Predicting Urban Micromobility Trip Distance Using Linear Regression and Penalized Modeling

Bryant Willoughby
University of Michigan, Ann Arbor

December 2025

1 Introduction

Shared micromobility systems—including docked and dockless bicycles, e-bikes, and electric scooters—have become an integral component of modern urban transportation networks. These services provide flexible, low-cost travel and complement public transit by enabling efficient first-mile and last-mile connections. Their widespread adoption has encouraged cities to use increasingly data-driven approaches for infrastructure planning, fleet management, and regulatory design. Because of this, understanding how far users travel, and how trip distance varies across temporal, spatial, and operational contexts, holds practical importance for both municipal agencies and private operators.

Trip distance is a central behavioral metric in micromobility research. It influences infrastructure decisions (e.g., protected lanes, charging stations), multimodal planning, and assessments of neighborhood-level mobility behavior. Yet trip distance is shaped by complex interactions among temporal patterns, spatial heterogeneity, and user choices, many of which are moderated by categorical factors such as vehicle type, time of day, and council district. Classical linear regression often struggles to capture such complexity, motivating a careful investigation of more flexible modeling approaches within a linear framework.

This study aims to predict the *log-transformed trip distance* of micromobility trips using predictors derived from a large real-world dataset. The modeling strategy proceeds in two stages. First, a sequence of increasingly flexible regression models is developed, culminating in a hierarchical interaction specification (denoted *M7*) that includes quadratic terms and interpretable domain-motivated interactions. Although M7 adheres to principled hierarchical modeling and offers substantive interpretability, diagnostic analysis reveals persistent nonlinear patterns that it cannot capture.

The second stage expands the feature space to include all pairwise interactions and quadratic terms, yielding a high-dimensional linear representation. Penalized regression methods—ridge, lasso, and elastic net—are applied to this expanded design matrix to perform shrinkage and automatic variable selection beyond what is feasible in manually constructed hierarchical models. Additional techniques such as weighted least squares (WLS), robust regression, and influence-point removal are explored but prove ineffective when the underlying mean function remains misspecified.

Across both stages, results consistently indicate structural nonlinearities and complex interaction effects that no linear model adequately represents. While penalized approaches provide modest predictive gains relative to M7, systematic residual curvature remains, highlighting both the limitations of linear modeling for this task and the intrinsic complexity of micromobility trip-distance behavior.

2 Data Description

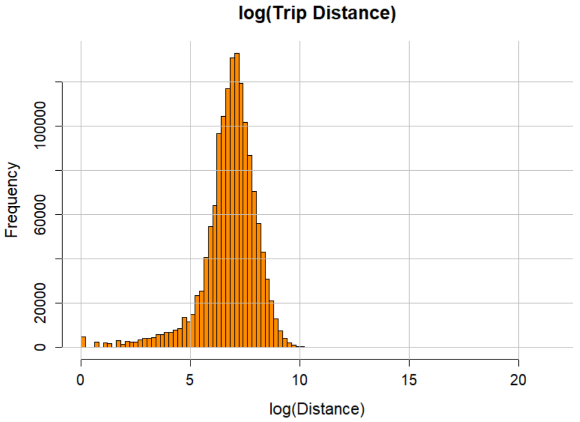
The dataset used in this study is sourced from the City of Austin's Open Data Portal City of Austin (2022), which provides administrative records of micromobility trips collected from dockless bicycle, e-bike, and e-scooter providers operating within the city. The full dataset spans April 2018 through April 2022 and contains more than fifteen million observations. Each entry corresponds to a single trip. Because both user behavior and provider operations vary substantially across years—particularly during the COVID-19 pandemic—the present analysis restricts attention to weekday trips recorded in calendar year 2018. This restriction yields a sample of approximately 1.67 million trips and avoids structural breaks that could otherwise obscure modeling insights.

To prepare the raw dataset for statistical modeling, a sequence of preprocessing steps was undertaken. Administrative identifiers, duplicated timestamp fields, and auxiliary metadata were removed to reduce unnecessary dimensionality. Trip distance and duration occasionally contained non-numeric characters, including units or stray punctuation; these fields were cleaned using regular-expression parsing to extract valid numeric components before conversion to floating-point values. Observations with non-positive distance or duration were discarded, as these typically arise from incomplete device logging or reporting errors. Start and end timestamps were converted to Unix time to provide a consistent numerical representation of temporal information. Missing council district values were rare and were imputed using the empirical mode for each district, a sensible choice given the categorical and low-missingness structure of these variables.

To balance interpretability with representational richness, several categorical variables were recoded into more analytically meaningful forms. The vehicle-type variable was simplified to a bicycle versus scooter indicator. The month of year was collapsed into three seasonal categories—

Spring/Early Summer, Summer, and Fall/Winter—and the hour of day was grouped into four time-of-day periods capturing overnight, morning, afternoon, and evening travel patterns. The day-of-week variable was restricted to weekdays and encoded as an ordered factor (Monday–Friday). These engineered predictors preserve essential behavioral structure without introducing excessive categorical complexity.

Because the raw distribution of trip distance is highly right-skewed, the modeling response variable is defined as the natural logarithm of distance. This log transformation substantially reduces skewness and stabilizes the conditional variance, yielding a response more compatible with linear modeling assumptions. Figure 1 (left panel) presents a histogram of log-transformed trip distance, illustrating the improved symmetry achieved through transformation. Table 1b (right panel) summarizes the final set of variables used throughout the modeling process.



(a) Distribution of log-transformed trip distance for weekday trips in 2018.

Variable	**Description**
Log Trip Distance	Response variable (log miles)
Trip Duration	Duration in seconds
start_ts	Unix start time
end_ts	Unix end time
councilstart	Start council district (imputed)
councilend	End council district (imputed)
vehicle_type	Bicycle vs. scooter
day_of_week	Weekday factor (Mon–Fri)
season	Three-level seasonal category
time_of_day	Four-level time-of-day period

(b) Summary of analyzed variables.

Figure 1: Summary of response distribution (left) and predictor metadata (right).

3 Methods

The overarching objective of this analysis is to model the conditional mean of the log-transformed trip distance

$$Y = \log(\text{Trip Distance})$$

using a structured set of temporal, spatial, and operational predictors. Let X_i denote the design (predictor) matrix for trip i and Y_i the corresponding log-distance. All models considered in this study estimate a parametric function $f(X_i;\theta)$ under the squared-error loss

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \left(Y_i - f(X_i;\theta)\right)^2 ,$$

with model performance evaluated on an independent test set.

The methodological workflow proceeds in three stages: (1) exploratory modeling using main-effects and progressively enriched linear specifications, (2) construction and assessment of a hierarchical interaction model (M7), and (3) high-dimensional penalized regression using an expanded feature space.

All analysis was conducted in R using the `tidyverse`, `MASS`, and `glmnet` packages. The cleaned dataset was partitioned into a 70% training subset and 30% test subset prior to model construction.

3.1 Exploratory Modeling and Preliminary Linear Specifications

Initial exploratory modeling examined the association between log-distance and core predictors including trip duration, time of day, weekday indicators, vehicle type, and council districts. Several main-effects linear models were fit to evaluate baseline explanatory power, inspect residual patterns, and determine whether linearity or homoscedasticity assumptions were violated.

The exploratory analysis revealed systematic curvature in both trip duration and start timestamp, motivating the inclusion of quadratic terms. Interactions among temporal and operational predictors appeared necessary to capture heterogeneous behavioral regimes (e.g., vehicle-type–specific duration effects; seasonally varying temporal patterns).

3.2 Hierarchical Model Construction and the M7 Specification

Based on exploratory diagnostics and theoretical considerations, a hierarchically structured linear model (denoted M7) was constructed to incorporate:

- main effects for all operational, temporal, and spatial variables,
- quadratic terms for standardized trip duration and start timestamp,
- selected two-way interactions adhering to marginality constraints (i.e., inclusion of interaction terms only when the associated main effects were present).

The resulting specification can be written as

$$\begin{aligned}Y_i = \beta_0 + \beta_1\,\text{tripduration}_{z,i} + \beta_2\,\text{start\_ts}_{z,i} + \beta_3\,\text{tripduration}^2_{z,i} + \beta_4\,\text{start\_ts}^2_{z,i}\\ + \text{councilstart}_i + \text{councilend}_i + \text{vehicle\_type}_i + \text{day\_of\_week}_i + \text{season}_i + \text{time\_of\_day}_i\\ + \text{selected two-way interactions between}\{\text{duration}, \text{timestamp}, \text{season}, \text{vehicle type}, \text{time of day}\}.\end{aligned}$$

This model served as the primary interpretable benchmark for all subsequent comparisons.

3.3 Diagnostic Assessment of M7

After fitting M7 by ordinary least squares, model assumptions were evaluated using standardized residuals, leverage values, externally studentized residuals, and Cook's distances. For an observation i, let X_i denote the corresponding row of the design matrix, e_i the ordinary residual, p the number of fitted parameters, and $\hat{\sigma}^2$ the estimated error variance. The primary diagnostics are defined as:

$$h_i = X_i^\top (X^\top X)^{-1} X_i, \qquad t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}, \qquad D_i = \frac{e_i^2}{p\,\hat{\sigma}^2}\,\frac{h_i}{(1-h_i)^2}.$$

Here, h_i measures the *leverage* of observation i, quantifying its influence in determining the fitted values; t_i is the *externally studentized residual*, which adjusts the raw residual by the observation's leverage; and D_i is *Cook's distance*, a joint measure of residual magnitude and leverage that approximates each observation's influence on the full vector of fitted coefficients.

To identify potentially influential observations, standard heuristic thresholds were applied:

$$h_i > \frac{2p}{n}, \qquad |t_i| > 3, \qquad D_i > \frac{4}{n}.$$

Observations exceeding *all three* thresholds were flagged. These cases were removed temporarily, and the M7 model was refit using the identical specification to assess whether reducing extreme leverage and outlier effects materially improved predictive performance.

3.4 Penalized Regression on an Expanded Feature Space

To probe whether a richer linear structure could better approximate $\mathbb{E}[Y \mid X]$, the predictor space was expanded to include all main effects used in M7, all pairwise interactions among these variables, and quadratic terms for the two continuous predictors. Formally, the augmented design matrix is

$$\mathcal{X} = \Big\{\text{main effects},\ X_j X_k \text{ for all } j < k,\ X_j^2 \text{ for continuous } X_j\Big\},$$

resulting in a high-dimensional representation containing 430 columns after encoding categorical factors. Because the dimensionality of $\mathcal{X}$ is large relative to the underlying signal, and categorical interactions produce substantial sparsity, the full design matrix was constructed using `sparse.model.matrix` to substantially reduce memory overhead.

Within this expanded linear space, three penalized regression estimators were considered, each defined as the minimizer of a regularized empirical risk under the same squared-error loss introduced at the start of the Methods section. Let $\beta \in \mathbb{R}^p$ denote the coefficient vector for the expanded feature set and write the penalized objective generically as

$$\hat{\beta}_{\lambda,\alpha} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(Y_i - X_i^\top \beta\right)^2 \; + \; \lambda \, P_\alpha(\beta) \right\},$$

where $\lambda > 0$ controls the overall penalty strength and α determines the form of the penalty. Ridge regression corresponds to the purely quadratic penalty

$$P_{\alpha=0}(\beta) = \|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2,$$

which shrinks coefficients continuously toward zero while retaining all predictors. Lasso regression replaces this with an ℓ_1 penalty

$$P_{\alpha=1}(\beta) = \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|,$$

which induces sparsity by shrinking many coefficients exactly to zero, thereby performing automatic variable selection. Elastic net regression combines these two regularizers,

$$P_\alpha(\beta) = (1-\alpha)\|\beta\|_2^2 \; + \; \alpha\|\beta\|_1, \qquad 0 < \alpha < 1,$$

and provides a compromise between ridge's stability and lasso's sparsity. In this study, the elastic net was evaluated using $\alpha = 0.5$ to balance these effects.

All three estimators were fit using `glmnet` with three-fold cross-validation to select the tuning parameter λ. To reduce computational burden while still exploring a meaningful range of values, the penalty path consisted of 50 candidate λ values with a lower bound defined by

$$\lambda_{\min} = \lambda_{\max} \cdot (\text{lambda.min.ratio}), \qquad \text{lambda.min.ratio} = 0.01,$$

where $\lambda_{\max}$ denotes the smallest value at which the lasso solution is the zero vector. For each estimator, two tuning parameters were retained for evaluation: the minimum-MSE value $\lambda_{\min}$ and the more parsimonious one-standard-error value $\lambda_{1\text{se}}$.

Because the dataset contains more than one million observations and a large interaction-expanded feature matrix, all penalized models were fit using sparse matrix operations and parallel computation. A compute cluster was registered using `doParallel` and `makeCluster`, enabling cross-validation folds to be evaluated concurrently. This combination of sparsity, parallelization, and reduced λ-grid resolution ensured that the penalized estimators were computationally feasible on a local CPU environment while still providing rigorous regularization paths for comparison.

4 Results

4.1 Model Fit and Residual Structure for the Hierarchical Model (M7)

The M7 specification represents the most expressive interpretable linear model constructed in this study, incorporating all relevant main effects, curvature in the two continuous predictors, and several theoretically justified two-way interactions. Despite this enriched structure, the residual-versus-fitted plot in Figure 2 reveals clear departures from linearity: residuals exhibit strong curvature, funnel-shaped heteroscedasticity, and systematic under- and over-prediction across different regions of the fitted-value range. These patterns indicate that the underlying mean function governing trip distance is not well approximated by any linear model of this form.

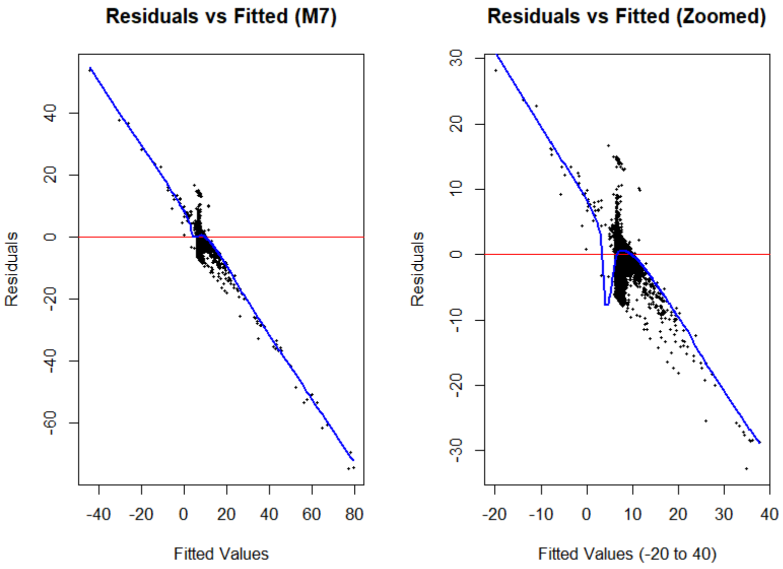


Figure 2: Residuals versus fitted values for the M7 model. Clear curvature and heteroscedasticity indicate structural misspecification.

Influence diagnostics (leverage, externally studentized residuals, and Cook's distance) identified 2,156 observations exceeding all three standard thresholds. Although these cases represent fewer than 0.2% of the training data, their removal provided an opportunity to assess whether extreme leverage or outliers were driving the observed residual pathology. After excluding these observations and refitting M7, the resulting residual plot (Figure 3) exhibited the same qualitative curvature and variance patterns as before. Thus, the misspecification arises from the functional form rather than the presence of atypical observations.

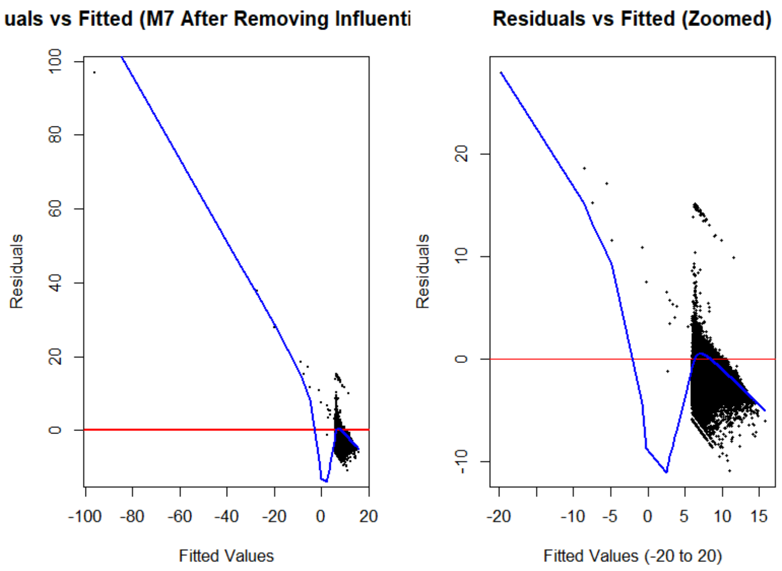


Figure 3: Residuals versus fitted values for M7 after removing influential observations. Structural curvature persists, confirming model misspecification.

Test-set RMSE results (1) reinforce this conclusion. Removing influential observations dramatically worsened predictive accuracy, suggesting that these points contain meaningful signal rather than noise. These findings establish that even carefully constructed hierarchical linear models cannot adequately represent the nonlinear relationships present in micromobility trip distance.

Model	RMSE
Original M7	1.290281
Cleaned M7	24.876217

Table 1: Test-set RMSE comparison for M7 before and after removing influential observations.

4.2 Exploratory Variance Modeling Attempts

Two additional modeling approaches were briefly explored to assess whether failures of the M7 specification arose from variance heterogeneity or heavy-tailed errors rather than from structural misspecification of the mean. First, a weighted least squares (WLS) model was fit using weights derived from a simple variance–mean relationship estimated from the absolute residuals of M7. Because the fitted values in M7 already reflect an incorrect mean function, the estimated weights inherit this misspecification, leading to distorted inference and an even more pronounced residual–fitted pattern. This reinforces that variance modeling cannot correct an incorrectly specified mean.

Second, motivated by the heavy-tailed behavior seen in QQ-plots of standardized residuals, a Huber M-estimator was fit as a form of robust regression. However, the algorithm failed to converge and produced unstable coefficient estimates. Robust procedures are designed to down-weight isolated outliers, not to compensate for systematic structural misspecification or the leverage imbalance created by large, uneven categorical designs. As with WLS, the robust regression attempt confirmed that the primary limitation lies in the mean function itself, not the error distribution.

Because neither method improved predictive performance or diagnostic behavior, neither is carried forward. Their inclusion here serves only to document that variance-focused remedies were considered but found inappropriate for this problem.

4.3 Penalized Regression Models on the Expanded Feature Space

Penalized regression methods were applied to an expanded design matrix containing all main effects, all two-way interactions, and quadratic terms. This produced 430 predictors encoded in a sparse matrix. Ridge, lasso, and elastic net were each tuned using three-fold cross-validation.

4.3.1 Lasso Regression

The cross-validation curve for the lasso model (Figure 4) displays a clear minimum at $\lambda_{\min} = 0.0573$. A more conservative estimate, based on the one-standard-error rule, selects $\lambda_{1\text{se}} = 0.1611$. The corresponding test-set RMSE values (Table 2) show that lasso marginally improves upon M7, but residual-versus-fitted diagnostics (not shown here) continue to exhibit pronounced curvature. Shrinkage alone cannot correct a fundamentally misspecified linear mean function.

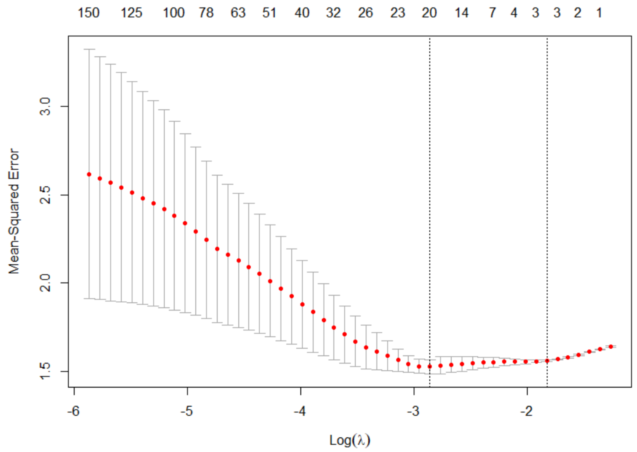


Figure 4: Cross-validation error for the lasso model, showing $\lambda_{\min}$ (left) and $\lambda_{1\text{se}}$ (right).

Model	RMSE
LASSO ($\lambda_{\min}$)	1.233579
LASSO ($\lambda_{1\text{se}}$)	1.248923

Table 2: Test-set RMSE for lasso regression.

4.3.2 Elastic Net Regression

Elastic net regression, combining ℓ_1 and ℓ_2 penalties, selected $\lambda_{\min} = 0.1146$ and $\lambda_{1\text{se}} = 0.3222$ (Figure 5). Like lasso, elastic net reduces the effective model complexity while yielding modest improvements in prediction (Table 3), yet the underlying residual curvature (not shown) persists. Performance is nearly identical to the lasso model, reinforcing that shrinkage alone cannot overcome mean-function misspecification.

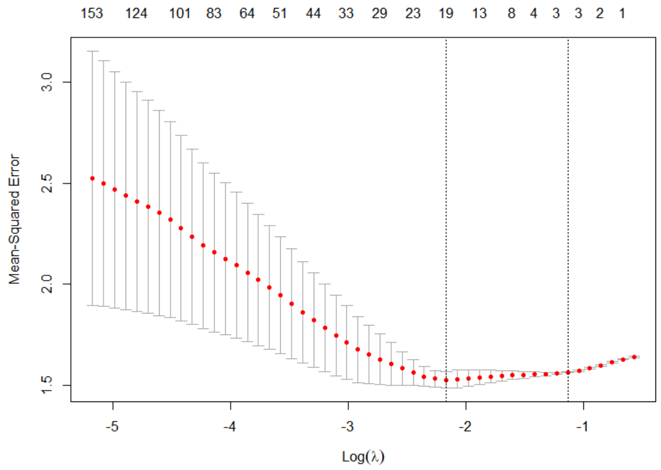


Figure 5: CV error for the elastic net model.

Model	RMSE
Elastic Net ($\lambda_{\min}$)	1.232250
Elastic Net ($\lambda_{1\text{se}}$)	1.249994

Table 3: Test-set RMSE for elastic net regression.

4.3.3 Ridge Regression

Ridge regression, which applies an ℓ_2 penalty that shrinks all coefficients while retaining every predictor, selected $\lambda_{\text{min}} = 10.5556$ and $\lambda_{\text{1se}} = 52.1625$ (Figure 6). As with the other penalized estimators, the residual patterns (not shown) remain non-random and structurally nonlinear, indicating persistent mean-function misspecification. Performance, again, is similar as compared to the other shrinkage methods (Table 4).

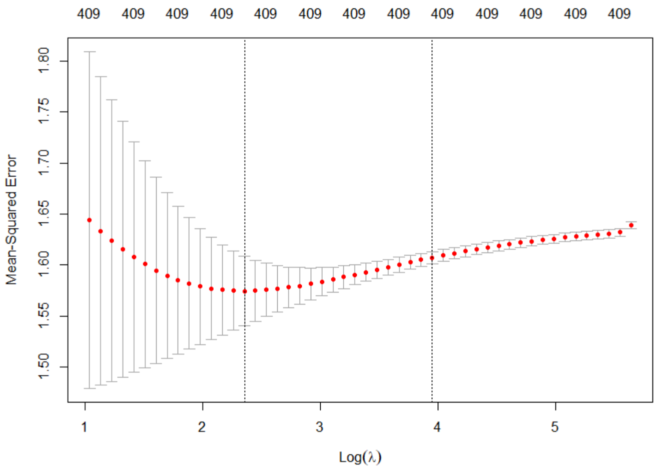


Figure 6: CV error for the ridge model.

Model	RMSE
Ridge (λ_{min})	1.236553
Ridge (λ_{1se})	1.261397

Table 4: Test-set RMSE for ridge regression.

4.4 Interpretability and Recurring Predictor Patterns Across Models

A central distinction between the hierarchical M7 model and the penalized regression models concerns interpretability. M7 respects hierarchical structure, ensuring that interaction terms are only included when their corresponding main effects are present, and preserving clear behavioral meaning for individual coefficients. Penalized estimators, in contrast, prioritize predictive performance over structural coherence: lasso and elastic net routinely violate hierarchy by selecting interaction terms while shrinking associated main effects to zero, and ridge retains all predictors, producing a dense model that is not interpretable in a classical inferential sense. Because all fitted models exhibit strong residual curvature and systematic mean-function misspecification, statistical significance and coefficient magnitudes must be interpreted cautiously; even when a predictor is repeatedly selected across models, its estimated effect may be biased in the presence of an incorrect functional form.

Despite these caveats, several predictors consistently emerge as influential across M7, lasso, and elastic net. Trip duration plays the most persistent role, often accompanied by interactions with Friday, Summer, and scooter usage—patterns that reflect distinct behavioral regimes (e.g., recreational vs. commuting trips). A quadratic effect in the start timestamp is also recurrent. Penalized models frequently retain specific council district interactions, particularly those involving District 9, which suggests localized spatial heterogeneity in trip behavior. These repeated selections provide insight into the multiscale temporal and spatial forces shaping micromobility travel.

5 Conclusion

This study applied an end-to-end linear modeling workflow to a large-scale micromobility dataset, progressing from interpretable hierarchical regression (M7) to high-dimensional penalized models including ridge, lasso, and elastic net. Despite expanding the linear feature space to include all pairwise interactions and quadratic terms, employing influence diagnostics, and testing alternative estimation strategies such as WLS and robust regression, all fitted models exhibited persistent residual curvature and systematic lack of fit. The modest improvements in RMSE offered by penalized regression came at the cost of interpretability and did not alter the fundamental conclusion that the conditional mean function of log trip distance cannot be captured adequately by any linear specification.

These findings underscore a structural limitation: travel behavior in micromobility systems is governed by nonlinear, context-dependent relationships that exceed the representational capacity of linear models, even when richly parameterized. While M7 remains useful as an interpretable descriptive model, and shrinkage methods offer slight predictive gains, none provide a satisfactory account of the underlying data-generating process. Future work should therefore consider nonlinear approaches—such as generalized additive models, spline-based methods, or neural networks—to better model the complex spatial–temporal dynamics observed in this dataset.

Appendix: Code and Reproducibility Materials

All data cleaning scripts, exploratory analyses, preliminary and advanced modeling files used to generate this report are publicly available at: `https://github.com/bryant-willoughby/ST-500-Regression-Analysis-Project`

References

City of Austin (2022). Shared micromobility vehicle trips (2018–2022). Accessed: 2025-11-18.