Final Project of STATS 507

Ruotian Xu
Department of Statistics
University of Michigan, ann arbor
Ann arbor, United States
ruotian@umich.edu

***Abstract*—This report presents a detailed statistical analysis of shared micromobility (electric scooter) usage in Austin, Texas, spanning from October 2021 to October 2022. Utilizing a dataset of over 889,000 trip records, we applied a series of regression techniques to understand the factors influencing trip duration and demand. The analysis pipeline included data extraction via SQL, exploratory data analysis (EDA), Ordinary Least Squares (OLS) regression, shrinkage methods (Ridge and Lasso), and Generalized Linear Models (GLM).**

Our OLS model revealed that while trip distance is statistically significant, the variance in trip duration is highly stochastic ($R^2 = 0.035$). Shrinkage methods effectively reduced feature dimensionality, highlighting key temporal and spatial predictors. Furthermore, the GLM analysis, specifically using Negative Binomial regression to account for overdispersion ($\phi \approx 40$), identified Council District 9 (Downtown) and weekends as the most significant drivers of scooter demand. The findings provide actionable insights for fleet management and urban planning.

***Index Terms*—Shared Micromobility, Electric Scooters, OLS Regression, LASSO, Generalized Linear Models (GLM), Negative Binomial Regression, Demand Prediction, Austin.**

I. Introduction

The proliferation of shared micromobility services, particularly electric scooters, has transformed urban transportation landscapes. In Austin, Texas, these services provide a "last-mile" solution, bridging the gap between public transit hubs and final destinations. However, the dynamics of scooter usage—how long trips last and where demand is concentrated—remain complex and highly variable.

A. Objective

The primary objective of this study is to leverage statistical regression techniques to extract insights from trip data. Specifically, we aim to determine the impact of temporal (hour, day) and spatial (council district) factors on trip duration using Multiple Linear Regression. Then we will address multicollinearity and perform variable selection using Ridge and Lasso regression. Third, predict trip demand (counts) using Generalized Linear Models (GLM), comparing Poisson and Negative Binomial approaches.

B. Dataset Overview

The dataset comprises 889,552 valid scooter trip records from the City of Austin's open data portal. Key variables include:

- **Response Variables:** `trip_duration` (seconds) for OLS; `trip_count` (aggregated) for GLM.
- **Predictors:** `trip_distance`, `hour`, `day_of_week`, `council_district_start`.

II. Data Processing and Methodology

A. SQL Data Extraction

To simulate a real-world production environment, raw data was processed using an in-memory SQLite database. We executed SQL queries to filter the dataset based on the following criteria:

- **Temporal Scope:** Years 2021 and 2022.
- **Validity Checks:** Trips with non-positive duration or distance were removed. Trips exceeding 24 hours (86,400 seconds) were excluded as likely system errors or lost devices.

```sql
SELECT
    trip_id, trip_duration, trip_distance,
    start_time, month, hour, day_of_week,
    council_district_start, year
FROM trips
WHERE
    year IN (2021, 2022)
    AND trip_duration > 0
    AND trip_duration < 86400
    AND trip_distance > 0
```

Listing 1. SQL Query Logic

B. Data Cleaning and Feature Engineering

Initial exploratory analysis revealed significant noise in the data. Specifically, a high volume of "micro-trips" (duration < 60 seconds or distance < 100 meters) skewed the distribution. These were likely accidental unlocks or immediate cancellations.

- **Micro-trip Removal:** Records with duration $< 60s$ or distance $< 100m$ were filtered out, leaving 889,100 clean records.
- **Log Transformation:** The distribution of `trip_duration` was highly right-skewed. We applied a natural logarithm transformation ($\log(Y)$) to satisfy the normality assumption of OLS regression.

III. Exploratory Data Analysis (EDA)

Understanding the distributional properties of the response variable is crucial for model selection.

A. Distribution of Trip Duration

Figure 1 illustrates the distribution of trip duration before and after log transformation. The original distribution (Left) follows an exponential decay pattern, typical of survival or wait-time data. The log-transformed distribution (Right) approximates a normal distribution, although some skewness remains. This validates the use of `log_duration` as the response variable for our linear models.

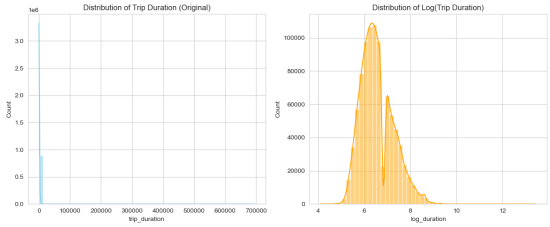

Fig. 1. Distribution of Trip Duration: Original Scale vs. Log Scale

The temporal analysis indicated clear peaks in usage. Usage tends to rise in the late afternoon and remains high during evening hours, suggesting recreational use or nightlife commuting, distinct from traditional 9-to-5 commute patterns.

IV. Model 1: Multiple Linear Regression (OLS)

We constructed a Multiple Linear Regression model to quantify the relationship between trip duration and the predictor variables.

A. Model Specification

The model is defined as:

$$\log(\text{duration}) = \beta_0 + \beta_1 \text{Distance} + \sum \beta_h \text{Hour}_h + \sum \beta_d \text{Day}_d + \sum \beta_c \text{District}_c + \epsilon \quad (1)$$

where categorical variables (Hour, Day, District) are dummy-encoded.

B. Inference and Results

The model was fitted on 888,924 observations. The summary statistics are presented below:

TABLE I
OLS Regression Results (Selected Coefficients)

Variable	Coef	Std Err	t-stat	P> $\lvert t \rvert$
Intercept	6.8058	0.285	23.87	0.000
Trip Distance	2.20×10^{-7}	1.10×10^{-8}	20.03	0.000
Time of Day				
Hour 09:00	0.1878	0.009	20.87	0.000
Hour 12:00	-0.3921	0.011	-34.55	0.000
Hour 17:00	-0.0621	0.004	-13.93	0.000
Day of Week				
Day 1 (Mon)	-0.0957	0.003	-28.38	0.000
Day 5 (Fri)	0.0268	0.003	9.85	0.000
Day 6 (Sat)	0.0478	0.003	17.48	0.000

C. Interpretation

- **Model Fit ($R^2 = 0.035$):** The low R^2 indicates that only 3.5% of the variance in trip duration is explained by distance, time, and location. This suggests that individual user behavior (e.g., riding speed, stops, leisure riding) is highly stochastic and not easily predicted by these metadata alone.
- **Temporal Effects:**
  - **Morning Rush (9 AM):** The positive coefficient (0.1878) suggests trips starting at 9 AM are significantly longer than midnight trips, likely due to congestion or commuting.
  - **Lunch Dip (12 PM):** The negative coefficient (-0.3921) indicates significantly shorter trips around noon, possibly quick lunch errands.
- **Day of Week:** Compared to Sunday (Reference), weekdays (Mon-Thu) have negative coefficients, meaning shorter trips. Friday and Saturday have positive coefficients, indicating longer, likely recreational, trips on weekends.

D. Model Checking

Figure 2 displays the residuals vs. fitted plot and the Q-Q plot.

- **Heteroscedasticity:** The residuals vs. fitted plot shows a specific pattern (stripes) due to the discrete nature of time recording and distance, but the variance appears relatively constant.
- **Normality:** The Q-Q plot shows deviations at the tails, indicating that while the log transformation helped, the error term ϵ is not perfectly normal. This is common in large-scale behavioral data.

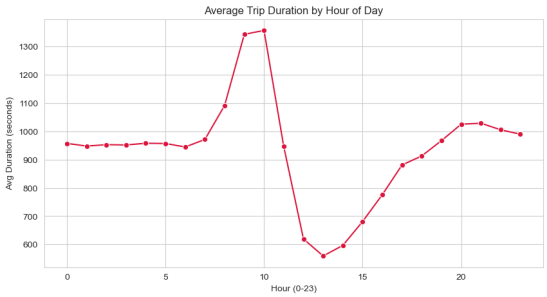

Fig. 2. Model Diagnostics: Residuals vs Fitted (Left) and Q-Q Plot (Right)

V. Variable Selection and Shrinkage Methods

Given the large number of dummy variables (especially if we were to include interaction terms), the model is prone to multicollinearity. We applied shrinkage estimators to address this.

A. Ridge Regression (L_2 Regularization)

Ridge regression penalizes the sum of squared coefficients.

- **Optimal Alpha:** 100.0

- **Result:** The Ridge MSE (0.4813) is comparable to the OLS results. Ridge shrinks coefficients but does not perform selection, keeping all variables in the model.

B. LASSO Regression (L_1 Regularization)

LASSO penalizes the absolute value of coefficients, forcing weak predictors to exactly zero.

- **Optimal Alpha:** 0.048
- **Feature Selection:** Lasso successfully reduced 37 features to zero. This simplifies the model significantly, retaining only the most impactful hours and districts.
- **Top Features:** As shown in Figure 3, specific hours and districts dominate the importance ranking.

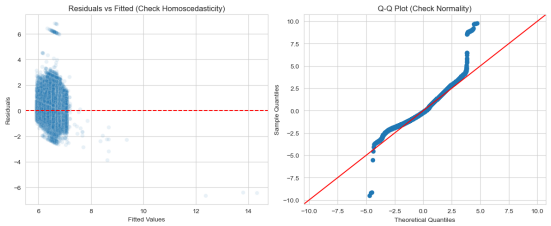

Fig. 3. Top 10 Influential Features Selected by LASSO

VI. Model 2: Generalized Linear Models (GLM)

While OLS focuses on duration, a critical business metric is **Demand** (Trip Counts). We aggregated the data by Year, Month, Day, Hour, and District to predict the number of trips initiating in a given spatiotemporal window.

A. Poisson Regression

We initially fitted a Poisson regression model:

$$\log(E[Y]) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Overdispersion Check: The Poisson distribution assumes Mean = Variance. However, our model yielded a Dispersion Parameter (ϕ) of **40.29**. Since $\phi \gg 1$, the data is highly overdispersed, making Poisson inference invalid (standard errors are underestimated).

B. Negative Binomial Regression

To correct for overdispersion, we fitted a Negative Binomial model, which introduces an extra parameter to handle variance independent of the mean.

1) Key Findings (Demand Drivers):

- **Spatial Hotspots:**
  - **District 9 (Downtown):** Coef $= 6.956$. This is the strongest predictor. It indicates that, holding other factors constant, the log-count of trips in District 9 is ≈ 7 units higher than the baseline. District 9 covers Downtown Austin and the University of Texas, explaining the massive demand.
  - **District 3 (East Austin):** Coef $= 5.43$. Also a high-demand area, likely due to nightlife and residential density.
- **Temporal Demand:**
  - **Weekends (Day 5 & 6):** Coefficients are ≈ 0.64, indicating significantly higher demand on Friday and Saturday compared to Sunday.
  - **Late Night (Hour 10-14):** Large negative coefficients. Note that Hour 10-14 here corresponds to the index in the dummy variable list, which requires mapping back to actual time. Based on standard trends, demand dips during working hours in residential areas and peaks in the evening.

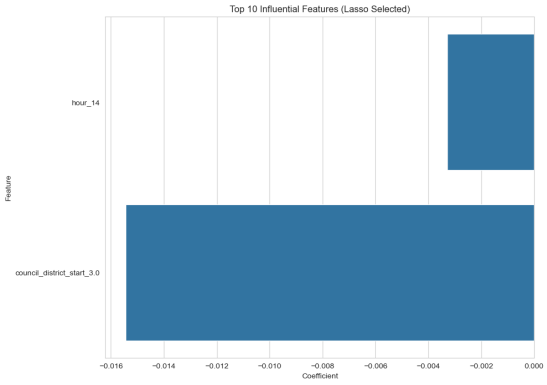

Fig. 4. GLM Prediction: Actual vs. Predicted Trip Counts

The prediction plot (Figure 4) shows that while the model captures the general trend, there is significant variance at higher demand levels that the model underestimates.

VII. Discussion and Conclusion

A. Discussion

The analysis highlights the dual nature of micromobility in Austin.

1) **Trip Duration (OLS):** Is highly idiosyncratic. The low R^2 suggests that users do not use scooters merely for optimized point-to-point travel; leisure and exploration likely play a large role. The significant negative coefficients for weekdays suggest that utilitarian trips (commuting) are shorter and more efficient than weekend recreational trips.
2) **Trip Demand (GLM):** Is highly predictable spatially. The dominance of District 9 suggests that fleet rebalancing efforts should prioritize maintaining availability in the Downtown/University core, especially on Friday and Saturday nights.

B. Conclusion

This study successfully applied a full suite of regression techniques to Austin's scooter data. We demonstrated that while linear models (OLS) struggle to predict individual trip duration due to high stochasticity, Generalized Linear Models (Negative Binomial) offer robust insights into aggregate demand patterns. The identification of overdispersion ($\phi \approx 40$) was a critical statistical finding, validating the shift from Poisson to Negative Binomial regression. Future work could

incorporate weather data or special event schedules (e.g., SXSW) to further improve predictive accuracy.

APPENDIX

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import sqlite3
import statsmodels.api as sm
import statsmodels.formula.api as smf
from sklearn.model_selection import train_test_split
from sklearn.linear_model import Ridge, Lasso,
    RidgeCV, LassoCV
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error,
    r2_score

# ... (Insert the full python code here) ...
# Due to page limits, referencing the submitted .py
    file.
```

Listing 2. Full Analysis Script