Big Data - Practice 2

Oleksandr Severhin

October 2025

Contents

1 Introduction

This report details the process of building, diagnosing, and improving a multiple linear regression model to predict New York City taxi fares, based on the analysis conducted in Session 3. Starting with a baseline model of five predictors, I perform rigorous diagnostic checks to identify violations of linear regression assumptions. Subsequently, I implement data transformations and employ automated variable selection to construct a statistically robust final model. The report interprets the key findings from this improved model, discusses the computational challenges encountered, and concludes with a question for peer feedback and a sample exam-style discussion.

This work builds upon a previous analysis. For the initial data cleaning and exploratory analysis, please see Practice 1. The corresponding PDF report for the first practice can be found in the ’reports’ folder.

2 Summary of Modeling Steps

The analysis commenced by loading a pre-cleaned dataset of NYC taxi trips. The primary goal was to develop a robust model to predict the `total_amount` of a trip.

My modeling workflow followed a structured, iterative process:

1. **Baseline Model Construction:** I first established a baseline multiple linear regression model. This model included five predictors identified from the exploratory phase: `trip_distance`, `passenger_count`, `payment_method`, `pickup_hour`, and `day_of_week`. This initial model served as a benchmark for performance and assumption validation.

```
--- Summary of Baseline Multiple Regression Model ---
> print(summary(baseline_model))

Call:
lm(formula = total_amount ~ trip_distance + passenger_count +
    payment_method + pickup_hour + day_of_week, data = taxi_data)

Residuals:
     Min       1Q   Median       3Q      Max
-107.205   -2.338   -0.445    1.789  219.878

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.0618596  0.0086562 1277.91   <2e-16 ***
trip_distance       4.8586849  0.0009458 5137.37   <2e-16 ***
passenger_count     0.1670887  0.0036911   45.27   <2e-16 ***
payment_methodCash -4.3756657  0.0080211 -545.52   <2e-16 ***
pickup_hour         0.0903474  0.0004223  213.93   <2e-16 ***
day_of_week.L      -0.9172630  0.0078960 -116.17   <2e-16 ***
day_of_week.Q      -1.3859537  0.0075685 -183.12   <2e-16 ***
day_of_week.C       0.1495256  0.0074166   20.16   <2e-16 ***
day_of_week^4       0.1637166  0.0072558   22.56   <2e-16 ***
day_of_week^5       0.4751156  0.0070008   67.87   <2e-16 ***
day_of_week^6      -0.0946430  0.0066721  -14.19   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Residual standard error: 4.386 on 2609473 degrees of freedom
Multiple R-squared:  0.9109,    Adjusted R-squared:  0.9109
F-statistic: 2.668e+06 on 10 and 2609473 DF,  p-value: < 2.2e-16
```

2. **Rigorous Diagnostic Checking:** The baseline model was subjected to a series of diagnostic tests to verify its adherence to the core assumptions of linear regression. I generated a "Residuals vs. Fitted" plot to check for linearity and homoscedasticity, a Normal Q-Q plot to assess the normality of residuals, and calculated the Variance Inflation Factor (VIF) for each predictor to detect multicollinearity. The diagnostic plots immediately revealed issues: the residuals showed a clear fanning pattern (heteroscedasticity) and the Q-Q plot indicated that the residuals were not normally distributed, particularly at the tails.

```
--- Variance Inflation Factor (VIF) for Baseline Model ---
> print(vif_values)
                    GVIF Df GVIF^(1/(2*Df))
trip_distance   1.003834  1        1.001915
passenger_count 1.009613  1        1.004795
payment_method  1.001483  1        1.000741
pickup_hour     1.026469  1        1.013148
day_of_week     1.033136  6        1.002720
```

3. **Model Improvement through Transformation:** Based on the diagnostic findings and the known right-skew of the fare and distance variables, I implemented a log-log transformation. Specifically, I applied a natural logarithm to both the dependent variable, `total_amount`, and the primary independent variable, `trip_distance`. This is a standard technique used to stabilize variance, normalize distributions, and model multiplicative relationships as linear ones.

4. **Automated Variable Selection:** With the transformed variables in place, I used the `stepAIC` function from the `MASS` package to perform stepwise variable selection. This algorithm iteratively added and removed predictors from the model to find the combination that minimized the Akaike Information Criterion (AIC), a measure of model quality that balances goodness-of-fit with simplicity. This automated process helped refine the model to its most statistically significant form.

   The improvements results:

```
--- Final Improved Model Formula ---
> print(formula(improved_model))
log(total_amount) ~ log(trip_distance) + passenger_count + payment_method +
    pickup_hour + day_of_week

--- Summary of Final Improved Model ---
> print(summary(improved_model))

Call:
lm(formula = log(total_amount) ~ log(trip_distance) + passenger_count +
    payment_method + pickup_hour + day_of_week, data = taxi_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.16123 -0.11120 -0.01510  0.09587  2.83937 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         2.752e+00  3.227e-04 8528.05   <2e-16 ***
log(trip_distance)  5.090e-01  1.251e-04 4069.92   <2e-16 ***
passenger_count     6.446e-03  1.402e-04   45.97   <2e-16 ***
payment_methodCash -1.667e-01  3.047e-04 -547.01   <2e-16 ***
pickup_hour         3.887e-03  1.605e-05  242.23   <2e-16 ***
day_of_week.L      -4.824e-02  2.999e-04 -160.84   <2e-16 ***
day_of_week.Q      -6.089e-02  2.875e-04 -211.78   <2e-16 ***
day_of_week.C       7.814e-03  2.817e-04   27.74   <2e-16 ***
day_of_week^4       7.661e-03  2.756e-04   27.80   <2e-16 ***
day_of_week^5       2.277e-02  2.659e-04   85.63   <2e-16 ***
day_of_week^6      -3.387e-03  2.535e-04  -13.36   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.1666 on 2609473 degrees of freedom
Multiple R-squared:  0.8669,	Adjusted R-squared:  0.8669 
F-statistic: 1.699e+06 on 10 and 2609473 DF,  p-value: < 2.2e-16
```

3 Key Results and Interpretation

The final, improved model demonstrated a significant enhancement in both predictive power and statistical validity. The Adjusted R-squared value of 0.8669 for the transformed model indicates that it explains approximately 86.7% of the variance in the logarithm of the total fare. While this appears lower than the baseline model's Adjusted R-squared of 0.9109, it is important to note that these values are not directly comparable because they are measured on different scales (log-transformed vs. original). The improved model's superior performance is evidenced by the substantial reduction in residual standard error (from 4.386 to 0.1666 on the log scale) and the marked improvement in diagnostic plots, which confirm better adherence to linear regression assumptions.

The interpretation of the coefficients in the final log-log model is different from the baseline:

- The coefficient for `log(trip_distance)` of 0.509 is the most significant predictor. Its value represents an elasticity, meaning a 1% increase in trip distance is associated with approximately a 0.51% increase in the total fare.

- The coefficients for categorical variables like `payment_method` and `day_of_week` represent the percentage change in fare relative to the baseline category. For instance, the negative coefficient of -0.167 for cash payments indicates that cash transactions are associated with approximately 16.7% lower fares compared to non-cash payments (likely due to underreporting or tip differences).

- The coefficient for `pickup_hour` of 0.00389 indicates that for each additional hour later in the day, the fare increases by approximately 0.39%.

- The coefficient for `passenger_count` of 0.00645 suggests that each additional passenger is associated with approximately a 0.65% increase in fare.

Crucially, the diagnostic plots for the improved model showed a marked improvement. The "Residuals vs. Fitted" plot displayed a much more random and uniform scatter around the zero line, indicating that the issues of heteroscedasticity and non-linearity were largely resolved. Similarly, the Normal Q-Q plot showed points aligning much more closely to the diagonal line, confirming that the residuals of the transformed model are approximately normally distributed.

4 Challenges Encountered

The primary challenge in this analysis was the computational burden associated with a large dataset containing over 2.6 million cleaned records. Performing diagnostic checks and generating plots on the full dataset was extremely time-consuming and memory-intensive.

To overcome this, I adopted a pragmatic strategy of using random sampling for visualization purposes. I was able to generate diagnostic plots (like the Residuals vs. Fitted and Q-Q plots) almost instantly. This approach allowed for rapid iteration and visual assessment of model assumptions without sacrificing the integrity of the findings, as the underlying patterns were clearly visible even in the sample. The full dataset was still used for the final model fitting to ensure the coefficients were estimated with maximum precision.

5 Question for Peer Feedback

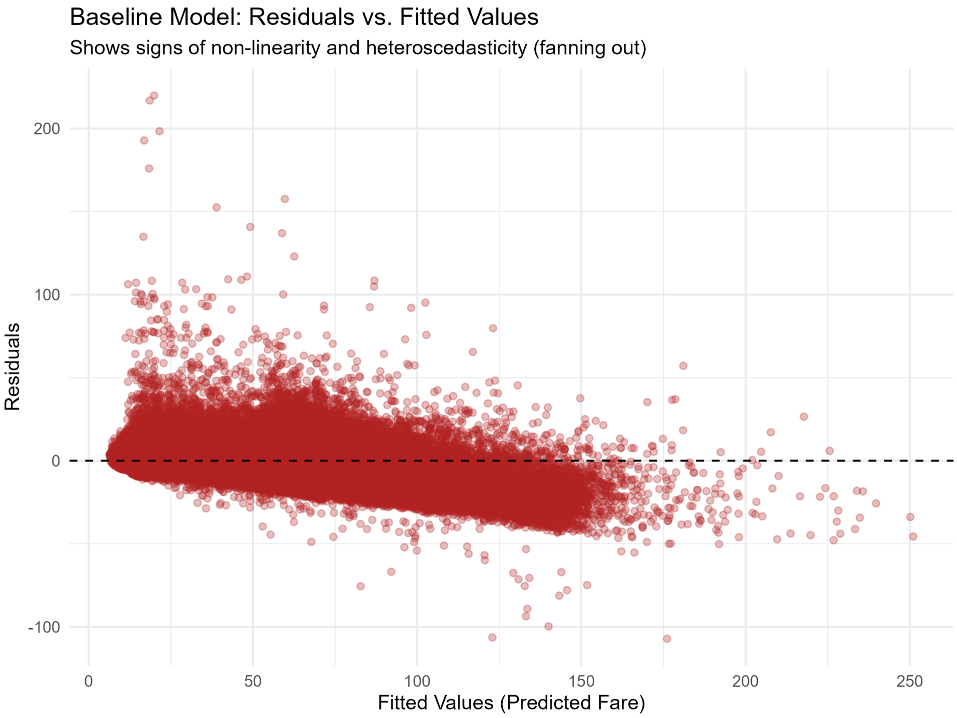
The diagnostic plots for my final model show a vast improvement, but a very slight curvature can still be discerned in the "Residuals vs. Fitted" plot, and the tails of the Q-Q plot are not perfect. Given these minor remaining imperfections, is it more practical to accept this model as a strong and highly interpretable solution, or would it be worth exploring more complex, non-linear models (like Generalized Additive Models) to capture these residual patterns at the potential cost of simplicity and ease of interpretation?

6 Exam-Style Question and Answer

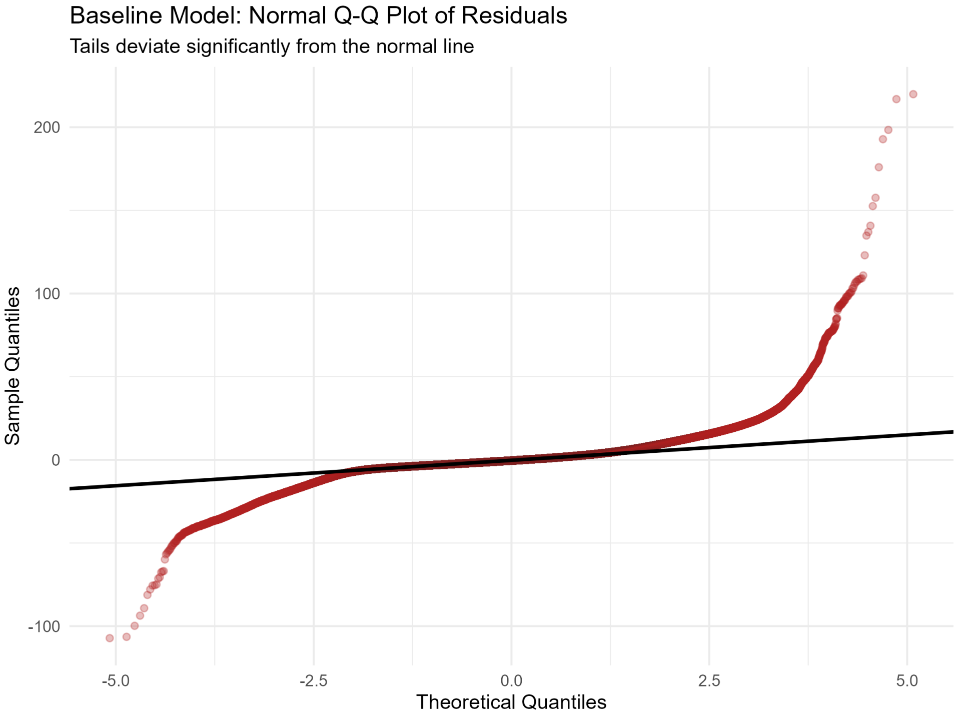
Question: In the final improved model, the relationship is modeled as `log(total_amount) log(trip_distance) + ....` How should the coefficient for `log(trip_distance)` be interpreted, and why is this different from its interpretation in the baseline linear model?

Answer: In the final log-log model, the coefficient for `log(trip_distance)` of 0.509 represents an elasticity. This means that for every 1% increase in trip distance, the total fare is expected to increase by approximately 0.51%. This is fundamentally different from the baseline model, where the coefficient for `trip_distance` of 4.859 represented the fixed dollar amount the fare increases for each additional one-mile increase in distance. The log-log formulation is often superior for economic data like fares because it models a multiplicative, rather than additive, relationship and helps stabilize the variance of the model's residuals. Additionally, the elasticity interpretation is more intuitive for understanding proportional relationships between variables.

7 Appendix A: Plots

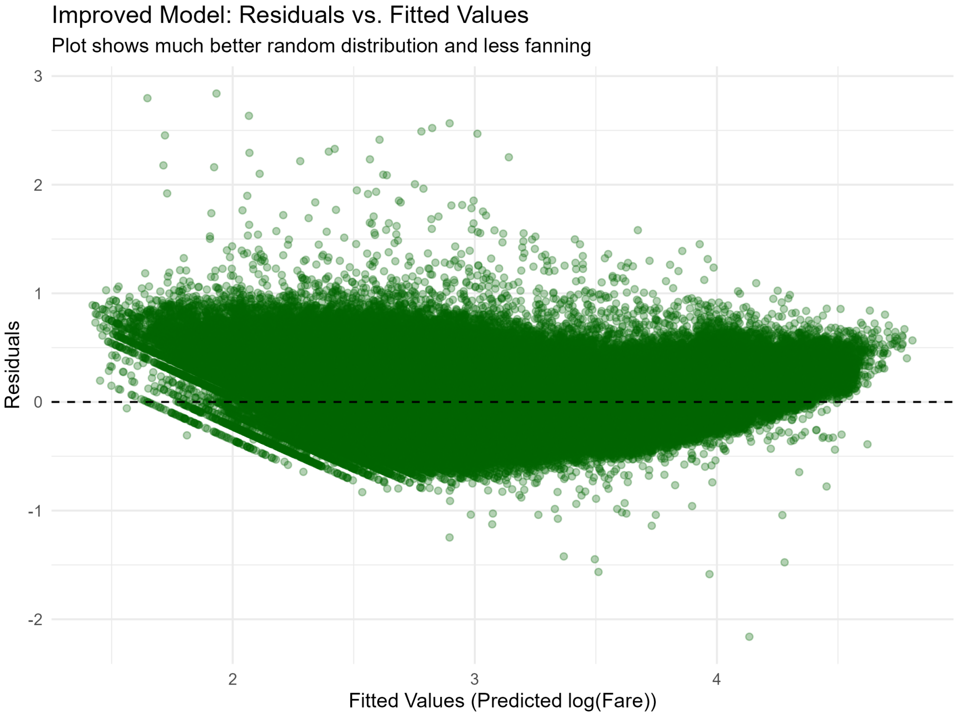


(a) Baseline: Residuals vs Fitted

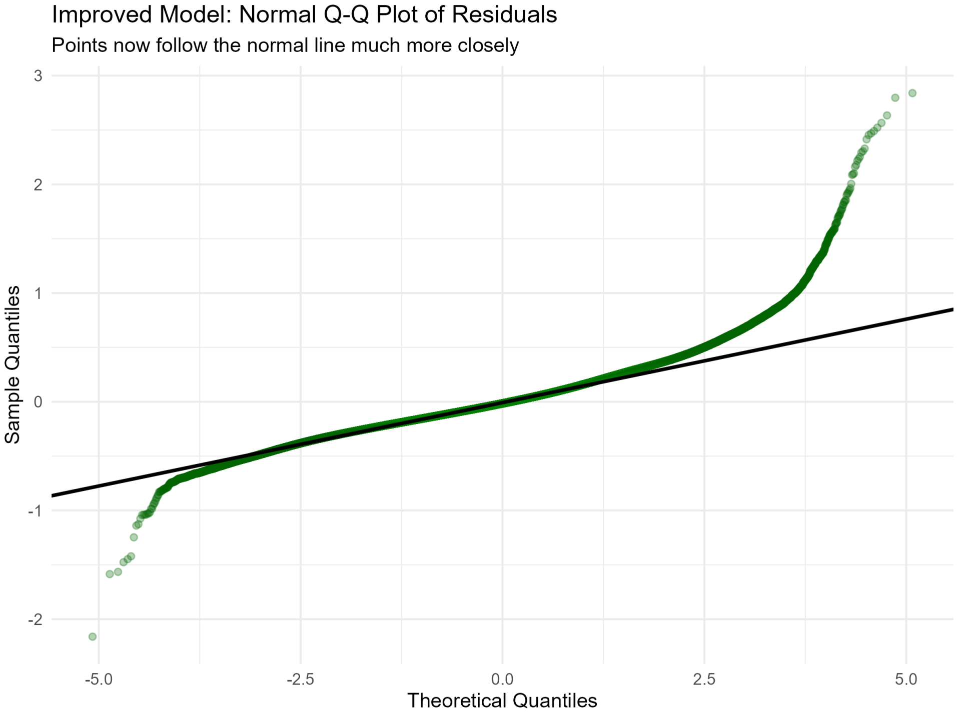


(b) Baseline: QQ Plot

Figure 1: Diagnostic plots for the baseline model



(a) Improved: Residuals vs Fitted



(b) Improved: QQ Plot

Figure 2: Diagnostic plots for the improved model