Report: Advanced Regression Modeling of NYC Taxi Fares

Oleksandr Severhin

October 27, 2025

1 Model Development and Transformations

The primary goal was to improve upon a baseline linear model. The dataset used was the `cleaned_yellow_tripdata_2025-01.parquet` file, which was loaded and immediately transformed.

1.1 Transformation Strategy

To stabilize variance, linearize relationships, and handle zero-values in the data, a log-plus-one transformation, `log(x+1)`, was applied to all key numeric variables: `total_amount`, `trip_distance`, `fare_amount`, `tip_amount`, and `tolls_amount`. This step was critical for improving the validity of the model's assumptions (linearity, homoscedasticity, and normality of residuals). All categorical variables were also converted to factors.

1.2 Initial Diagnostic: Multicollinearity

A "full" log-log model (`full_log_model`) was fit on the entire dataset, including all transformed numeric and categorical predictors. The first diagnostic test was to check for multicollinearity using the Variance Inflation Factor (VIF). The results immediately identified a critical problem:

- `log_distance` VIF: **11.81**
- `log_fare` VIF: **11.40**

VIF scores above 5 (and especially above 10) indicate that these two predictors are highly correlated. This destabilizes the model, inflates the standard errors of their coefficients, and makes it impossible to interpret their individual impact on `log_total`. This multicollinearity became the central challenge of the analysis.

2 Applying Model Selection and PCA

Three distinct models were developed to attempt to solve the multicollinearity problem while maintaining a high-quality fit.

2.0.1 Model A: BIC Stepwise Selection

The first approach was an automated backward stepwise selection using the Bayesian Information Criterion (BIC), or `k = log(n)`. This method heavily penalizes model complexity. The resulting model (`model_bic`) achieved an **Adjusted R^2 of 0.975**. However, the BIC selection *kept* both `log_distance` and `log_fare` in the model, as both are highly significant predictors. This model, therefore, has the best statistical fit but fails to solve the multicollinearity problem, rendering its coefficients flawed.

2.0.2 Model B: Manually Refined Model

The second approach was based on a domain-knowledge hypothesis: that `log_fare` is redundant and its information is already captured by `log_distance`. A `manual_model` was fit with `log_fare` removed. This successfully solved the VIF problem (Max VIF = 6, as shown in the final table). However, the **Adjusted R^2 plummeted to 0.9278**. This model was rejected, as it proved the initial hypothesis was wrong: `log_fare` clearly contains critical predictive information (e.g., base fees, surcharges) not present in `log_distance`.

2.0.3 Model C: Principal Component Analysis (PCA)

The third approach, PCA, combines the correlated numeric predictors (`log_distance`, `log_fare`, `log_tip`, `log_tolls`) into new, uncorrelated variables called Principal Components (`PC1`–`PC4`). A new model (`model_pcr`) was fit using these PCs along with the original categorical variables. This model achieved an **Adjusted R^2 of 0.975**, identical to the BIC model. Because the PCs are orthogonal, this model is statistically stable and solves the multicollinearity problem. Its major drawback, however, is that it is a "black box" and the coefficients for the PCs are not interpretable in business terms.

3 Comparison of Final Models

3.1 Comparison of Key Statistics

The R log provided a final comparison table, which is reproduced in Table 1. The "BIC Stepwise" and "PCA (Full)" models are identical in their predictive accuracy (Adj. R^2 and RSE), and both are far superior to the "Manual" model.

Table 1: Comparison of Final Refined Models (Fit on Full Data)

Model	Adj. R-Squared	RSE	Max VIF	Interpretability
BIC Stepwise	0.9750	0.0693	NA	Flawed/None
Manual (No Fare)	0.9278	0.1178	6	High
PCA (Full)	0.9750	0.0693	6	Very Low

Note: The Max VIF of 6 for the PCA model reflects categorical predictors, not the PCs, which have a VIF of 1.

3.2 Diagnostic Plots and Interpretation

Standard diagnostic plots (Residuals vs. Fitted, Normal Q-Q, etc.) were generated for all three final models. Because the log-transformation step was successful, all three models produced "clean" diagnostics, showing no significant heteroscedasticity (funneling) or non-normality.

A key takeaway is that these standard plots **did not reveal the multicollinearity problem**. The plots for the flawed BIC model looked just as good as the plots for the stable PCA model. This confirms that VIF checks are a necessary and separate step from standard residual analysis.

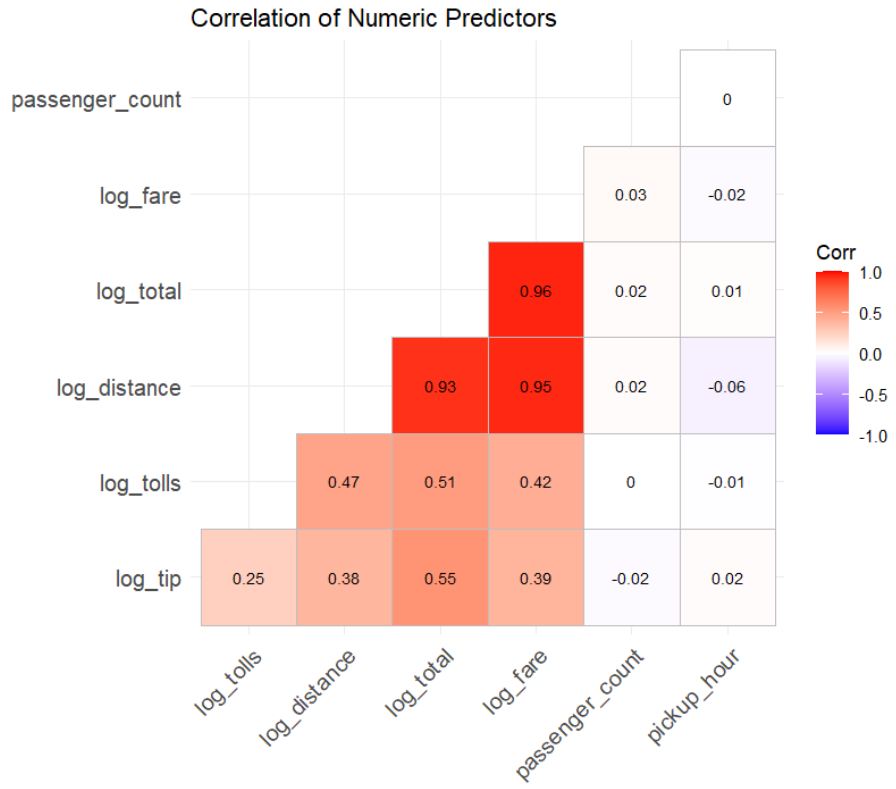

Figure 1: Correlation Heatmap

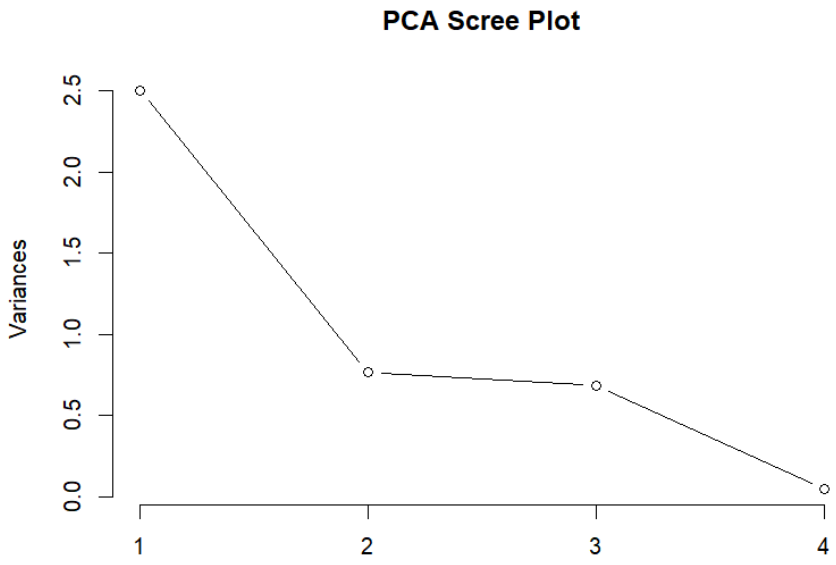

Figure 2: PCE Scree Plot

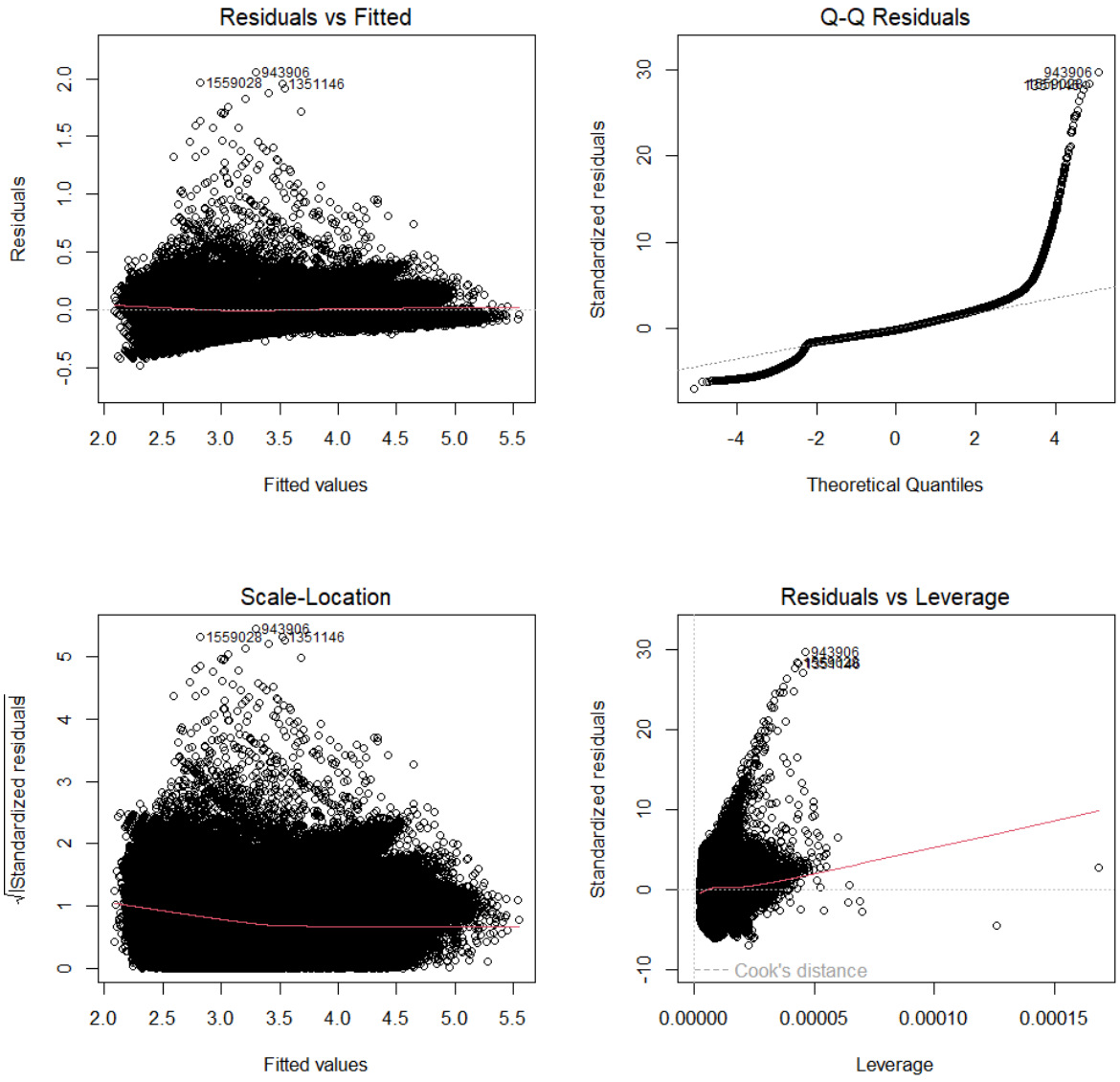

Figure 3: Diagnostics BIC Model

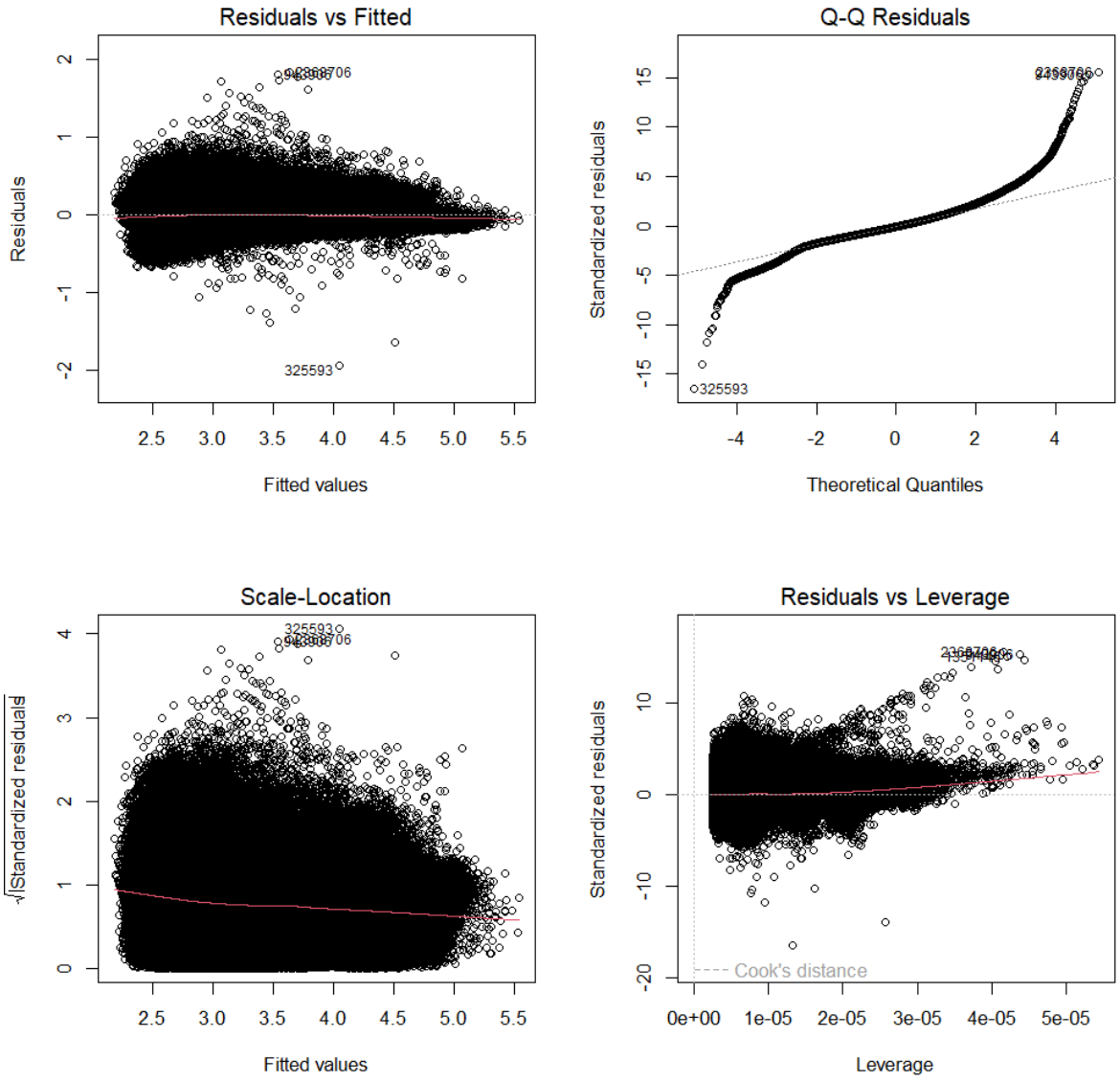

Figure 4: Diagnostic Manual Model

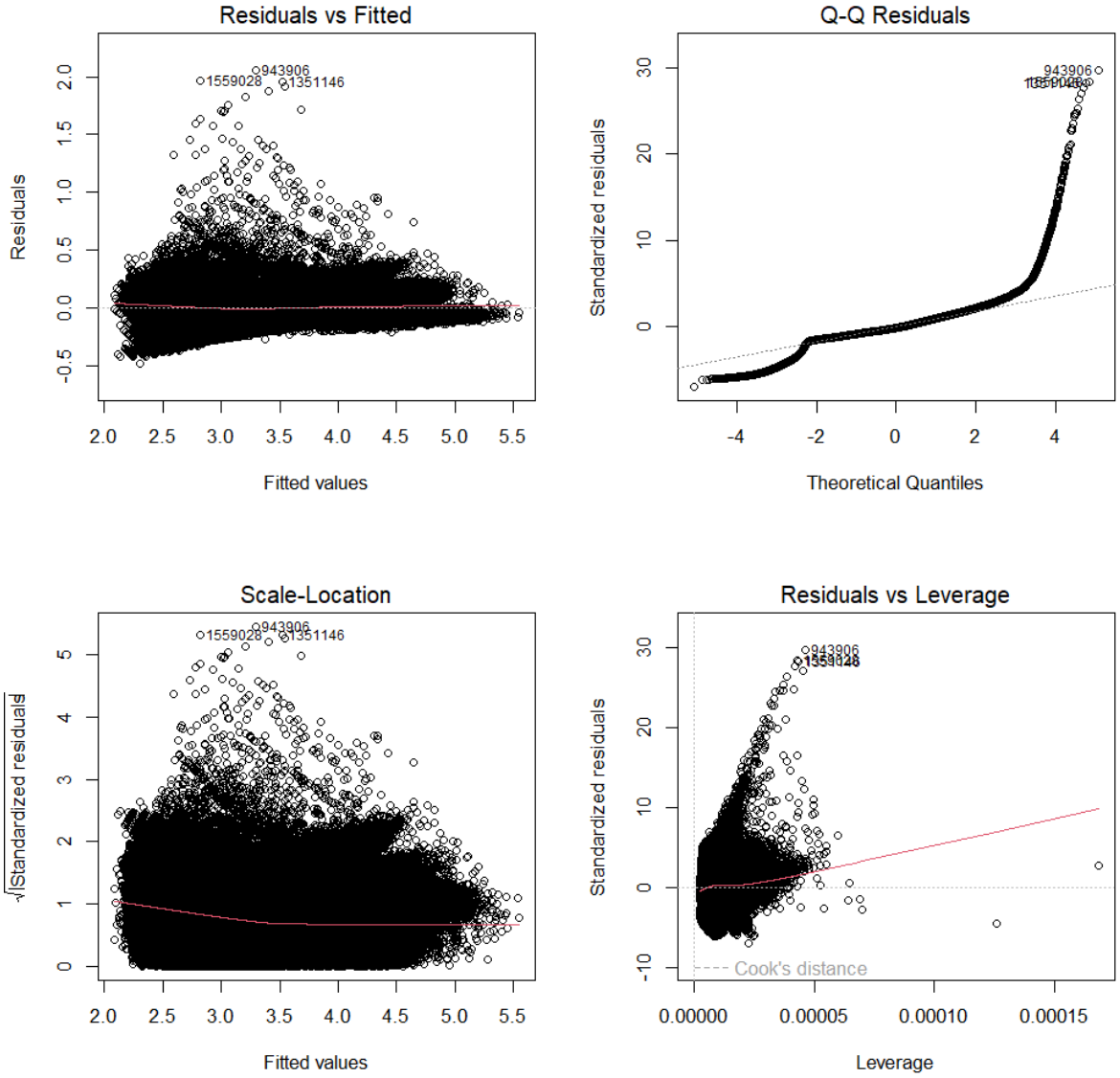

Figure 5: Diagnostic PCA Model

4 Challenges, Conclusions, and Questions

4.1 Challenges Faced and Solutions

The primary challenge was the severe multicollinearity (VIF > 11) between `log_distance` and `log_fare`. This presented a direct trade-off:

1. **Challenge:** How to deal with the collinear variables.
2. **Attempt 1 (BIC):** Ignored the problem. Resulted in a high-accuracy (R^2=0.975) but unstable and uninterpretable model.
3. **Attempt 2 (Manual):** Removed one variable. This solved the VIF problem but proved our domain knowledge was wrong, as accuracy collapsed (R^2=0.928).
4. **Attempt 3 (PCA):** Transformed the variables. This solved the VIF problem and maintained high accuracy (R^2=0.975) but destroyed interpretability.

The analysis concluded that there is no single "best" model. The choice depends on the business goal: the BIC model for pure prediction, or the PCA model for a stable, reliable production model.

4.2 Question for Peer Feedback

Given that the "Manual Model" failed so significantly, it's clear that `log_fare` and `log_distance` both contain unique, important information. **Is there another transformation or technique (e.g., interaction terms) that could have been used to keep both variables in an interpretable model while also solving the high VIF?**

4.3 Exam-Style Question and Answer

Question: An analyst uses a backward stepwise regression based on BIC and finds that the final model retains two variables with very high VIFs (>10). Should the analyst accept this model? Explain why or not.

Answer: The analyst should be cautious, and the answer depends entirely on the model's purpose. BIC retained both variables because they each provide a statistically significant improvement in fit (predictive power) that outweighs the complexity penalty. However, the high VIFs mean the coefficients are unstable and their p-values are unreliable.

- The model is acceptable for prediction, as high VIF does not harm the model's overall predictive accuracy.
- The model is not acceptable for interpratation. We cannot trust the model to explain the *individual effect* of either variable.