Big Data - Practice 03

yehorbolt

Table of contents

---

1. Introduction and Setup

This analysis refines the log-log regression model developed previously. The goal is to improve model stability and statistical validity using advanced techniques, including model selection based on the **Bayesian Information Criterion (BIC)**, in-depth diagnostics, and Principal Component Analysis (PCA) to address multicollinearity.

---

2. Initial "Full" Model and Multicollinearity

We start with a "full" model including all potentially relevant predictors to check for multicollinearity.

Table: VIF Scores for Full Model

	GVIF	Df	GVIF^(1/(2*Df))
log_distance	9.888568	1	3.144609
log_fare	10.264509	1	3.203827
log_tip	2.881900	1	1.697616
log_tolls	1.235349	1	1.111463
passenger_count	1.016157	1	1.008046
payment_type	2.411730	1	1.552974
pickup_hour	1.025404	1	1.012622
day_of_week	1.053559	6	1.004357
VendorID	1.008402	1	1.004192

Finding: The VIF scores for `log_distance` (**9.89**) and `log_fare` (**10.26**) confirm significant multicollinearity.

3. Model Improvement I: Stepwise Selection (BIC)

We use backward stepwise selection based on **BIC**.

```
# Determine n for BIC calculation
n_obs <- nrow(taxi_model_data)

# Run backward stepwise selection using BIC
step_model_bic <- stepAIC(full_model,
                          direction = "backward",
                          trace = FALSE,
                          k = log(n_obs))

# Show the summary of the model selected by BIC
summary(step_model_bic)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.48897 -0.04688 -0.01405  0.04115  2.51100

Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)           1.180e+00  5.976e-04 1974.523  < 2e-16 ***
log_distance          1.078e-02  1.782e-04   60.524  < 2e-16 ***
log_fare              6.521e-01  2.661e-04 2450.805  < 2e-16 ***
log_tip               1.435e-01  1.179e-04 1216.836  < 2e-16 ***
```

```
log_tolls             1.202e-01  1.333e-04  901.152  < 2e-16 ***
passenger_count       9.935e-04  6.476e-05   15.341  < 2e-16 ***
payment_type2         2.745e-02  2.176e-04  126.157  < 2e-16 ***
pickup_hour           2.442e-03  8.108e-06  301.145  < 2e-16 ***
day_of_weekTuesday    6.399e-03  1.964e-04   32.574  < 2e-16 ***
day_of_weekWednesday  1.129e-03  1.876e-04    6.021 1.73e-09 ***
day_of_weekThursday   2.984e-03  1.857e-04   16.065  < 2e-16 ***
day_of_weekFriday     3.505e-03  1.876e-04   18.684  < 2e-16 ***
day_of_weekSaturday  -2.720e-02  1.947e-04 -139.739  < 2e-16 ***
day_of_weekSunday    -1.668e-02  2.044e-04  -81.622  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07705 on 2617548 degrees of freedom
Multiple R-squared:  0.9717,    Adjusted R-squared:  0.9717
F-statistic: 6.92e+06 on 13 and 2617548 DF,  p-value: < 2.2e-16
```

Interpretation: Stepwise selection using BIC removed `VendorID`. The resulting `step_model_bic` is simpler but **still retains both `log_distance` and `log_fare`**, meaning multicollinearity persists.

---

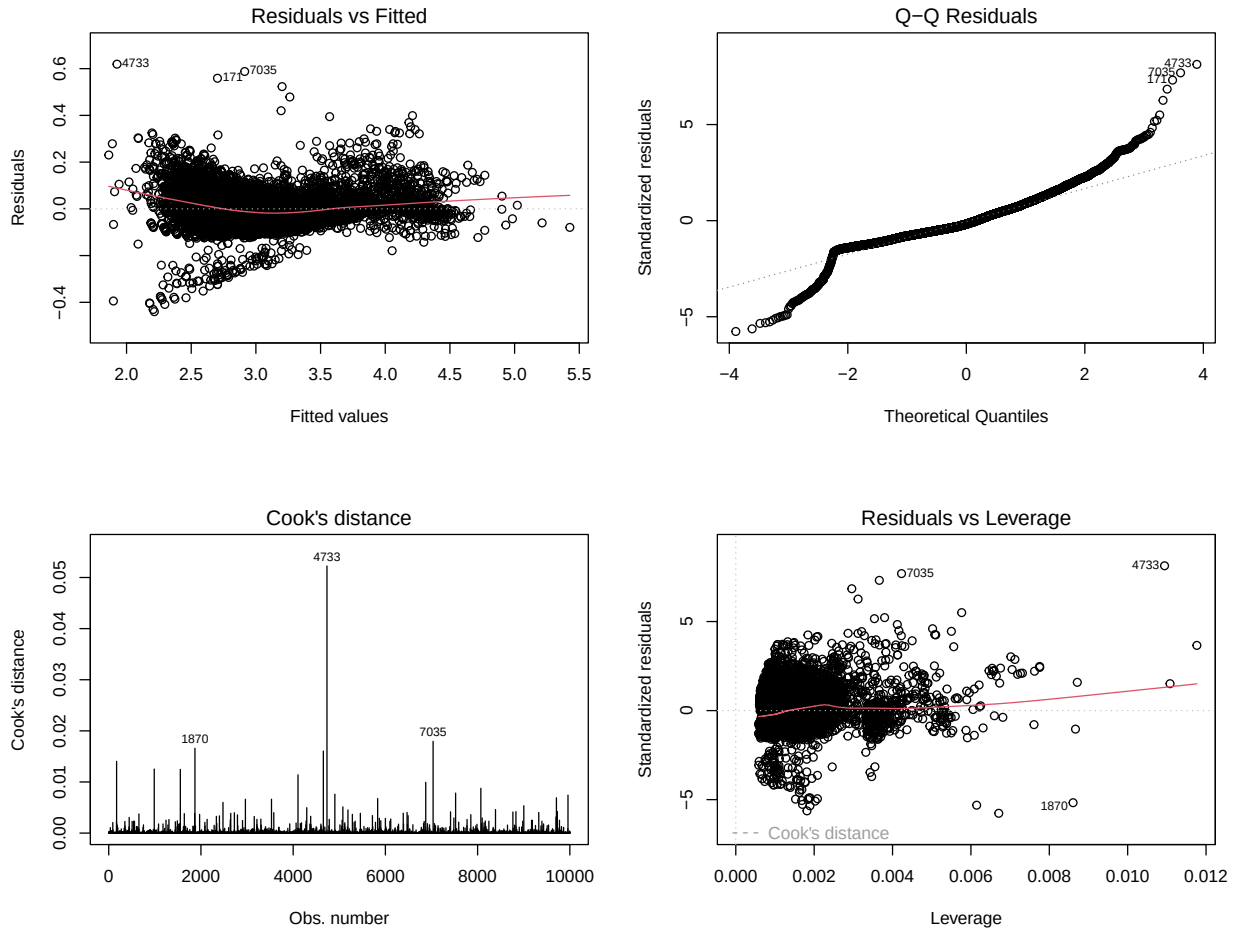
4. Advanced Diagnostics on the BIC-Selected Model

Diagnostics are run on a sample for efficiency.

```r
# Create a sample for fast diagnostics
set.seed(123)
diag_sample_bic <- taxi_model_data %>% slice_sample(n = 10000)

# Re-fit the BIC-selected model formula on the sample
step_model_diag_bic <- lm(formula(step_model_bic), data = diag_sample_bic)
```

```r
# --- Plot diagnostics ---
par(mfrow = c(2, 2))
plot(step_model_diag_bic, which = 1) # Residuals vs Fitted
plot(step_model_diag_bic, which = 2) # Normal Q-Q
plot(step_model_diag_bic, which = 4) # Cook's Distance
plot(step_model_diag_bic, which = 5) # Residuals vs Leverage
```




```r
par(mfrow = c(1, 1))
```

Diagnostic Interpretation: The plots are generally reasonable, with no extreme outliers strongly influencing the BIC model after log transformations.

---

5. Model Improvement II: Principal Component Analysis (PCA)

PCA is applied to the 4 correlated numeric variables.

Interpretation: PCA transforms the correlated variables. PC1 captures ~61.7% of the original numeric variance. We create two PCA models for comparison.

---

6. Final Model Comparison (BIC vs. PCA)

This table compares the final candidate models. The `step_model_bic` is compared against the two PCA variants.

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

Model	# Predictors	Adj. R^2	RSE	F-statistic	F p-value	AIC	BIC
Stepwise (BIC)	13	0.9717	0.0770	6919683	< 0.001	-5990895	-5990703
PCA (Full)	14	0.9717	0.0770	6425453	< 0.001	-5990908	-5990703
PCA (Reduced)	12	0.9554	0.0968	4673070	< 0.001	-4798183	-4798005

Summary of Results and Interpretation

Both the `Stepwise (BIC)` model and the `PCA (Full)` model achieved nearly identical high Adjusted R-squared values (~0.9717) and very similar BIC scores.

However, they differ in usability: * **Stepwise (BIC) Model:** Retains original variables, making it *partially* interpretable. But the coefficients for `log_distance` and `log_fare` are unreliable due to high VIFs. * **PCA Model:** Successfully eliminates multicollinearity, but its coefficients (e.g., `PC1`) are blends of the original variables, making them difficult to explain in a direct, real-world sense.

The `PCA (Reduced)` model is clearly inferior, with a significantly lower Adj. R^2. The choice between the BIC and PCA (Full) models depends on the project goal: predictive accuracy (both are equal) vs. interpretability (both are flawed in different ways).

7. Challenges Encountered

The primary challenge was the **strong multicollinearity** between `log_distance` and `log_fare`. This forced a trade-off. Stepwise selection (BIC) prioritized statistical fit over coefficient stability, keeping both correlated variables. PCA resolved the multicollinearity but at the cost of interpretability. A practical challenge was running diagnostics on the full dataset, which was overcome by using a representative random sample for plotting.

8. Question for Peer Feedback

Given that the Stepwise (BIC) model retained highly collinear predictors but achieved the best BIC score, while PCA resolved multicollinearity but sacrificed interpretability: **In a scenario where explaining the *individual impact* of distance vs. base fare is important, would it be statistically justifiable to manually remove `log_fare` (even though BIC suggests keeping it) to create a third, more interpretable model, potentially sacrificing a tiny amount of predictive accuracy?**

9. Exam-Style Question and Answer

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

Answer: The analyst should be cautious about accepting this model, especially if interpreting individual coefficient effects is important. While BIC selected this model as optimal based on balancing fit and complexity (in-sample goodness-of-fit penalized by model size), the high VIF scores mean the coefficients for those two variables are unstable and have inflated standard errors. This makes their individual effects unreliable and potentially misleading, even if the overall model's predictive power is strong.