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

By Grace He, Chris Lee

Background

One of MTA's main forms of transportation is the **subway**. However there are some **problems** with the current system:

1. Inaccurate information about wait times
2. Delays due to subway failure

Solution:

1. Model **additional wait times** in terms in terms of available covariates
2. Model **failures** in terms available covariates

For the decision-maker, we can see how the additional wait times passengers experience translate to revenue gain or loss.

Description of Data

1. MTA Subway Customer Journey Focused Metrics: Beginning 2015
    - month
    - division
    - period
    - num_passengers
    - additional_platform_time
    - additional_train_time
2. MTA Subway Mean Distance Between Failures: Beginning 2015
    - month
    - car_class
    - mdbf

Discussion of Models

Three Models:

1. Weibull Regression on additional_platform_time (APT)
2. Weibull Regression on additiona_train_time (ATT)
3. Negative Binomial Regression on mean distance between failures

Covariates: Division, Period, Number of Passengers, Season

Results

	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.

Results

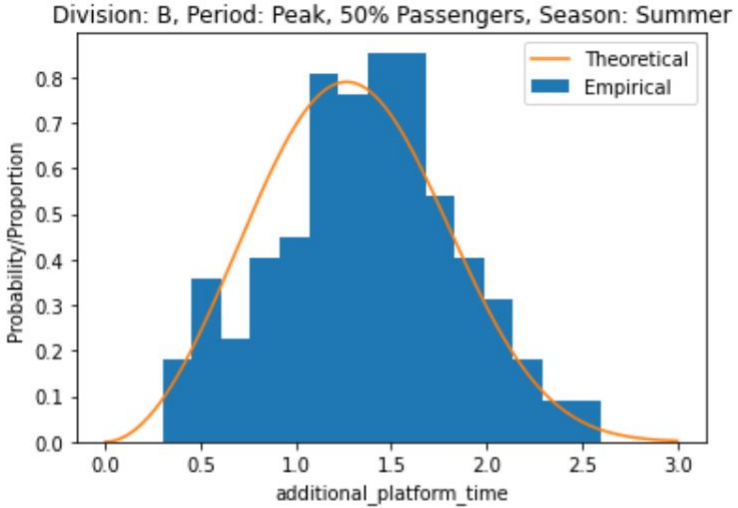


Figure 1: The resultant plot for the distribution of APT given x overlayed on a histogram of the empirical data.

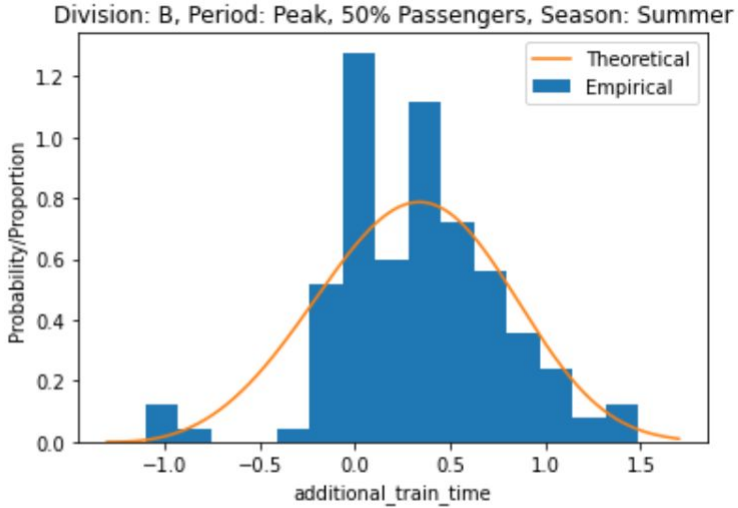


Figure 2: The resultant plot for the distribution of ATT given x overlayed on a histogram of the empirical data.

Results

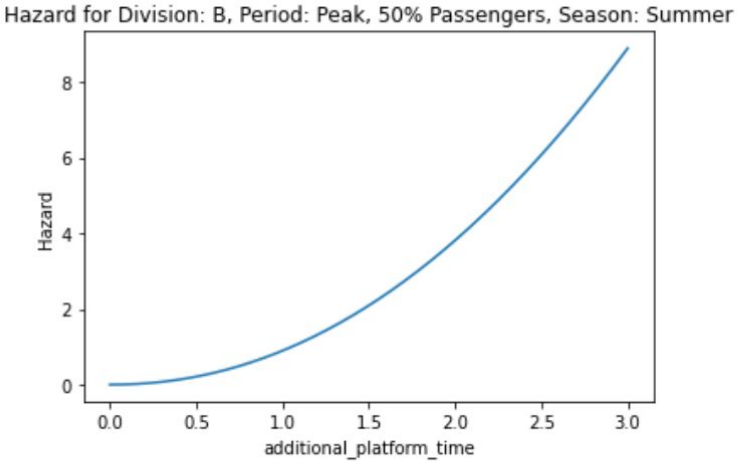


Figure 3: The hazard function for APT given x.

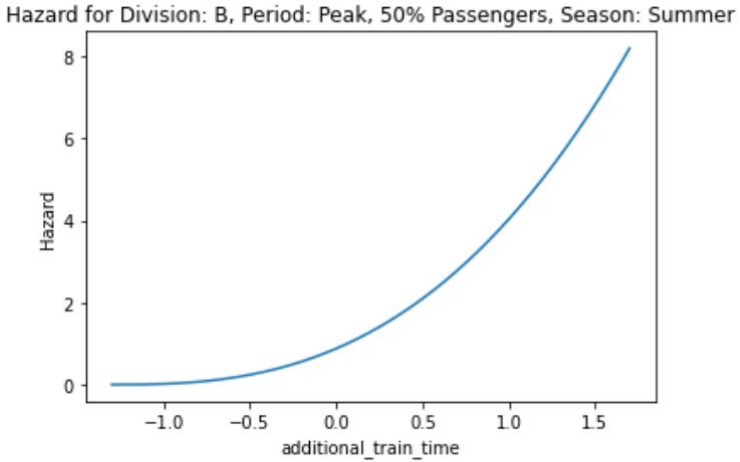


Figure 4: The hazard function for ATT given x.

Optimization

Problem: **Maximizing MTA's revenue**

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

which is equivalent to **maximizing number of passengers.**

Let's assume that when the additional wait time is too long, a proportion of passengers will decide not to take the subway in favor of another form of transportation:

$$L(t) = \frac{1}{1+e^{-t}}$$

Optimization

Thus, we need to find the number of passengers that maximizes the expected number of actual customers.

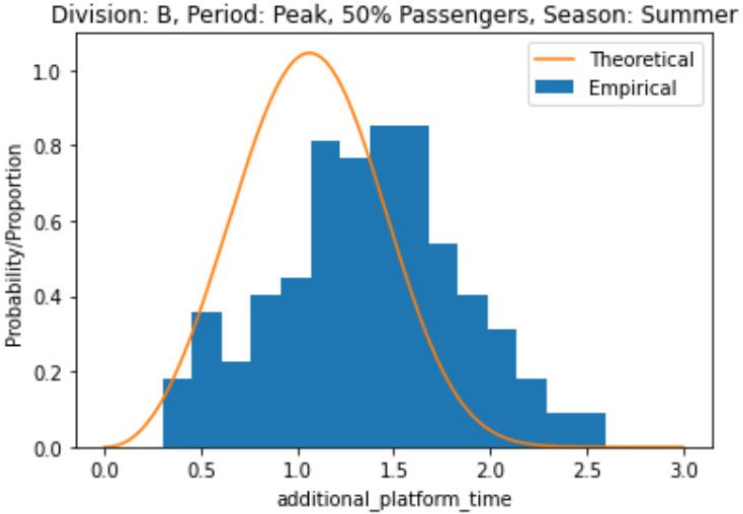
$$P = \arg\max_{p} \sum_{S \in \mathbb{S}(C \setminus \text{num\_passengers})} \mathbb{P}(S) \int_{t=0}^{\infty} (1 - L(t)) \cdot p \cdot f(t \mid S \cap p) \partial t,$$

We assume independence between the two Weibull distributions of APT and ATT. Using convolution, we get:

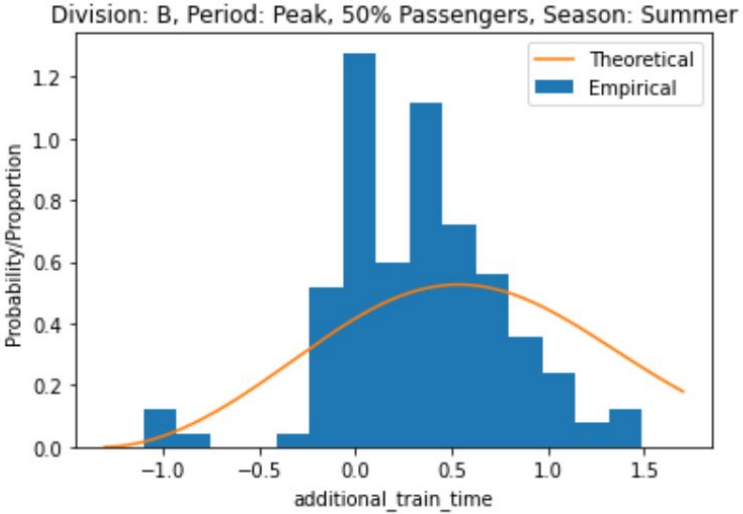
$$f(t \mid S \cap p) = \int_{0}^{t} f_{\text{APT}}(apt \mid S \cap p) f_{\text{ATT}}(t - apt \mid S \cap p) \partial apt.$$

Limitations

- Unstable models



(a) APT.



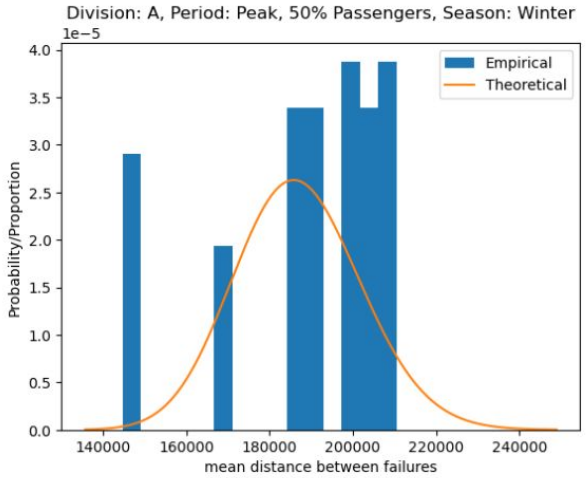
(b) ATT.

Figure 5: "Bad" distribution estimates.

Limitations

Mean Distance Between Failures (MDBF)

- Insufficient data for MDBF optimization
- Assume MDBF follows NBD



	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	12.4356	0.026	477.648	0.000	12.385	12.487
division	-0.4522	0.005	-100.072	0.000	-0.461	-0.443
period	0.0010	0.004	0.251	0.802	-0.007	0.009
log_num_passengers	-0.0203	0.002	-11.668	0.000	-0.024	-0.017
spring	0.0021	0.006	0.356	0.722	-0.010	0.014
summer	-0.0030	0.006	-0.511	0.609	-0.014	0.008
autumn	-0.0239	0.007	-3.632	0.000	-0.037	-0.011
alpha	0.0066	0.000	27.660	0.000	0.006	0.007

Limitations

Optimization:

1. Independence between Weibull distributions
   a. Correlation of 0.311
2. Loss function
   a. Overestimates how many passengers would leave
3. Computational power
   a. Double numerical integration approximation
   b. Maximization

Conclusion

Using our Weibull regressions, we can now:

- Quantify and interpret expected wait times for passengers
- Give passengers more accurate representation of waiting times
- Derive theoretical equations for optimal revenue

For future work, we aim to:

- Find a method to find consistent regressions on APT and ATT
- Include correlation in our optimization
- Better model the loss function
- Increase the efficiency of our optimization via better algorithms and/or higher computational power