# Practice 6: Generalized Linear Models Analysis of NYC Taxi Data

Oleksandr Severhin

November 2025

## 1 Introduction and Data Description

This report extends previous linear regression analyses of the NYC Yellow Taxi dataset (January 2025) by applying Generalized Linear Models (GLMs). While Ordinary Least Squares (OLS) assumes normally distributed errors and constant variance, taxi data often violates these assumptions—costs are strictly positive and right-skewed, and tipping behavior is binary.

We processed the dataset to create two distinct modeling tasks:

1. **Total Cost Prediction (Continuous):** Modeling `total_amount` ($Y_{cost}$) using trip distance, fare, tolls, and temporal factors.
2. **Tipping Likelihood (Binary):** Modeling `is_tipper` ($Y_{tip}$), a binary variable indicating whether a passenger left a tip ($> 0$) or not. This analysis was restricted to credit card trips to ensure data validity.

## 2 Model Specification and Justification

### 2.1 Gamma GLM for Total Cost

For the continuous response $Y_{cost}$, we selected a Gamma distribution with a Log link function.

- **Family (Gamma):** The target variable '*total_amount*' is strictly positive and right-skewed. The Gamma distribution naturally handles heteroscedasticity, where variance increases with the mean ($\text{Var}(Y) \propto \mu^2$), a common characteristic of financial cost data.
- **Link (Log):** The log link function, $\eta = \log(\mu)$, ensures that predicted costs are always positive ($e^{\eta} > 0$) and allows coefficients to be interpreted multiplicatively, which is intuitive for pricing models.

### 2.2 Binomial GLM for Tipping Behavior

For the binary response $Y_{tip}$, we selected a **Binomial distribution** with a **Logit link function**.

- **Family (Binomial):** The outcome is dichotomous (Tip vs. No Tip).
- **Link (Logit):** The logit link, $\eta = \log(\frac{p}{1-p})$, maps the linear predictor to the probability interval $[0, 1]$, avoiding the invalid predictions (e.g., negative probabilities) that can occur with OLS.

## 3 Results and Interpretation

### 3.1 Total Cost Model Performance

The Gamma GLM achieved a Test RMSE of **3.203**, marginally outperforming the log-transformed OLS model (RMSE **3.211**). While the predictive gains are modest, the Gamma GLM offers a theoretical advantage: it models the cost directly on the original scale without requiring bias-correction transformations.

Table 1: Performance Comparison: Gamma GLM vs. OLS

| Model | Distribution | Link | Test RMSE |
|---|---|---|---|
| OLS (Log-Transformed) | Gaussian | Identity (Log Y) | 3.211 |
| **Gamma GLM** | **Gamma** | **Log** | **3.203** |

Diagnostics confirm the appropriateness of the Gamma family. The plot of Pearson residuals shows a random scatter around zero with no clear "fanning" pattern, indicating that the Gamma distribution successfully accounted for the non-constant variance.

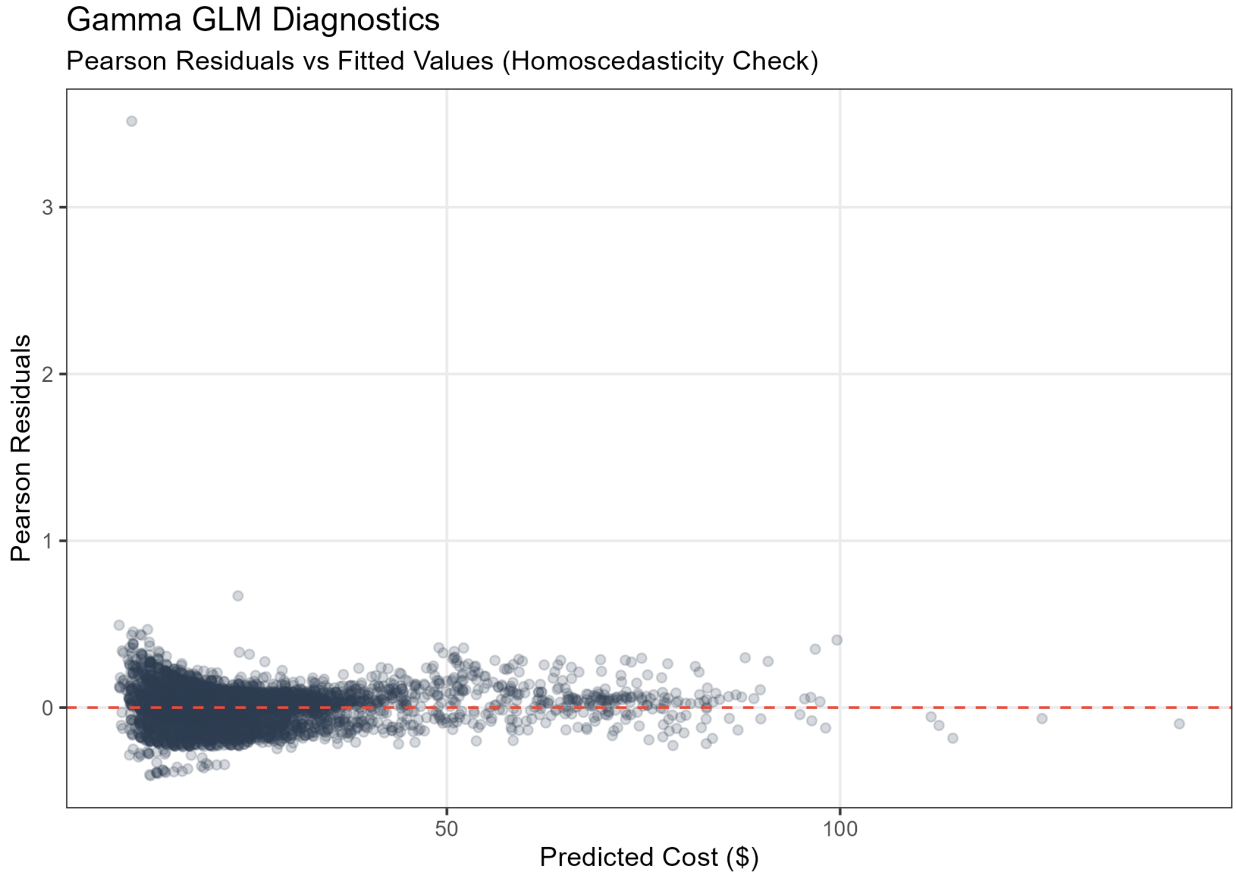


Figure 1: Gamma GLM Diagnostics: Pearson Residuals vs. Fitted Values. The random scatter indicates that the heteroscedasticity has been adequately modeled.

## 3.2 Tipping Behavior Model Interpretation

The Binomial GLM provided insights into the drivers of tipping. The intercept was extremely high (Odds Ratio $\approx 19.7$), reflecting the base reality that the vast majority of credit card passengers do tip.

Key coefficient interpretations (Odds Ratios):

- **Pickup Hour (OR $\approx$ 1.02):** For every hour later in the day, the odds of receiving a tip increase by roughly 2%.
- **Trip Distance (OR $\approx$ 0.97):** Interestingly, longer trips are associated with a slightly *lower* likelihood of tipping (3% decrease in odds per mile), perhaps due to higher total costs squeezing the passenger's willingness to pay extra.

However, the model's predictive power was limited. The McFadden pseudo-$R^2$ was very low (0.0059), and the ROC curve area (AUC = 0.558) indicates the model is only slightly better than random guessing. This suggests that tipping is highly stochastic or driven by unobserved variables (e.g., driver friendliness, car cleanliness) rather than the observed trip metrics.

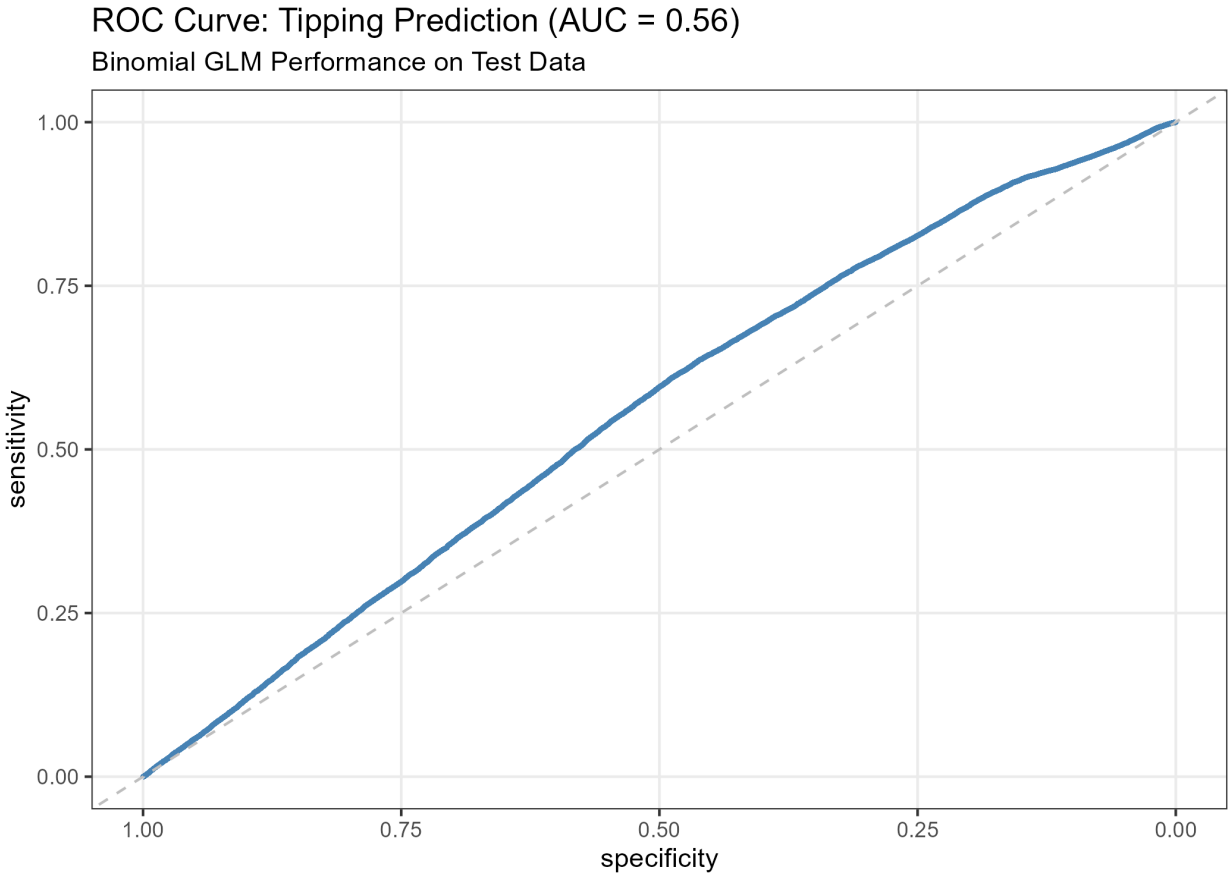


Figure 2: ROC Curve for Tipping Prediction. An AUC of 0.56 indicates weak discrimination capability.

# 4 Peer Feedback Question

**Question:** "In my Gamma GLM analysis, the predictive performance (RMSE) was virtually identical to the Log-Log OLS model. Given that OLS is computationally faster and easier to explain to non-technical stakeholders, is the theoretical correctness of the Gamma GLM (handling heteroscedasticity directly) sufficient justification to prefer it over OLS in a business context?"

# 5 Answer to Exam Question

**Question:** How should I interpret a log-odds coefficient for a categorical predictor?

**Answer:** A log-odds coefficient $\beta_j$ for a categorical predictor (e.g., 'ColorRed') represents the difference in the log-odds of the outcome occurring between that category ('Red') and the reference category (e.g., 'Blue'), holding all other variables constant. To make this interpretable, we exponentiate the coefficient ($e^{\beta_j}$) to get the Odds Ratio. If $e^{\beta_j} = 2$, it means the odds of the event occurring for the 'Red' group are **2 times** the odds for the reference group.