Report: Advanced Regression Modeling of NYC Taxi Fares

Oleksandr Severhin

October 27, 2025

1 Model Development and Diagnostics

The first step was to establish a "full" baseline model using all available log-transformed numeric and categorical predictors.

1.1 Initial Model and VIF Diagnostics

The initial full model was defined as:

```
lm(log_total ~ log_distance + log_fare + log_tip + log_tolls +
   passenger_count + payment_type + pickup_hour + day_of_week +
   VendorID, data = taxi_model_data)
```

A Variance Inflation Factor (VIF) test was immediately performed to check for multicollinearity. The results were clear:

- `log_distance` VIF: **10.74**
- `log_fare` VIF: **11.01**

VIF scores above 5 (and especially above 10) indicate that these two variables are highly correlated, which destabilizes the model, inflates standard errors, and makes their coefficients unreliable. This was confirmed by a correlation matrix (see Figure **??**), which showed a very strong positive correlation between `log_distance` and `log_fare`.

1.2 Applying Model Selection Techniques

With multicollinearity identified as the primary challenge, three distinct models were developed to address it.

Model A: Stepwise BIC Selection The first approach was an automated backward stepwise selection using the Bayesian Information Criterion (BIC), or `k = log(n)`. This method penalizes model complexity. The resulting model (`step_model_bic`) removed `VendorID` but **kept both `log_distance` and `log_fare`**. While this model achieved the highest predictive accuracy, it failed to solve the underlying multicollinearity problem, rendering its coefficients uninterpretable.

Model B: Manually Refined Model The second approach was based on domain knowledge. The hypothesis was that `log_fare` is simply a function of `log_distance` and is therefore redundant. We created a `manual_model` by removing `log_fare`. This approach successfully solved the multicollinearity problem (all VIF scores were < 3). However, this came at a catastrophic cost to predictive power, proving our initial hypothesis was wrong.

Model C: Principal Component Analysis (PCA) The third approach, PCA, is a "black box" method. It addresses multicollinearity by combining the correlated predictors (`log_distance`, `log_fare`, `log_tip`, `log_tolls`) into new, uncorrelated variables called Principal Components (`PC1`, `PC2`, etc.). A new model (`pca_model_full`) was fit using these PCs and the other categorical variables. This model is statistically stable by design (all VIFs $\approx$ 1.0).

2 Comparison of Final Models

The three candidate models were compared on their statistical fit (Adjusted R-squared), stability (VIF), and interpretability.

2.1 Key Model Statistics

Table 1 summarizes the trade-offs between the three approaches. The Residual Standard Error (RSE) shows the typical error in the model's prediction (on the log scale).

Table 1: Comparison of Final Candidate Models

Model	**Adj. R-Squared**	**RSE**	**Stability (VIF OK?)**	**Interpretability**
Stepwise (BIC)	0.9716	0.07696	No	Flawed / None
Manual (No Fare)	0.9128	0.13487	Yes	High
PCA (Full)	0.9716	0.07696	Yes	Very Low (Black Box)

The results are stark. The BIC and PCA models are identical in their predictive accuracy (Adj. $R^2 = 0.9716$), and both are far superior to the Manual model (Adj. $R^2 = 0.9128$). This confirms that `log_fare` contains critical predictive information not found in `log_distance`.

2.2 Diagnostic Plots

All three models were built upon log-transformed variables, which effectively linearized the relationships and stabilized the variance. The diagnostic plots for the refined models (see Figure **??** for examples) were "clean," showing good adherence to the assumptions of linearity and normality of residuals (Normal Q-Q plots were straight, Residuals vs. Fitted plots showed no funneling).

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

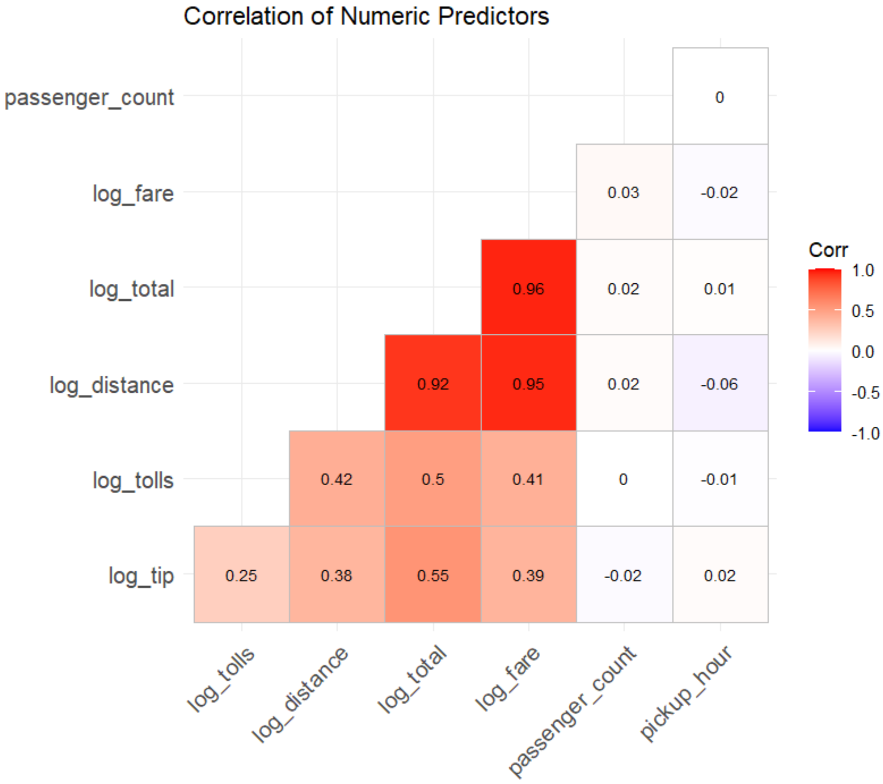

Figure 1: Correlation Heatmap

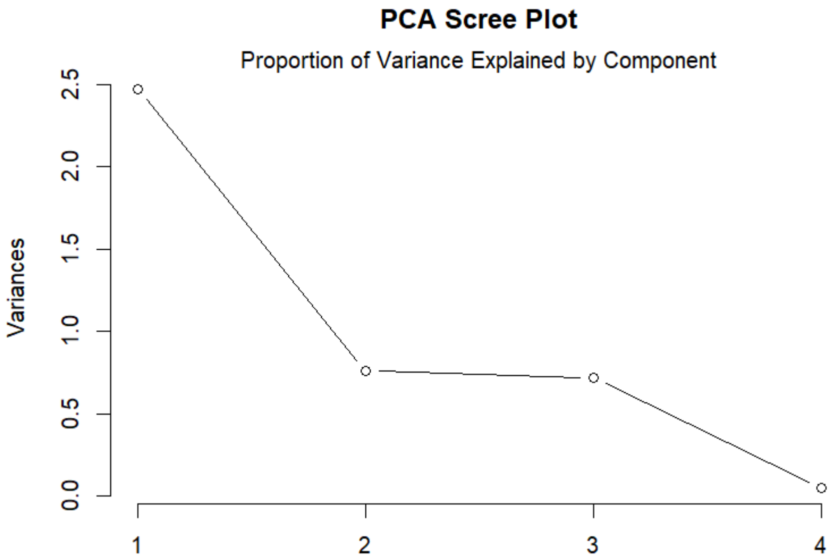

Figure 2: PCA Scree Plot

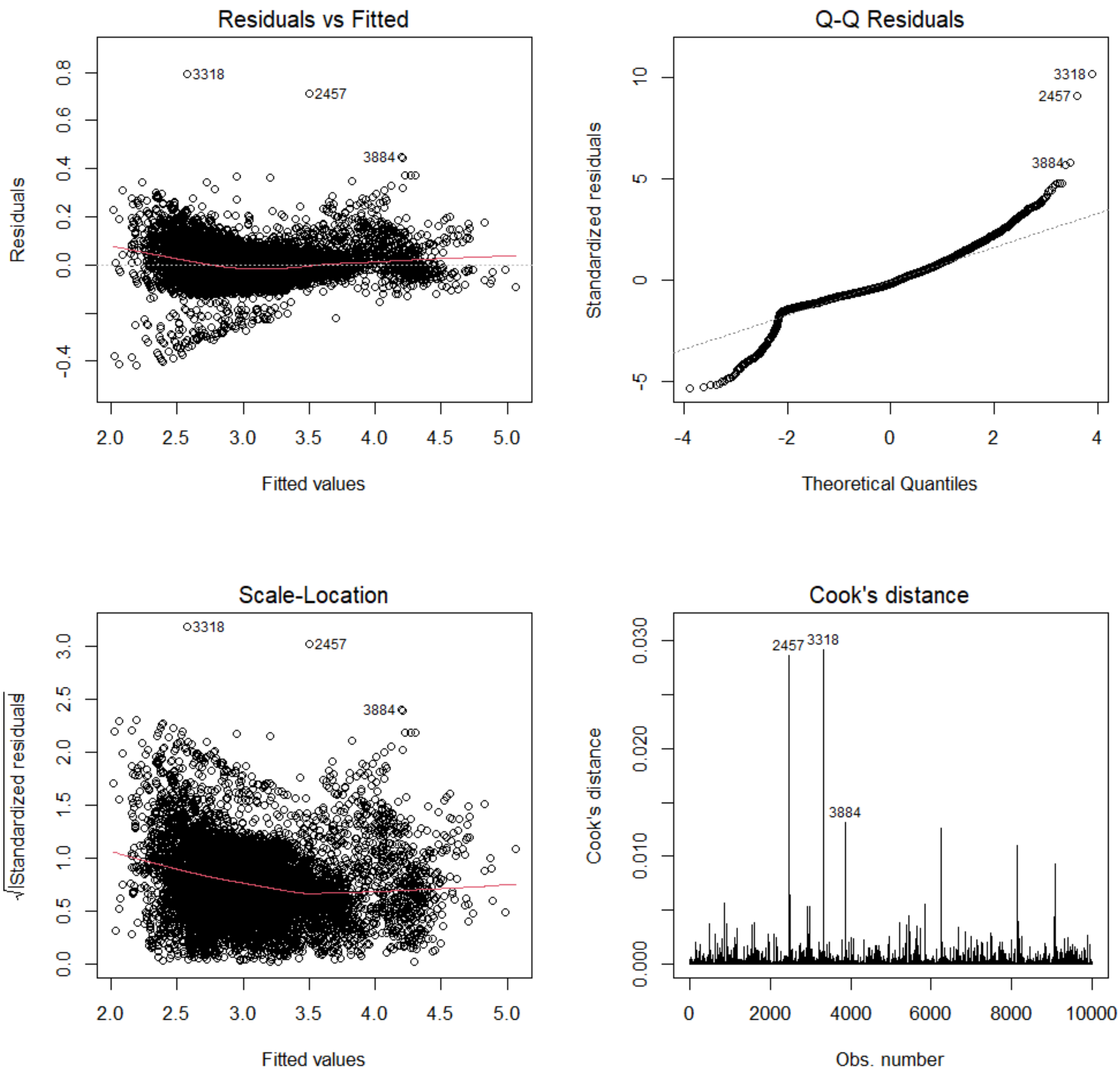

Figure 3: Diagnostics BIC Model

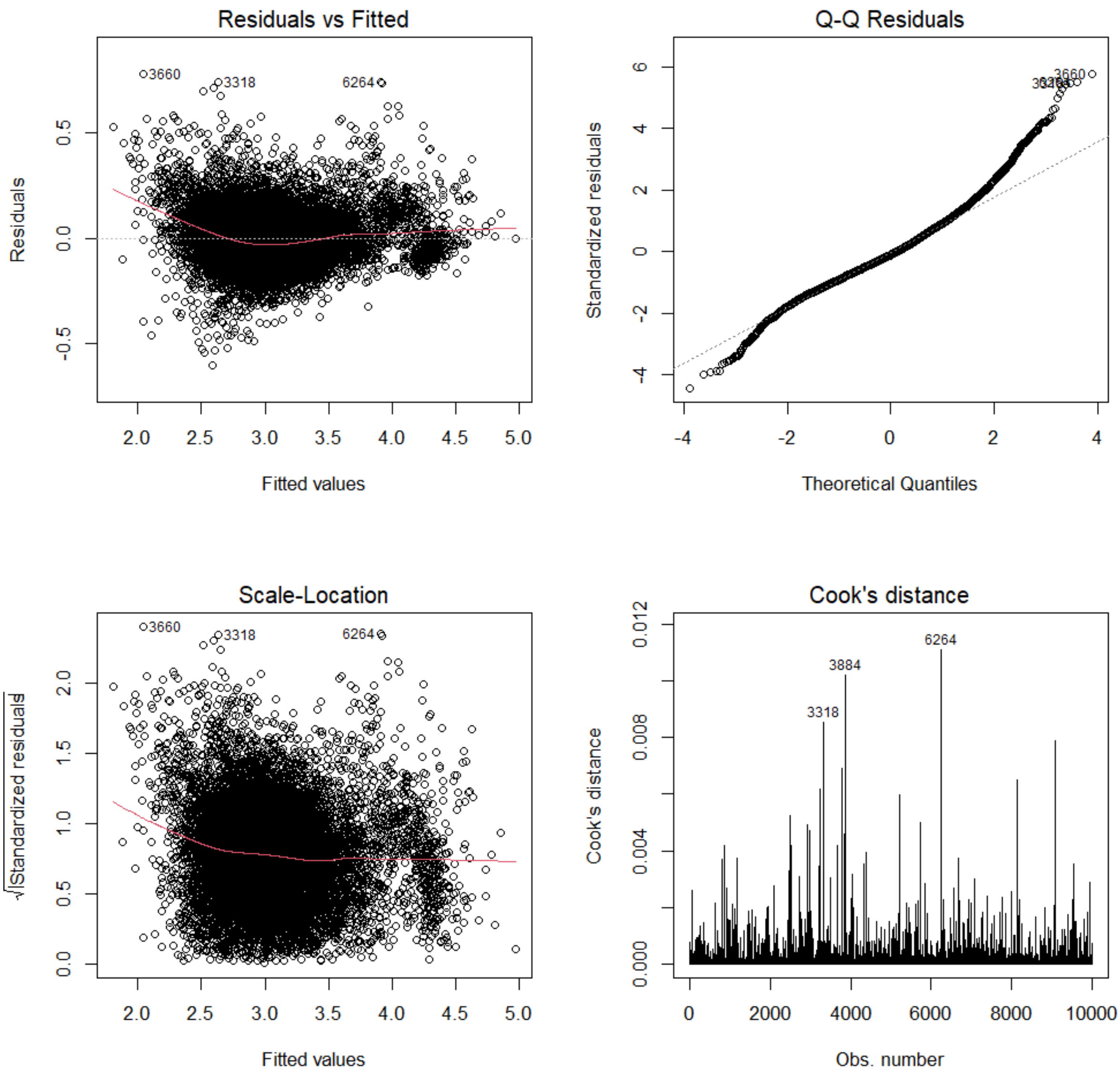

Figure 4: Diagnostics Manual Model

3 Challenges and Conclusions

The primary challenge was the trade-off between predictive accuracy and model interpretability.

1. **The Multicollinearity Problem:** The data clearly showed that `log_distance` and `log_fare` were highly collinear.

2. **The Failure of Domain Knowledge:** Our hypothesis that `log_fare` was redundant was proven wrong. Removing it destroyed the model's predictive power (Adj. R^2 dropped from 0.97 to 0.91). This means `log_fare` must contain critical information (e.g., base fees, surcharges) not captured by distance alone.

3. **The Final Trade-off:** We were forced to choose between an accurate but uninterpretable model (BIC) and an equally accurate but stable "black box" model (PCA).

Ultimately, there is no single "best" model.

- For pure prediction, the **BIC model** is sufficient.
- For a stable production model, the **PCA model** is the most reliable choice, as it is mathematically stable and equally accurate.
- We failed to find a model that was both highly accurate and interpretable.

3.1 Question for Peer Feedback

Given that the BIC model had the highest predictive accuracy, but the VIF scores for `log_distance` and `log_fare` were over 10: **Is it ever acceptable to keep highly collinear variables in a model if the model's sole purpose is prediction, not interpretation?**

3.2 Answer to Exam-Style Question

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 on the purpose. BIC retained the variables because both provided a significant improvement in *fit* that outweighed 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 INTERPRETATION**. We cannot trust the model to explain the *individual effect* of either variable.