Toronto Transit Commission (TTC) Streetcar Delay

Ruxin Liu

11/5/2020

Abstract

In the city of Toronto, it is very common to experience streetcar delays and it would be helpful to understand what factors could impact the delayed time. In order to answer the question, a multilevel negative binomial model is fitted in this study. The results found that both the delay-causing incident and the delay-happening day have impact on how long the delay in minutes will be. After improving this model, it will be useful for passengers to have a general idea about their waiting time.

Introduction

When walking in the street in Toronto, Canada, it is very common to see the red and white streetcars driving around. These streetcars are operated by the Toronto Transit Commission (TTC) and there are in total 10 different streetcar routes, which bring large convenience and efficiency to the transportation system in the city of Toronto (Wikipedia 2020). However, as many other public traffic systems do, TTC streetcars also have frequent delays. The City of Toronto's Open Data Portal has collected TTC streetcar delay information from the year 2014 to the year 2020, which includes the location, the incident, the date, the route, the length of the delay and etc (Open Data Portal 2020).

In this study, the major question is that what factors could potentially impact the delayed time. In order to answer this, a multilevel negative binomial regression is fitted to the data to explore the relationships between the delayed incident, the delayed day, and the length of the delay, while accounting for different streetcar routes.

Method

Data Content

The data was downloaded from The City of Toronto's Open Data Portal and was cleaned and processed in R (all the detailed codes are in the file: Supplementary R code.Rmd). The data was subsetted on the 10 routes that are part of the Toronto streetcar system, which are route number 501, 503, 504, 505, 506, 508, 509, 510, 511 and 512. In total, there are 76886 observations. Table 1 below displays the major information collected in this data, where Min.Delay is the length of the delay in minutes to the schedule for the following streetcar.

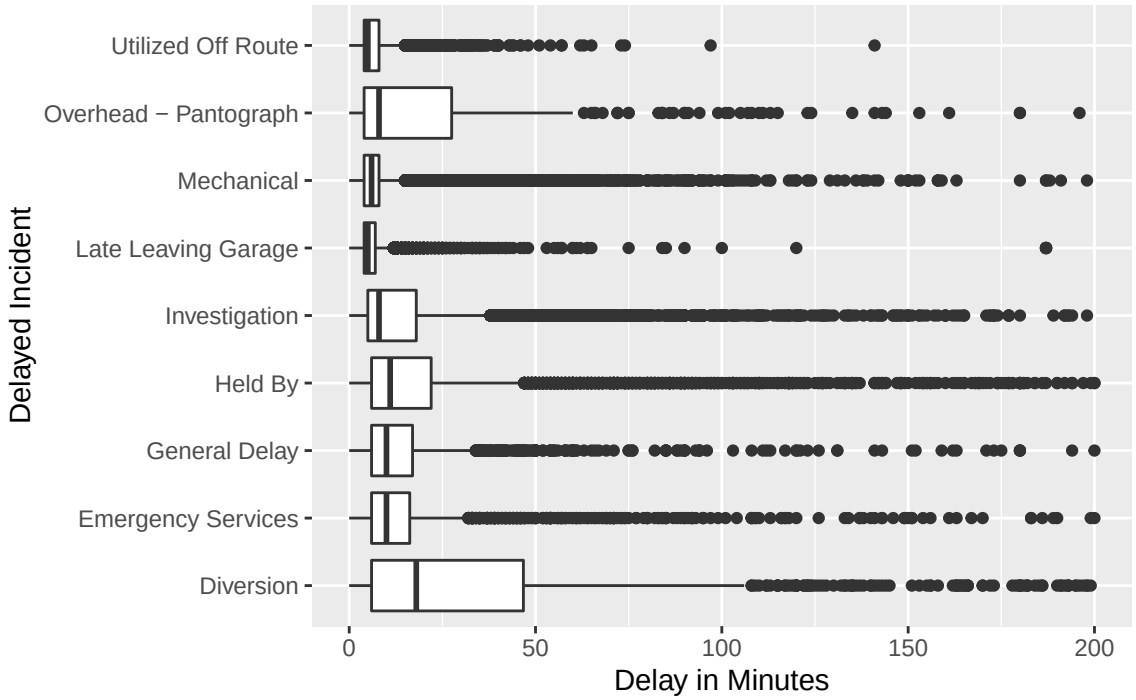
Variable Selection

In order to explore the potential factors that can impact the delayed time, exploratory data analysis (EDA) is performed. From Fig.1, it is very clear that the distribution of the delayed time related to different delay-causing incidents varies a lot, especially when there is a diversion happening, the delayed time is much longer compared to the other incidents. Also, lots of the outliers are observed from the plot.

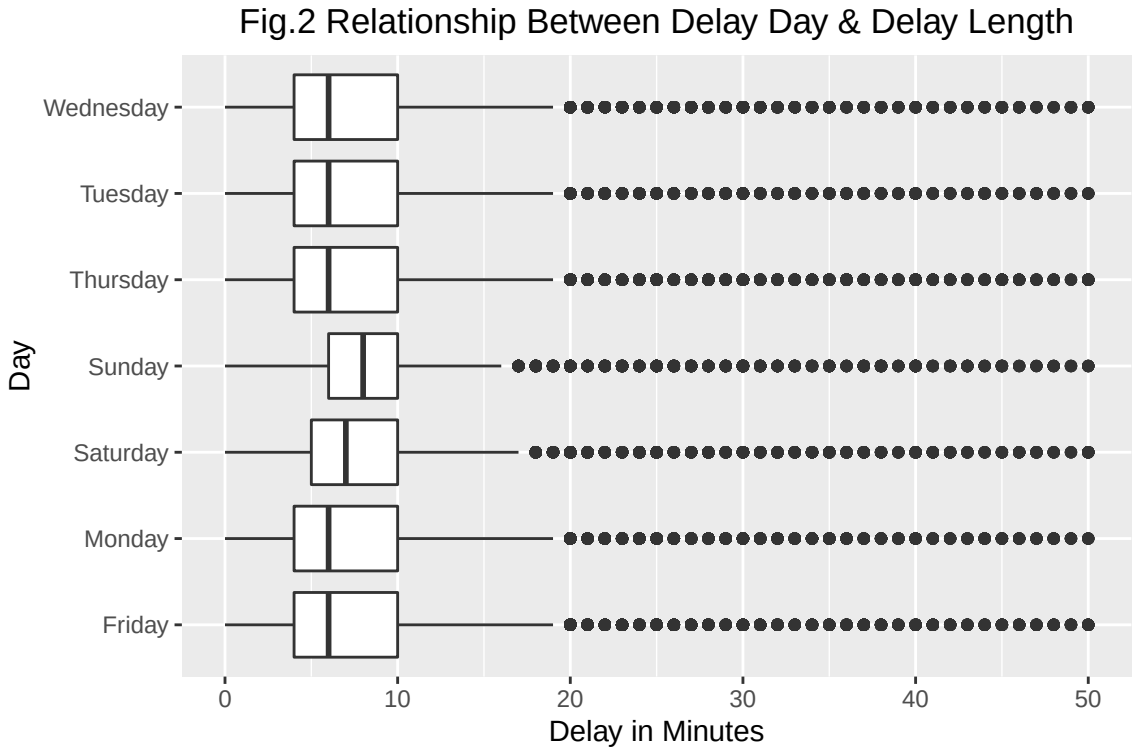
Table 1: Data Content

	Report.Date	Route	Time	Day	Incident	Min.Delay
1	2014-01-02	505	6:31:00 AM	Thursday	Late Leaving Garage	4
500	2014-03-02	504	1:46:00 PM	Sunday	Investigation	6
1000	2014-03-13	506	8:17:00 PM	Thursday	Held By	31
9900	2014-12-07	506	10:58:00 PM	Sunday	Mechanical	8
13000	2015-02-23	501	10:32:00 AM	Monday	Investigation	24

Fig.1 Relationship Between Incident & Delay Length



From Fig.2, although for all 7 days in the week, there are many outliers and the 3rd quartiles are relatively at the same level, the 1st quartiles of Saturday and Sunday are higher compared to the other days. Therefore, an variable "weekend" is created to indicate whether the day is a weekend or not to capture the difference in streetcar delay time between weekdays and weekends.

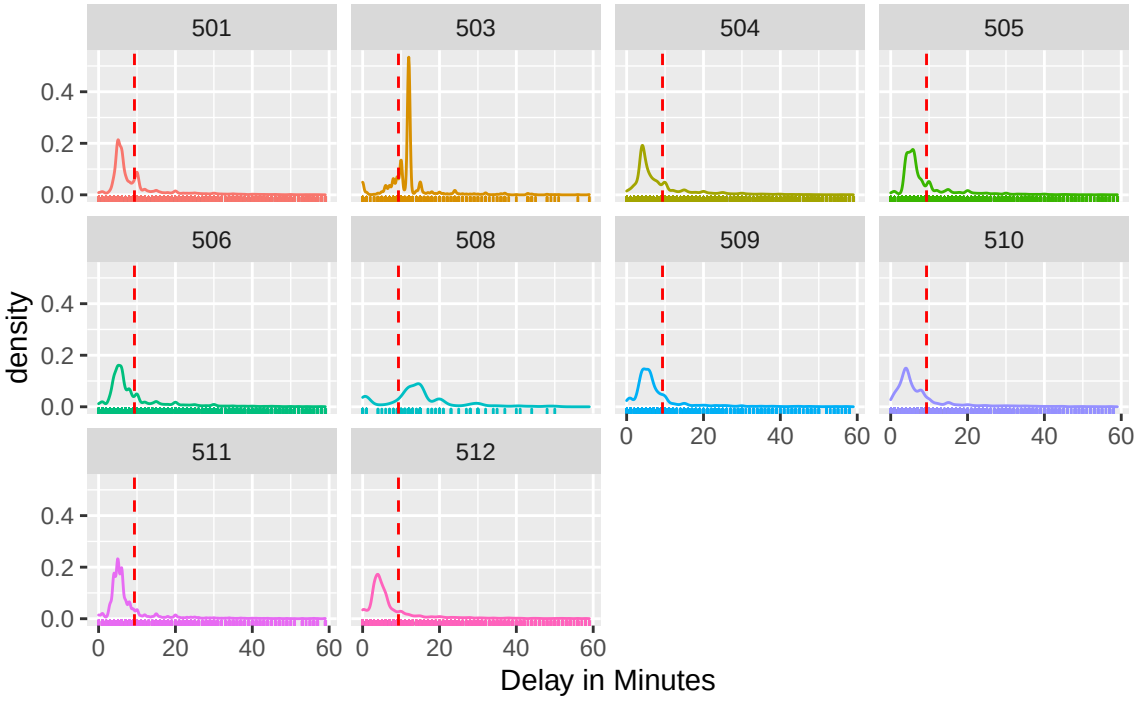


Model Selection & Validation

The response variable in this project is the length of delay in minutes (Min.Delay), which is a continuous and numerical variable. However, the data does not have a normal distribution even after transformations, and since the delayed time is recorded to the nearest minute, the variable Min.Delay performs like a discrete variable. Therefore, the response variable is considered as the count data (the number of delayed minutes) and the negative binomial regression is performed instead of the linear model. The negative binomial regression includes the incident and weekend variables as the predictors and they are all statistically significant at the α level of 5%, which confirms with the EDA that they are important factors to consider. And by checking the residuals and plotting Rootograms, the negative binomial model is confirmed to be the more appropriate model for this data (The detailed model validation process is in the Appendix).

From Fig.3, it is clear that the distributions of the delayed time are quite different among the 10 routes, where the red dashed line is the mean minutes of delay. In order to capture this in the model, each route will have its own intercept. Therefore, the final model for this study is a multilevel negative binomial regression fitted with the *lme4* package in R, where the delayed time in minutes is the response, delayed incident and weekend indicator are the predictors, with each route having a varying intercept (glmer.nb(Min.Delay ~ Incident + weekend + (1 | Route))).

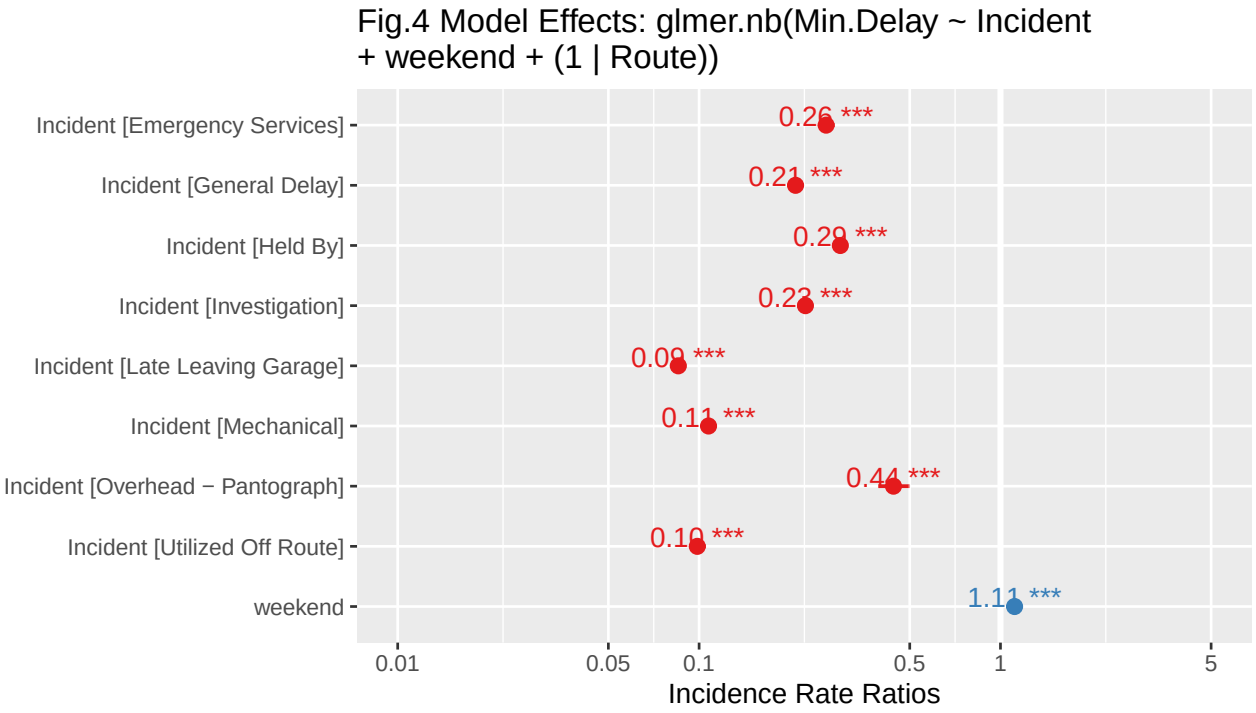
Fig.3 Distribution of the delay length among different routes



Result

Based on the results of the multilevel negative binomial model, each route has a slightly different intercept and the estimated average intercept is 4.34, which means that when the delayed incident is Diversion (the reference group) and the delay happens on weekdays (weekend = 0), the delayed time is expected to be 76.8 minutes ($e^{4.34} = 76.8$). From Fig.4, the effects and their 95% confidence intervals of the model are shown, where the effects are calculated by taking the exponential of the estimated coefficients, and all the effects are statistically significant at α level of 5%.

For all other incidents, the estimated delayed time is shorter compared to when there is a diversion, which is consistent with the patterns shown in Fig.1. For example, when holding other variables constant, if the delayed incident is General Delay, the delayed time is expected to be 79% (1 - 0.21) shorter compared to the delayed incident of Diversion. When holding other variables constant, if the delay happens on weekends (weekend = 1), the delayed time is expected to be 11% (1.11 - 1) longer compared to weekdays, which is consistent with the patterns shown in Fig.2. Therefore, the delay-causing incident and the day of the delay both have impact on the streetbus delayed minutes.



Discussion

Based on the EDA and the results of the multilevel negative binomial model, there is strong evidence that the delay-causing incident and the delay-happening day both have influence on how long the delay will be. More specifically, when the delay is due to diversion, the average delayed time in minutes is expected to be the longest, while when the delay is due to late leaving garage, the average delayed time in minutes is expected to be the shortest. Also, when the delay happens on weekdays, the average delayed time in minutes is expected to be shorter than on weekends. Since this study performs a multilevel model with random intercepts for all 10 routes, we can also know the difference between the routes. With this conclusion, it might be helpful for passengers to have a general idea about how long they need to wait for the following streetcar.

Future Directions

- Although the current multilevel negative binomial model is the most appropriate one after exploring other options, the residual plot and the Rootogram both suggest that the model could be improved more. One future improvement could be doing research and setting a delayed time cut-off point, which indicates whether the delayed time is long or short, and then try to fit the multilevel logistic model.
- The other improvement could be done for the future is to use the Gaussian approach and fit the model with stan functions, which allows posterior predictive checks.
- The original data set does not contain lots of information. For future studies, one improvement could be adding in other potential confounding variables, such as the weather.

Appendix

EDA

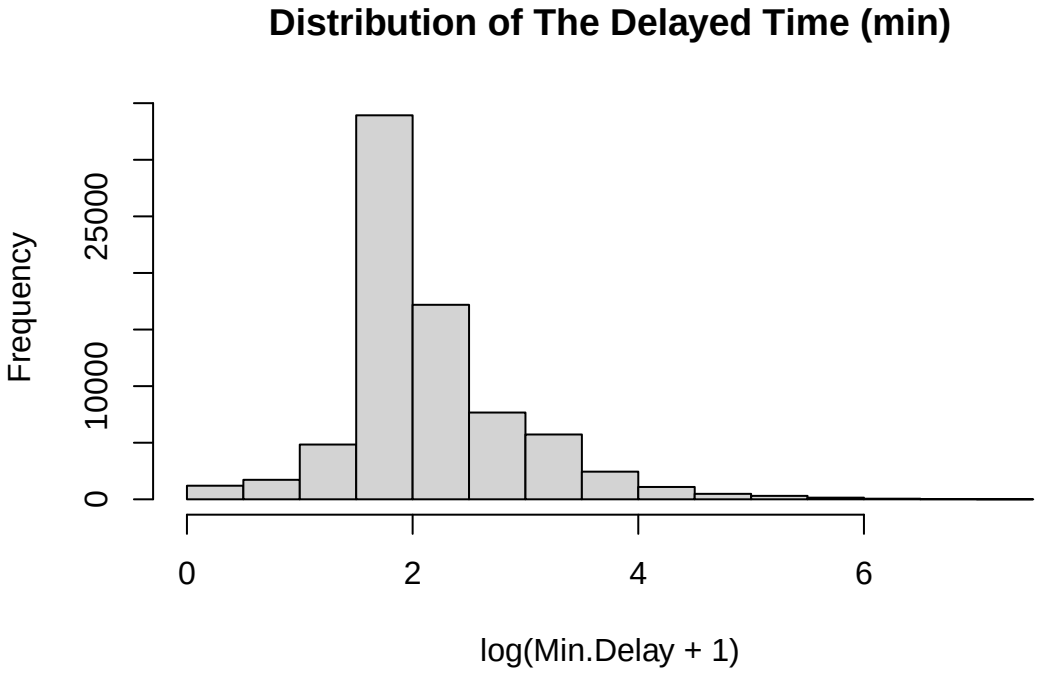


Figure 1: The distribution of the delayed time (Min.Delay) is not very normally distributed after log transformation, therefore suggesting that a linear model is not appropriate for this data.

Table 2: The average delayed length and the maximum delayed length both vary a lot among different incidents, which suggests that incident is a important variable to consider.

Incident	mean(min)	mean(hr)	max(min)	max(hr)
Diversion	77.00	1.28	1294	21.57
Emergency Services	20.03	0.33	999	16.65
General Delay	15.70	0.26	1008	16.80
Held By	22.37	0.37	999	16.65
Investigation	16.75	0.28	1400	23.33
Late Leaving Garage	6.66	0.11	600	10.00
Mechanical	8.23	0.14	1294	21.57
Overhead - Pantograph	30.19	0.50	358	5.97
Utilized Off Route	7.57	0.13	141	2.35

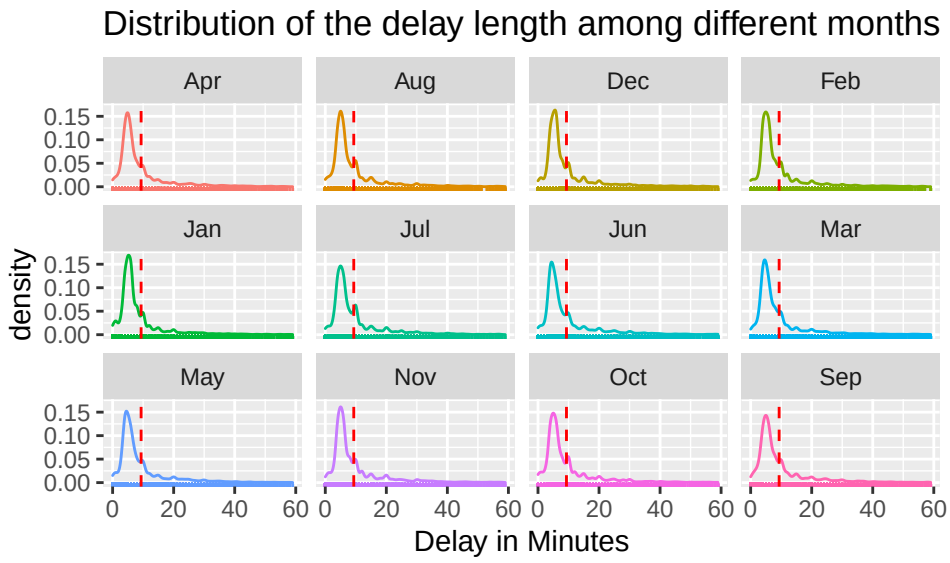


Figure 2: From this plot, there is no obvious difference between the 12 months, therefore the month is not used for grouping in the model.

Model Fitting & Validation

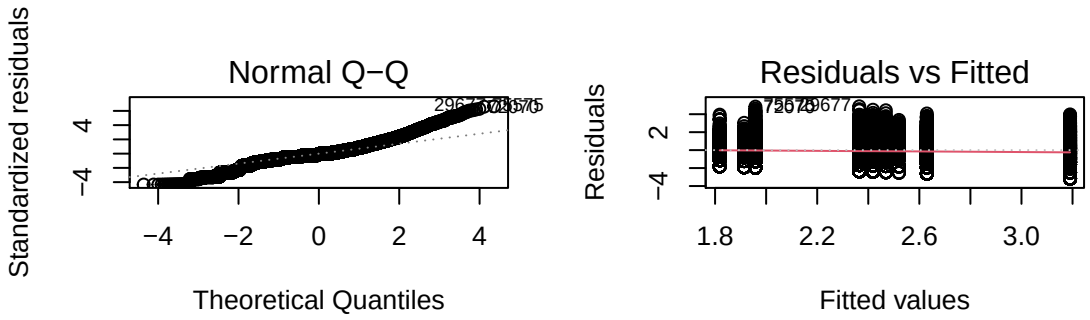


Figure 3: Initial Linear Regression: lm(log(Min.Delay + 1) - Incident). Assumptions are violated, suggesting that linear model is not appropriate.

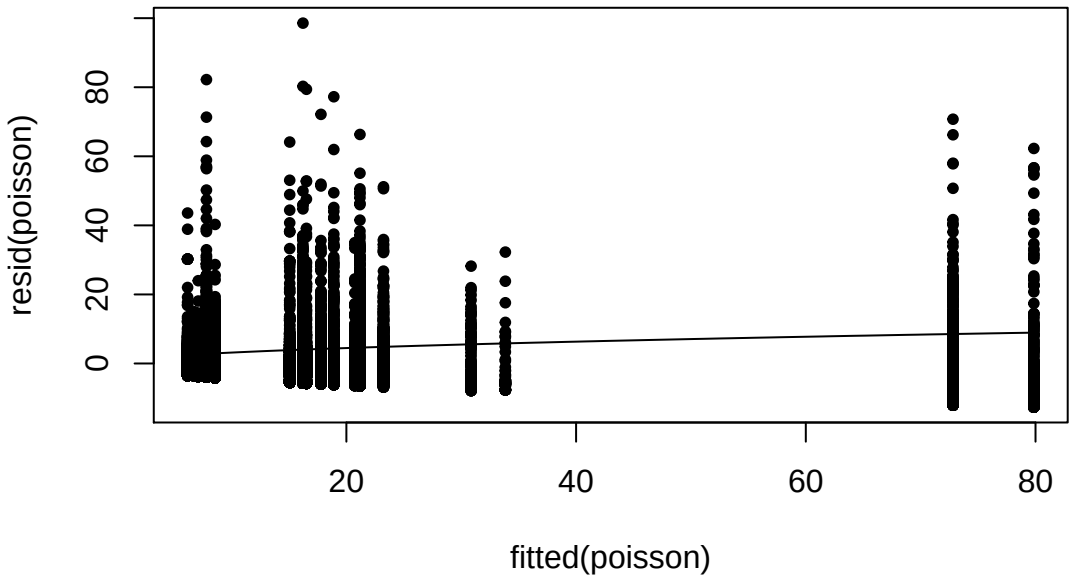


Figure 4: In the initial poisson regression: glm(Min.Delay - Incident + weekend, family = poisson), the model's standard deviation is expected to have a square root relationship with the fitted mean, which can be compared with the curve in the plot. Since the majority of the points fall above the line, it indicates the existence of overdispersion, and therefore poisson model may not be the best.

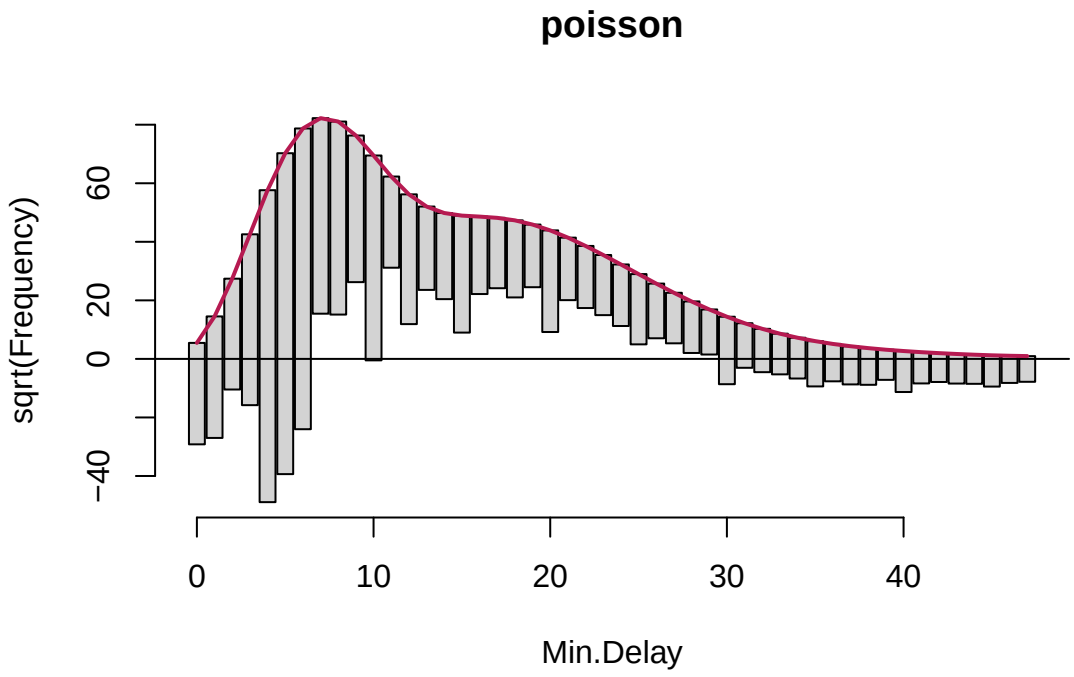


Figure 5: The departures from expected counts are considerably large, again suggesting the poisson model is not the best.

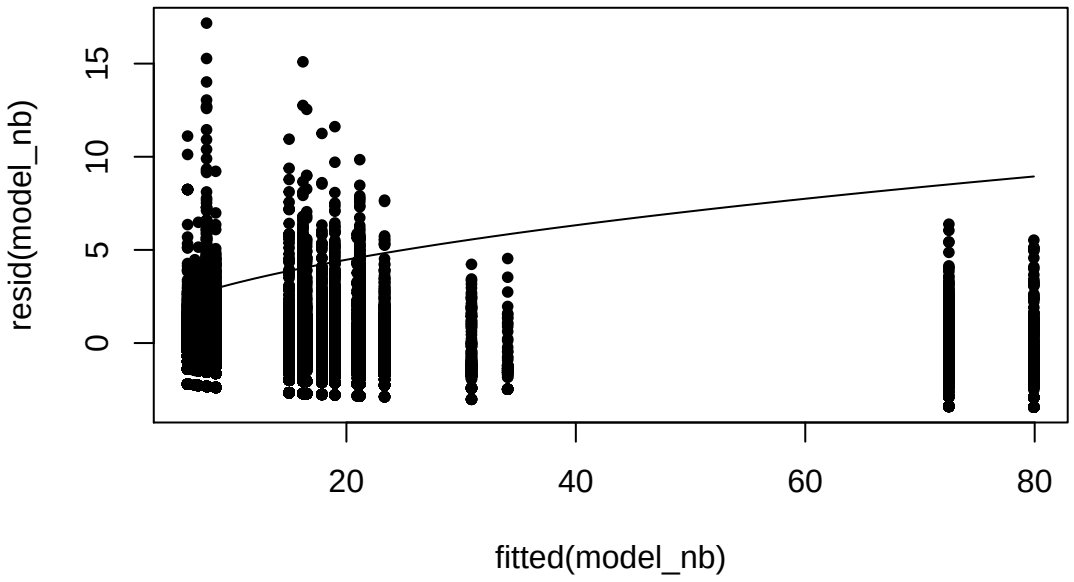


Figure 6: The initial negative binomial model is glm.nb(Min.Delay - Incident + weekend). Although there are still some points that fall above the line, the overdispersion has been fixed a lot compared to the poisson model, which suggests that the negative binomial model works better.

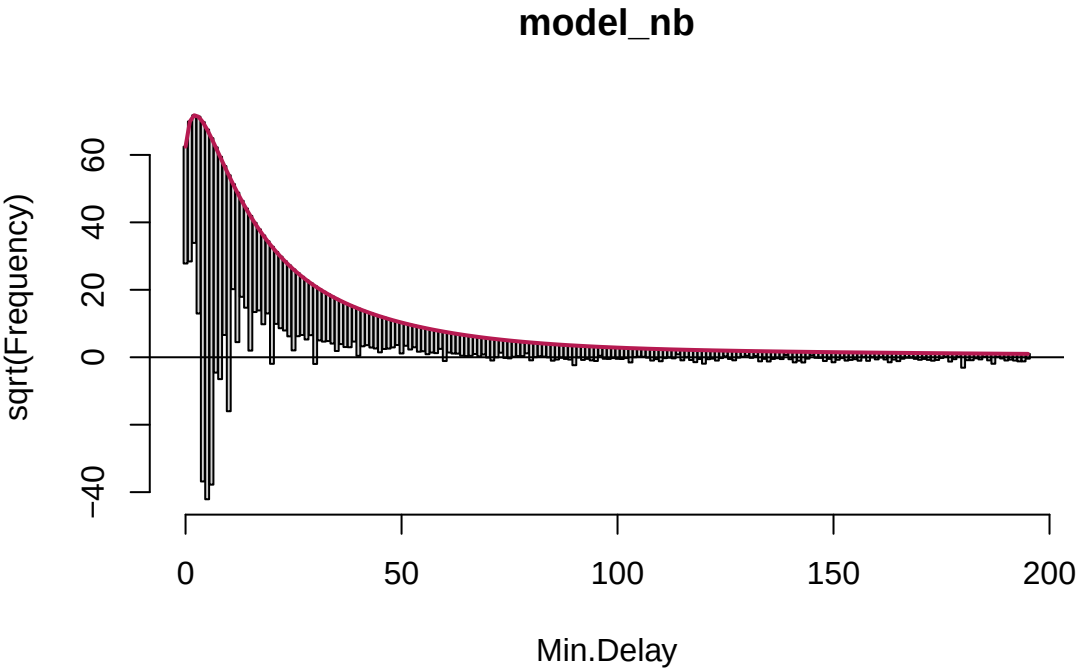


Figure 7: Although the roorogram does not look perfect, compared to the poisson model, the departures from expected counts are much smaller with the negative binomial model.

\begin{table}

\caption{The 95% confidence interval for the coefficients of the initial negative binomial model.}

	Estimate	2.5 %	97.5 %
(Intercept)	4.2838787	4.2344520	4.3340870
IncidentEmergency Services	-1.3396918	-1.3990500	-1.2808002
IncidentGeneral Delay	-1.5754069	-1.6301800	-1.5212637
IncidentHeld By	-1.2315364	-1.2845620	-1.1791990
IncidentInvestigation	-1.4981604	-1.5506809	-1.4463454
IncidentLate Leaving Garage	-2.4662586	-2.5204038	-2.4127697
IncidentMechanical	-2.2286676	-2.2794834	-2.1786135
IncidentOverhead - Pantograph	-0.8529988	-0.9692475	-0.7340489
IncidentUtilized Off Route	-2.3244533	-2.3841294	-2.2652542
weekend	0.0972311	0.0823506	0.1121420

\end{table}

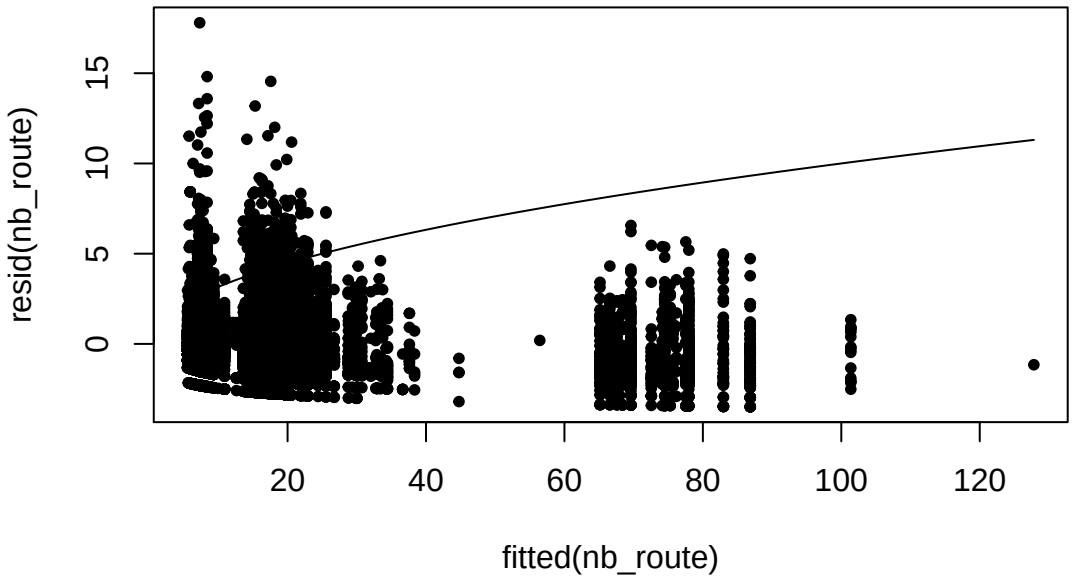


Figure 8: The final multilevel negative binomial regression model is glmer.nb(Min.Delay - Incident + weekend + (1 | Route)). Although there are still some overdispersion at the lower counts, the residuals overall look fine.

```
# Model Summary
summary(nb_route)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: Negative Binomial(1.4869)  ( log )
## Formula: Min.Delay ~ Incident + weekend + (1 | Route)
##    Data: delay_final
##
##       AIC       BIC    logLik  deviance  df.resid
##  521570.8  521681.8 -260773.4  521546.8     76804
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
##  -1.209  -0.605  -0.359   0.036 135.822
##
```

```
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Route  (Intercept) 0.04449  0.2109
## Number of obs: 76816, groups:  Route, 10
##
## Fixed effects:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    4.341164   0.069280   62.66   <2e-16 ***
## IncidentEmergency Services    -1.332035   0.029384  -45.33   <2e-16 ***
## IncidentGeneral Delay         -1.565737   0.027000  -57.99   <2e-16 ***
## IncidentHeld By               -1.223542   0.026074  -46.92   <2e-16 ***
## IncidentInvestigation         -1.490216   0.025829  -57.70   <2e-16 ***
## IncidentLate Leaving Garage   -2.461434   0.026696  -92.20   <2e-16 ***
## IncidentMechanical            -2.230405   0.024935  -89.45   <2e-16 ***
## IncidentOverhead - Pantograph -0.817444   0.058297  -14.02   <2e-16 ***
## IncidentUtilized Off Route    -2.316469   0.029603  -78.25   <2e-16 ***
## weekend                        0.107766   0.007599   14.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) IncdES IncdGD IncdHB IncdnI IncLLG IncdnM IncO-P IncUOR
## IncdntEmrgS -0.277
## IncdntGnrlD -0.300  0.752
## IncdntHldBy -0.312  0.780  0.849
## IncdntInvst -0.316  0.788  0.858  0.889
## IncdntLtLvG -0.308  0.763  0.830  0.861  0.870
## IncdntMchnc -0.328  0.816  0.889  0.922  0.931  0.901
## IncdntOvr-P -0.131  0.338  0.367  0.381  0.385  0.374  0.399
## IncdntUtlOR -0.277  0.686  0.748  0.775  0.783  0.758  0.812  0.336
## weekend     -0.029  0.022  0.006  0.022  0.024  0.043  0.028  0.009  0.027
```

```
# Random effects
ranef(nb_route)
```

```
## $Route
##     (Intercept)
## 501  0.01521893
## 503  0.27766104
## 504 -0.09822933
## 505 -0.14305104
## 506 -0.03053647
## 508  0.50945769
## 509 -0.11591307
## 510 -0.16454620
## 511 -0.12764112
## 512 -0.12214203
##
## with conditional variances for "Route"
```

Bibliography

Online Resources:

1. Wikipedia. (2020). *Toronto streetcar system* [online]. Available from: https://en.wikipedia.org/wiki/Toronto_streetcar_system#Route_numbers [accessed 29 November 2020]

2. Open Data Portal. (2020). *TTC Streetcar Delay Data* [online]. Available from: https://open.toronto.ca/dataset/ttc-streetcar-delay-data/ [accessed 5 November 2020]

R package:

1. Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

2. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

3. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

4. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

5. Hao Zhu (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra

6. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

7. Lüdecke D (2020). sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.6, <URL: https://CRAN.R-project.org/package=sjPlot>.

8. Christian Kleiber, Achim Zeileis (2016). Visualizing Count Data Regressions Using Rootograms. The American Statistician, 70(3), 296–303. doi:10.1080/00031305.2016.1173590