INFO2950 Phase III

Andrew Chung, hc893

November 5, 2025

Preregistraton Statement I

Hypothesis

We hypothesize that ridership recovery rates have increased over time (2020-2024) since the COVID-19 pandemic, but that the recovery trajectories for commuter rail services (LIRR, MNR) are significantly more pronounced than for urban transit agencies (NYC Subway, Buses). We also hypothesize that there are significant, persistent differences between weekday and weekend recovery patterns.

Analysis

I plan to transform the current data set from wide to long format, forming $1776 \times 4 = 7104$ rows for each agency-`DateTime` combination that contain the following:

- the response variable (`recovery_rate`)
- dummy binary variables representing the type of transit agency and year since 2020
- an indicator variable `is_weekend` (The reference values will be NYC Subway and the year 2020, respectively.)
- Interaction terms between transit agency and year (e.g., LIRR * 2022) to test the differing trajectories.

I will train a multiple linear regression model. I plan to account for the chronological nature of my data by checking the Durbin-Watson statistic for autocorrelation, and using robust standard errors if significant (e.g., if the D-W statistic is < 1.5 or > 2.5). My hypothesis tests will check for the significance of the following terms:

- Year Effect: $H_0 : \beta_y = 0, H_A : \beta_y > 0; y \in \{2021, \cdots, 2024\}$
- Weekend Effect: $H_0 : \beta_{\text{weekend}} = 0, H_A : \beta_{\text{weekend}} \neq 0$
- $H_0 : \forall(k, j)\, \gamma_{kj} = 0, H_A :$ at least one $\gamma_{kj} \neq 0$, where k denotes year and j denotes transit agency
  - This F-test examines whether temporal recovery patterns differ significantly from the subway across non-reference transit agencies overall.
  - For specific insights on individual agency-level differences from the subway, I will examine the signs and significance of individual $\gamma_k j$ terms.

To make direct comparisons between non-reference groups, I plan on 3 post-hoc linear contrast F-tests (LIRR vs MNR, LIRR vs. Bus, MNR vs. Bus). We will apply a Bonferroni correction for this family of 3 tests.

Preregistraton Statement II

Hypothesis

The relationship between weekend status and ridership recovery has evolved differently over time for commuter rail services (LIRR/MNR) compared to urban transit agencies (NYC Subway/Buses). We hypothesize that in recent years (2023-2024), this difference has become more pronounced, indicating a structural shift in use patterns of commuter vs. urban transit agencies on weekends versus weekdays.

Analysis

Similar to above, I will dummy-code years and weekend status into indicator variables; here, however, instead of encoding each transit agency I plan to construct a simple binary variable indicating commuter status (LIRR/MNR) vs. non-commuter (Subways/Buses). (This grouping is justified by our Phase II EDA, which showed a +0.93 correlation between LIRR and MNR recovery rates, suggesting they behave as a distinct 'commuter' block.)

I plan to build interaction terms between `is_commuter * is_weekend`, `is_commuter * year`, `is_weekend * year`, and `is_commuter * is_weekend * year` (hierarchy principle), then train a multiple linear regression model with `recovery_rate` as the response. In the same fashion as above, I plan to account for the chronological nature of my data by checking the Durbin-Watson statistic for autocorrelation, and using robust standard errors if significant. My hypothesis tests will check for the significance of the following terms (W = weekend, C = commuter):

- $H_0 : \delta_{WCk} = 0; H_A : \delta_{WCk} \neq 0$, where $k \in \{2021, \cdots, 2024\}$
  - A significant coefficient for a certain year would imply that the difference between weekend and weekday recovery for commuter rail changed in that year relative to the 2020 baseline.
  - The overall F-test will examine whether the 3-way interaction term is significant at all.
- I additionally plan to compare the averages of the δ_{WCk} coefficients between 2021-2022 and 2023-2024 to address the temporal aspect of my hypothesis test.

Potential Model Diagnostics

Before final interpretations for either model, we will perform the following diagnostic checks to ensure our model assumptions are fulfilled:

- Plot residuals vs. fitted values to check for non-linear patterns or heteroskedasticity (fanning-out shapes)
- Use Q-Q plot to check for normally distributed residuals
- Calculate the correlation matrix to check for multicollinearity
- Use Cook's distance (d) to identify and study any leverage/high influence points

Questions for Reviewers

1. I proposed dummy coding for year per your feedback. Would you also recommend exploring a continuous variable through polynomial regression to better capture non-linear curves of recovery, or would the annual indicator variables suffice?
2. For the second hypothesis we decided to group MTA transit agencies into urban (NYC subway, buses) and commuter (LIRR, MNR), per the EDA which showed high correlation (+0.93). Do we lose explanatory power with this grouping?
3. I plan to adopt the Durbin-Watson statistic for autocorrelation and Newey-West corrected standard errors for robustness. Would this be sufficient to account for the time series nature of my data or should I consider a full ARIMA model in place of traditional regression? Should we also consider robust standard errors beyond Newey-West in event of heteroskedascity?
4. Our models have complex interactions (3-way). For our final notebook how could I effectively interpret and communicate these? Is a standard coefficient table sufficient or would a plot of predicted values be useful?

Appendix

Below are the full mathematical regression models for hypotheses I and II.

Hypothesis I

$$Y_i = \beta_0 + \beta_w x_{w,i} + \sum_j \beta_j x_{j,i} + \sum_k \beta_k x_{k,i} + \sum_j \sum_k \gamma_{jk}(x_{j,i} \times x_{k,i}) + \varepsilon_i$$

where w denotes the `is_weekend` dummy variable, $j \in \{2021, 2022, 2023, 2024\}$ denotes year, $k \in \{\text{Buses}, \text{LIRR}, \text{MNR}\}$ denotes non-subway transit agency. $\mathbf{x}_w, \mathbf{x}_j, \mathbf{x}_k$ are indicator variables for weekend status, year and transit agency.

Hypothesis II

$$Y_i = \beta_0 + \beta_w x_{w,i} + \beta_c x_{c,i} + \sum_k \beta_k x_{k,i} + \gamma_{wc}(x_{w,i} \times x_{c,i}) + \sum_k \gamma_{wk}(x_{w,i} \times x_{k,i}) + \sum_k \gamma_{ck}(x_{c,i} \times x_{k,i}) + \sum_k \delta_{wck}(x_{w,i} \times x_{c,i} \times x_{k,i}) + \varepsilon_i$$

where w denotes the `is_weekend` dummy variable, c denotes the `is_commuter` dummy variable, and $k \in \{2021, 2022, 2023, 2024\}$ denotes year. $\mathbf{x}_w, \mathbf{x}_c, \mathbf{x}_k$ are indicator variables for weekend status, commuter status, and year.