Improving MTA rider experience via analyzing contributing factors of additional wait times

Grace He, Christopher Lee

December 28, 2022

1 Introduction

While MTA transportation can be convenient for New Yorkers, delays and failures often frustrate riders. Oftentimes, the displayed wait times for subways are incorrect and riders end up waiting longer than they are promised. Moreover, some riders are even sometimes stranded between stations as unprecedented failures arise. Through these issues, MTA subways end up being an unreliable source of transportation. With the limited resources that the MTA has in labor and materials, decision-makers have to efficiently choose where to focus their efforts in order to meet the public needs as well as to maximize their own profits.

In this study, we aim to solve both fronts of this problem. We estimate models for the additional wait times passengers experience and how they translate to revenue gain or loss for the decision maker. We are thus able to provide wait time estimates as well as optimize the revenue gained based on the data given, providing value and crucial information for both consumers and decision-makers.

2 Methodology

Overview We run two regression models with covariates for subway additional platform time (APT), additional train time (ATT). More specifically, we assume APT and ATT to follow Weibull distributions as the values are both continuous and positive[1]. We use the MTA Subway Customer Journey-Focused Metrics: Beginning 2015 dataset.

2.1 Data Preprocessing The featured datasets include data from January 2015 to August 2022. However, we suspect that due to the circumstances surrounding the COVID-19 pandemic, the year 2020 will present data that is uncharacteristic of "normal" years. We therefore omit data from dataset that falls within 2020.

We log-scaled num_passengers to allow our models to converge more consistently. We also incorporate seasonality by appending dummy variables to the existing datasets. More precisely, we add dummy variables for 3 of the seasons (spring, summer, and fall) and use the remaining season (winter) as a "baseline" covariate.

In total, there are 8 relevant columns: division, period, log_num_passengers, spring, summer, autumn, APT, and ATT. Division, period, spring, summer, and autumn are all vectors containing indicator variables[2].

[1] As it turns out, some of the values in ATT are negative. However, we account for this by shifting the data by some positive amount δ. More details in section 2.3.

[2] 1 for an entry in division means division B, and 0 means division A. 1 for an entry in period means peak, and 0 means offpeak.

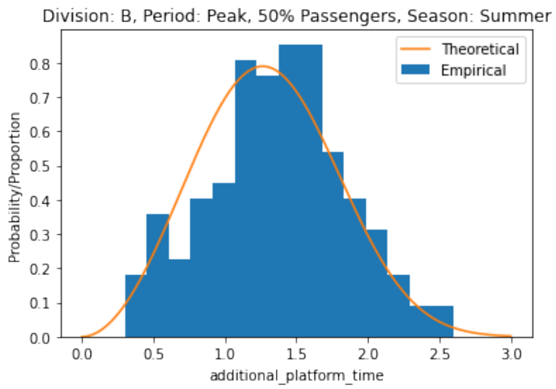


Figure 1: The resultant plot for the distribution of APT given x overlayed on a histogram of the empirical data.

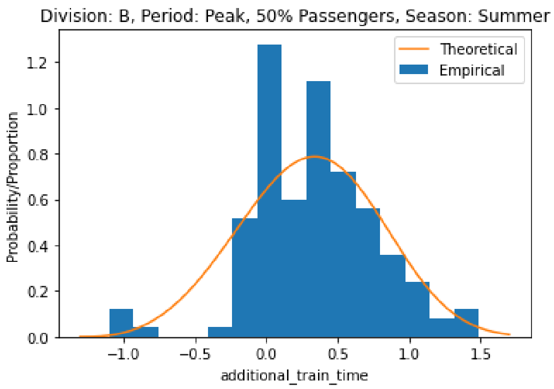


Figure 2: The resultant plot for the distribution of ATT given x overlayed on a histogram of the empirical data.

For the follow sections, assume we have the design matrix X with the following order of columns:

$$X = \begin{bmatrix} \mathbf{1} & \text{division} & \text{period} & \text{log\_num\_passengers} & \text{spring} & \text{summer} & \text{autumn} \end{bmatrix},$$

where $\mathbf{1}$ denotes the ones vector.

2.2 Additional Platform Time We assumed APT $\sim Weibull(c, \lambda)$ and attempted to recover estimates for the parameters c and λ. We set $c = \exp(X\beta_c)$ and $\exp(X\beta_\lambda)$, where β_c, $\beta_\lambda \in \mathbb{R}^7$ (since $X \in \mathbb{R}^{n\times(d+1)}$, where n is the number of rows and d is the number of covariates). Below are the values we recovered[3] for β_c and β_λ:

	$\beta^{(0)}$	$\beta^{(1)}$	$\beta^{(2)}$	$\beta^{(3)}$	$\beta^{(4)}$	$\beta^{(5)}$	$\beta^{(6)}$
β_c	-0.5034	0.0811	-0.0962	0.1210	0.0153	-0.2099	-0.0148
β_λ	-0.1231	-0.3498	0.1676	-0.0653	0.1040	0.1152	-0.0555

With these values, we could estimate APT for a specific subset of people. For example, if we wanted to find the distribution of APT for people taking the subway in division B during a peak period when there are the median number of passengers in the summer, we would have the following:

$$x = \begin{bmatrix} 1 & 1 & \log M & 0 & 1 & 0 \end{bmatrix},$$

where M is the median number of passengers. Then we would have our crystallized parameters as $c = \exp(x \cdot \beta_c)$ and $\lambda = \exp(x \cdot \beta_\lambda)$. Figure 1 is a plot of the resultant distribution overlayed on a histogram of the empirical data matching this specific subset of people.

Now suppose we wanted to find the likelihood that the subway arrives given that it hasn't arrived since time t. We can find this by simply using the hazard function h, where $h(t)$ is defined as the following for the Weibull distribution:

$$h(t) = c\lambda t^{c-1}.$$

Figure 3 displays the hazard for the given x. We observe that as more time passes, the likelihood of the subway arriving increases, which intuitively makes sense.

[3] See Appendix A.1 for all raw model outputs.

2.3 Additional Train Time Since there are negative values present in the ATT column of the dataset, we shifted the ATT values by δ, where

$$\delta := \left| \min_{k \in \text{ATT}} k \right| + \epsilon$$

for some small $\epsilon > 0$. For our experiments, we used $\epsilon = 1 \times 10^{-1}$. After running the regression, we shifted the predicted values by the same amount.

Similar to APT, we assumed ATT $\sim Weibull(c, \lambda)$. We ran the Weibull regression with covariates with parameters $c = \exp(X\beta_c)$ and $\lambda = \exp(X\beta_\lambda)$, where β_c, $\beta_\lambda \in \mathbb{R}^7$. Below are the values we recovered for β_c and β_λ:

	$\beta^{(0)}$	$\beta^{(1)}$	$\beta^{(2)}$	$\beta^{(3)}$	$\beta^{(4)}$	$\beta^{(5)}$	$\beta^{(6)}$
β_c	-0.0563	-0.2203	-0.1589	0.1102	0.0097	-0.0082	-0.1766
β_λ	-0.0015	0.1057	0.2849	-0.1694	0.0164	0.0499	0.1861

Again, we can now use these values to estimate the distribution of ATT for a specific subset of people. Using the same x as in section 2.2, we obtain the plot in Figure 2.

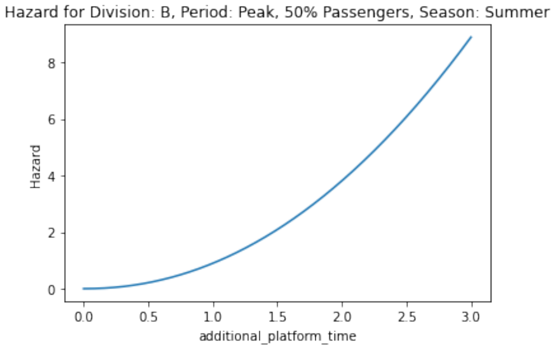


Figure 3: The hazard function for APT given x.

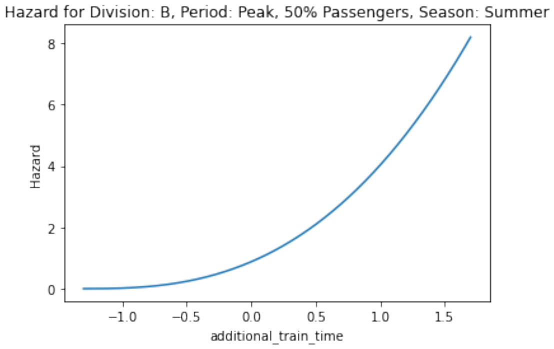


Figure 4: The hazard function for ATT given x.

We can also find the likelihood of a subway arriving at time t given that it has not arrived before via the same hazard function in section 2.2. Figure 4 contains the resultant plot. We find once more that as more time passes, the likelihood of the subway arriving increases, which makes intuitive sense.

3 Optimization

As a decision-maker, arguably one of the most substantial factors in considering any important decision is the revenue of said decision. Thus, in the context of this study, it is pertinent to observe the relationship between revenue (R) and APT and ATT.

More formally, define $R = 2.75P$, where P is the number of passengers (2.75 represents the price of one ticket). We want to find the maximum revenue, or in other words,

$$\max(R) = 2.75 \max(P).$$

Thus in order to maximize the revenue, we need to maximize the number of passengers. We assume that if the additional wait time t is too long (where $t = \text{apt} + \text{att}$), the number of passengers will decreases since consumers will be more inclined to choose an alternative form of transportation that is faster. We can model this by a loss rate $L(t) = \frac{1}{1+e^{-t}}$. For example, if a subway takes $t = 3$ additional minutes to come by, we expect $L(3) \times P$ passengers to leave.

This means we want to find P, which is defined as the following:

$$P = \arg\max_{p} \sum_{S \in \mathbb{S}(C \backslash \text{num\_passengers})} \mathbb{P}(S) \int_{t=0}^{\infty} (1 - L(t)) \cdot p \cdot f(t \mid S \cap p) \partial t,$$

where $f(t \mid S \cap p)$ is the distribution of total wait time given a specific set of covariates S and a specific number of passengers p, $C\backslash$num_passengers is the set of all covariates minus num_passengers, and $\mathbb{S}(\cdot)$ is the permutation of all *possible* values the input can take. For instance, one such permutation would be

$$\{\text{division, period, spring, summer, autumn}\} \rightarrow \{0, 1, 0, 1, 0\},$$

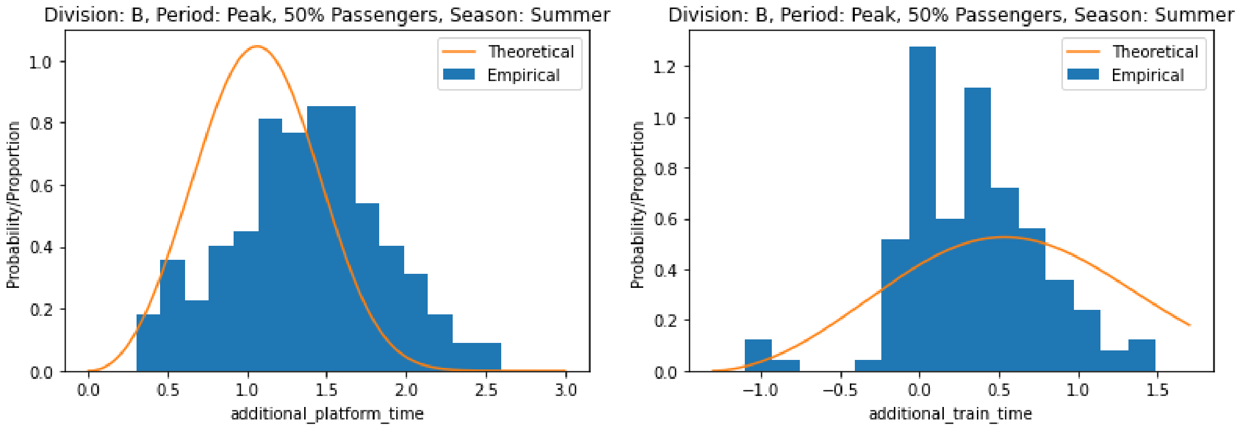
which would represent division A during a peak period in the summer. Note that spring, summer, and autumn are mutually exclusive: if spring is 1, summer and autumn must be 0.

The intuition for this equation comes from iterating through all potential wait times for given sets of covariates to get the expected number of customers. To find $f(t \mid S \cap p)$, we assume APT and ATT to be independent, implying $f_{\text{APT, ATT}}(apt, att \mid S \cap p) = f_{\text{APT}}(apt \mid S \cap p) f_{\text{ATT}}(att \mid S \cap p)$. Thus we find that

$$f(t \mid S \cap p) = \int_{0}^{t} f_{\text{APT}}(apt \mid S \cap p) f_{\text{ATT}}(t - apt \mid S \cap p) \partial apt.$$

4 Limitations

One of the major limitations of our study is that our Weibull regression models for APT and ATT rarely converge to the same values. Because of this, we sometimes encounter "bad" estimates for c and λ, resulting in "bad" fits to the data. This is shown in Figures 5a and 5b. More generally, this inability to converge to specific values implies our models are unstable for the given data and covariates.



(a) APT.

(b) ATT.

Figure 5: "Bad" distribution estimates.

We had also considered using the MTA Subway Mean Distance Between Failures: Beginning 2015 dataset in order to perform some type of optimization in resource allocation for subway failures. However, this would require having data on the current resource allocation for the MTA, which we could not find. Refer to Appendix A.2 for more information on modeling mean distance between failures.

While the hazard functions for APT and ATT provided in sections 3.2 and 3.3 correspond to what we intuitively expect (that as additional wait time increases, the likelihood of observing the subway arriving increases), the functions do not give actual *likelihoods*. We can thus only see the *relationship* between additional wait time and observing the subway arrival, not the actual underlying probabilities.

Another limitation is the assumption of independence between APT and ATT in our optimization model even though there is a significant correlation of 0.331. For future studies, we may refer to Yacoub et al. 2005 where the authors derive a closed-form solution for the probability density of two correlated Weibull distributions to increase the accuracy of the optimization model by accounting for correlation. Furthermore, the loss function can be better modeled as the current loss function most likely overestimates the number of passengers lost when faced with a certain wait time. Moreover, since the optimization requires numerical integration methods and maximization techniques, the optimization requires large computational power.

5 Conclusion

Using our models for APT and ATT and the recovered coefficients for the covariates, we can predict distributions APT and ATT for a given set of covariates. For instance, if a passenger is riding a subway in Division B, during a peak period, when there is the median number of passengers, in the summer, we can find the expected wait time for their trip by adding the expected APT and ATT from the predicted distributions using our models. The MTA may find this information useful in updating the actual wait time by increasing it by the expected predicted wait time, giving subway passengers a more accurate measure of their wait.

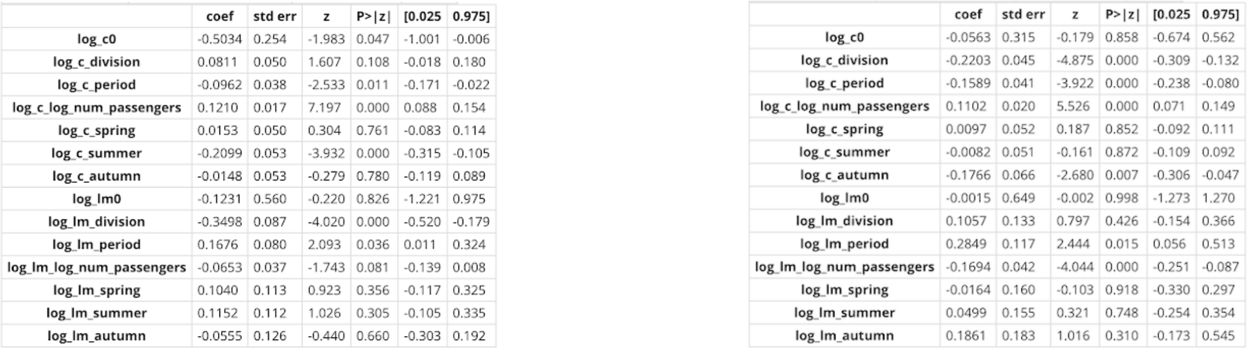
Similarly, since the mean of a Weibull distribution is increasing in c and decreasing in λ, we can interpret the recovered coefficients from our regressions keeping all else equal. Thus, based on our model, the expected APT is longer for division B than it is for division A while the expected ATT is longer for division A than it is for division B. This suggests that subways in division A (lines 1, 2, 3, 4, 5, 6, and 7) may have longer ATT but shorter APT while subways in division B (lines A, B, C, D, E, F, G, J, L, M, N, Q, R, W, and Z) may have shorter ATT but longer APT. Moreover, the expected APT and expected ATT for peak periods are longer than that of off-peak periods, which makes sense intuitively. For a larger number of passengers, both expected APT and ATT are longer, which aligns with our intuition as well. Lastly, for seasonal coefficients with a p-value less than 0.10, during the summer, the expected APT and ATT are shorter than in other seasons. Thus, we are able to quantify and interpret the expected additional wait times for passengers.

While we were unable to obtain conclusive results for optimized revenue as the program took too much time to run, we were able to implement methods to estimate the integrals and find the maximizing number of passengers. With more efficient algorithms and better computational power, the MTA can use this information to increase its revenue.

Overall, we hope that the findings in this paper can help not only the MTA but also subway riders and the surrounding community of New York.

Appendix A

A.1 Figures 6a and 6b contain the raw model outputs for APT in section 2.2 and ATT in section 2.3, respectively.



	coef	std err	z	P>|z|	[0.025	0.975]
log_c0	-0.5034	0.254	-1.983	0.047	-1.001	-0.006
log_c_division	0.0811	0.050	1.607	0.108	-0.018	0.180
log_c_period	-0.0962	0.038	-2.533	0.011	-0.171	-0.022
log_c_log_num_passengers	0.1210	0.017	7.197	0.000	0.088	0.154
log_c_spring	0.0153	0.050	0.304	0.761	-0.083	0.114
log_c_summer	-0.2099	0.053	-3.932	0.000	-0.315	-0.105
log_c_autumn	-0.0148	0.053	-0.279	0.780	-0.119	0.089
log_lm0	-0.1231	0.560	-0.220	0.826	-1.221	0.975
log_lm_division	-0.3498	0.087	-4.020	0.000	-0.520	-0.179
log_lm_period	0.1676	0.080	2.093	0.036	0.011	0.324
log_lm_log_num_passengers	-0.0653	0.037	-1.743	0.081	-0.139	0.008
log_lm_spring	0.1040	0.113	0.923	0.356	-0.117	0.325
log_lm_summer	0.1152	0.112	1.026	0.305	-0.105	0.335
log_lm_autumn	-0.0555	0.126	-0.440	0.660	-0.303	0.192

(a) APT.

	coef	std err	z	P>|z|	[0.025	0.975]
log_c0	-0.0563	0.315	-0.179	0.858	-0.674	0.562
log_c_division	-0.2203	0.045	-4.875	0.000	-0.309	-0.132
log_c_period	-0.1589	0.041	-3.922	0.000	-0.238	-0.080
log_c_log_num_passengers	0.1102	0.020	5.526	0.000	0.071	0.149
log_c_spring	0.0097	0.052	0.187	0.852	-0.092	0.111
log_c_summer	-0.0082	0.051	-0.161	0.872	-0.109	0.092
log_c_autumn	-0.1766	0.066	-2.680	0.007	-0.306	-0.047
log_lm0	-0.0015	0.649	-0.002	0.998	-1.273	1.270
log_lm_division	0.1057	0.133	0.797	0.426	-0.154	0.366
log_lm_period	0.2849	0.117	2.444	0.015	0.056	0.513
log_lm_log_num_passengers	-0.1694	0.042	-4.044	0.000	-0.251	-0.087
log_lm_spring	-0.0164	0.160	-0.103	0.918	-0.330	0.297
log_lm_summer	0.0499	0.155	0.321	0.748	-0.254	0.354
log_lm_autumn	0.1861	0.183	1.016	0.310	-0.173	0.545

(b) ATT.

Figure 6: Raw model outputs.

A.2 We assume mean distance between failures (MDBF) to follow a Negative Binomial distribution as the values are positive and discrete. In similar fashion to regression on APT and ATT, we assume MDBF $\sim NBD\left(\gamma, \frac{\alpha}{\exp(X\beta)}\right)$, where X here is the same design matrix as defined in section 2.1 except without the ones vector as the first column (thus $X \in \mathbb{R}^{n\times d}$, where before $X \in \mathbb{R}^{n\times(d+1)}$). Below are the estimated parameter values:

γ	α	β_1	β_2	β_3	β_4	β_5	β_6
151.14	0.0008	-0.4522	0.0010	-0.0203	0.0021	-0.0030	-0.0239

Suppose we wanted to find the distribution of MDBF for people in division A during a peak period during the winter when there are the median number of passengers. Then we have

$$x = \begin{bmatrix} 0 & 1 & \log M & 0 & 0 & 0 \end{bmatrix},$$

where M is the median number of passengers and we would have MDBF $\sim NBD\left(\gamma, \frac{\alpha}{\exp(x\cdot\beta)}\right)$. Figure 7 contains the resultant plot overlayed on a histogram of the empirical data.

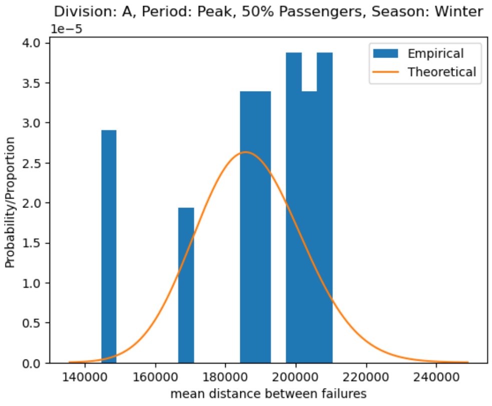


Figure 7: The resultant plot for the distribution of MDBF given x overlayed on the empirical data.