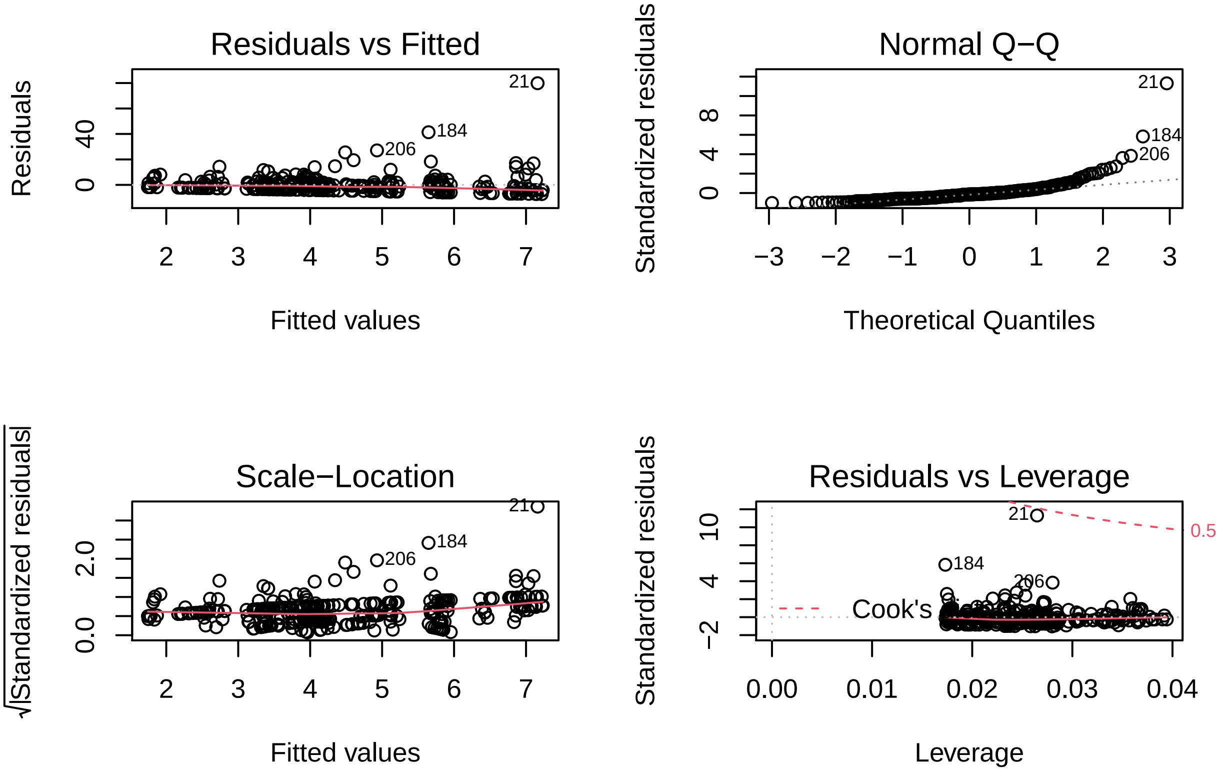


Figure 5. [Above] Linear model assumption checking plots