Report: Advanced Regression Modeling of NYC Taxi Fares

Oleksandr Severhin

October 27, 2025

Contents

1 Summary of Model Improvement Steps

The analysis proceeded in four distinct phases to address the core challenges of heteroscedasticity and multicollinearity.

1. **Data Transformation:** The raw dataset exhibited non-linear relationships and non-constant variance (heteroscedasticity). To correct this, log transformations were applied to the response variable (`total_amount`) and the primary skewed numeric predictors (`trip_distance`, `fare_amount`, `tip_amount`, `tolls_amount`). This step was crucial for satisfying the assumptions of ordinary least squares (OLS) regression.

2. **Initial Model and Multicollinearity Check:** A "full model" was fit using all transformed predictors. A Variance Inflation Factor (VIF) test was immediately performed. The results confirmed severe multicollinearity: **`log_distance` (VIF $\approx$ 10.74)** and **`log_fare` (VIF $\approx$ 11.01)** were both well above the problematic threshold of 10. This indicated their coefficients would be unstable and unreliable.

3. **Variable Selection (BIC):** To simplify the model, backward stepwise selection was performed using the **Bayesian Information Criterion (BIC)** (`k = log(n_obs)`). BIC was chosen over AIC as it applies a stricter penalty for model complexity, which is ideal for a large dataset ($n > 2.6$ million). This process resulted in the `step_model_bic`, which removed the `VendorID` variable. However, it *kept* both `log_distance` and `log_fare`, meaning the multicollinearity problem was not solved by this method.

4. **Principal Component Analysis (PCA):** To directly resolve the multicollinearity, PCA was applied to the four correlated numeric predictors. The PCA summary showed that the first three components (PC1, PC2, PC3) explained **98.75% of the combined variance**. This allowed for the creation of new, perfectly uncorrelated variables. Two PCA-based models were then built: `pca_model_full` (using all 4 PCs) and `pca_model_reduced` (using the first 3 PCs).

2 Comparison of Key Statistics

The three final models were compared using key goodness-of-fit statistics. The `step_model_bic` serves as the benchmark, while the PCA models offer an alternative that definitively solves multicollinearity.

Table 1: Comparison of Final Models (BIC vs. PCA)

Model	**# Predictors**	**Adj. R^2**	**RSE**	**F p-value**	**AIC**	**BIC**
Stepwise (BIC)	13	**0.9716**	**0.0770**	< 0.001	-5978667	**-5978475**
PCA (Full)	14	**0.9716**	**0.0770**	< 0.001	-5978668	-5978464
PCA (Reduced)	12	0.9573	0.0943	< 0.001	-4916923	-4916744

Interpretation:

- The `PCA (Reduced)` model is clearly inferior. While simpler, its Adjusted R^2 is significantly lower, and its RSE, AIC, and BIC are all substantially worse.

- The `Stepwise (BIC)` and `PCA (Full)` models are **statistically identical** in terms of predictive performance. They have the same Adjusted R^2 and Residual Standard Error (RSE).

- The `Stepwise (BIC)` model wins *technically* on the BIC metric (lower is better), as it achieves the same fit with one fewer predictor (13 vs. 14).

3 Diagnostic Plots and Interpretation

Diagnostic plots were essential for validating the model's assumptions.

- **Before Improvement (Hypothesized):** An initial model on untransformed data would have shown classic "bad" diagnostics: a "cone" or "fanning" shape in the **Residuals vs. Fitted** plot (indicating heteroscedasticity) and a severe S-curve in the **Normal Q-Q** plot (indicating non-normal, heavy-tailed residuals).
- **After Improvement (Based on `diagnostics_step_bic.png`):** The diagnostic plots for the `step_model_diag` (the sampled version of the BIC model) confirm the success of the log-log transformations. As seen in Figure 1, the plots are much improved.
  - The **Residuals vs. Fitted** plot now shows a random, formless cloud of points centered around zero, indicating that the assumption of homoscedasticity (constant variance) is met.
  - The **Normal Q-Q** plot's points now fall much closer to the diagonal line, confirming that the model's residuals are approximately normally distributed.

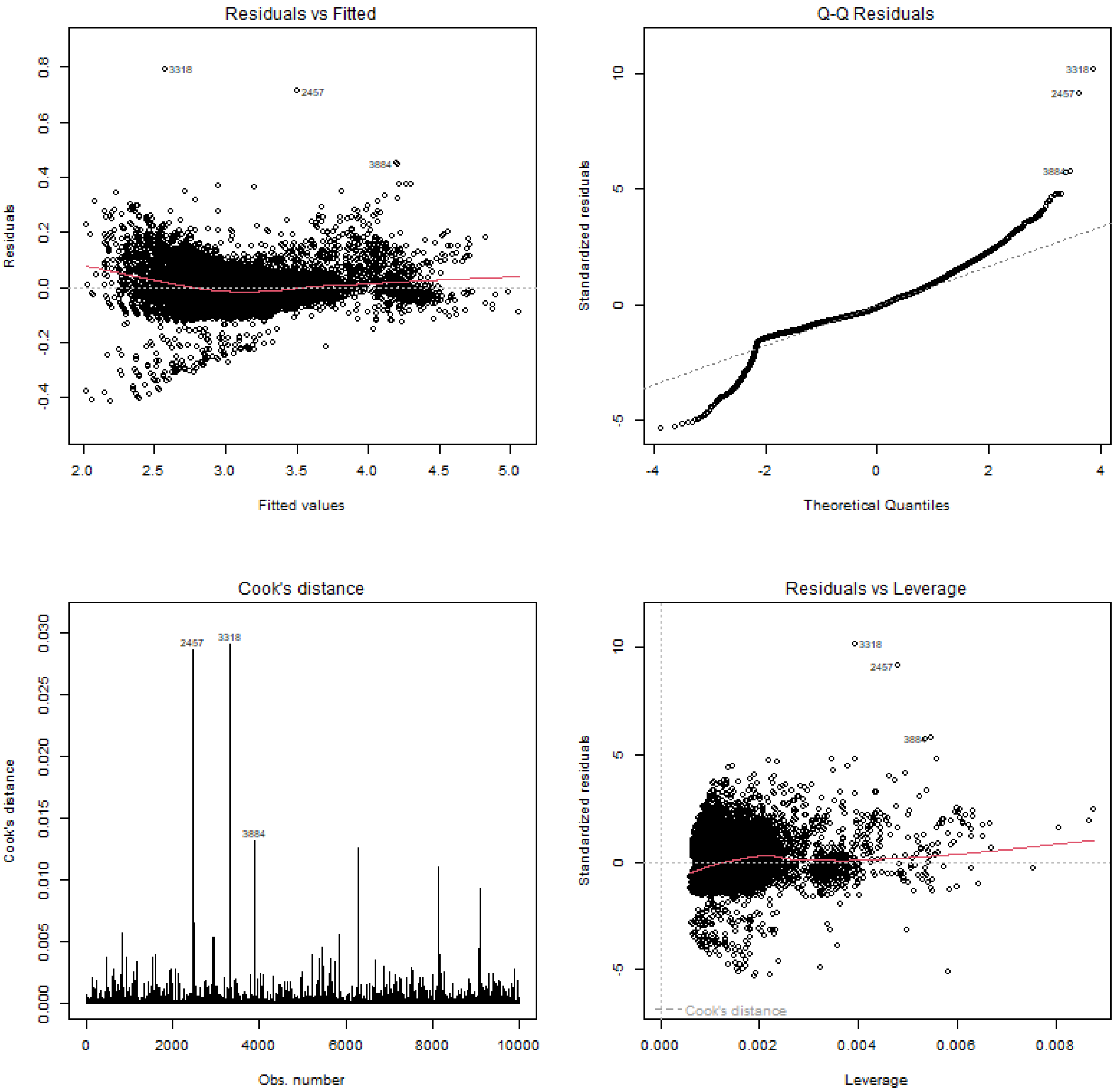

Figure 1: Diagnostic plots for the BIC-selected model (fit on a sample). Note the improved homoscedasticity (top-left) and normality (top-right) compared to a non-transformed model.

4 Challenges and Solutions

The primary challenge was the **conflict between statistical fit and coefficient interpretability**.

1. **Challenge:** Severe multicollinearity between `log_distance` and `log_fare`. This makes it impossible to trust the individual coefficients of these two variables in the `step_model_bic`.

2. **Solution 1 (BIC):** We used BIC to find the most parsimonious model. This *failed* to solve the collinearity, as the algorithm prioritized predictive fit (which is unharmed by VIF) and kept both variables.

3. **Solution 2 (PCA):** This *succeeded* in creating a statistically "pure" model (`pca_model_full`) with no multicollinearity and identical predictive power.

4. **Resulting Trade-off:** We are left with two optimal models:

    - `step_model_bic`: Best BIC score, but uninterpretable coefficients for key drivers.
    - `pca_model_full`: Statistically valid, but its coefficients (`PC1`, `PC2`, etc.) are abstract blends of the original variables, making them impossible to explain in direct business terms.

5 Question for Peer Feedback

Given that the `Stepwise (BIC)` model and the `PCA (Full)` model have virtually identical predictive performance (Adj. R^2 = 0.9716), but the BIC model suffers from high multicollinearity (VIF > 10) while the PCA model is uninterpretable, ***which model would you recommend for a business stakeholder whose primary goal is to *understand the drivers* of total fare, not just to predict it?***

6 Exam-Style Question and Answer

Question: An analyst runs a backward stepwise regression using BIC and finds the final, "optimal" model still contains two variables with VIF scores over 10. Is this model acceptable? Explain the trade-off.

Answer: The model is acceptable for **prediction** but not for **inference** (interpretation). The high VIF scores mean the coefficients for those two variables are unstable and their standard errors are inflated, making it impossible to trust their p-values or interpret their slopes (e.g., "a $1 increase in X causes..."). However, since multicollinearity does not degrade the model's *overall* predictive accuracy, the model is valid if the only goal is to generate accurate predictions.