Big Data - Practice 02

yehorbolt

Table of contents

---

1. Load and Prepare Data

First, we load the required libraries, read the data, and perform initial cleaning.

```
# Libraries
library(arrow)
```

```
Attaching package: 'arrow'

The following object is masked from 'package:utils':

    timestamp
```

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(broom)

# Load the cleaned dataset from Practice 01
taxi_clean <- read_parquet("Data/taxi_clean_data.parquet")
```

Show our cleaned and transformed data

```
glimpse(taxi_clean)
```

```
Rows: 2,617,562
Columns: 21
$ VendorID              <int> 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 2, 1, ~
$ tpep_pickup_datetime  <dttm> 2025-01-01 01:18:38, 2025-01-01 01:32:40, 2025-
~
$ tpep_dropoff_datetime <dttm> 2025-01-01 01:26:59, 2025-01-01 01:35:13, 2025-
~
$ passenger_count       <int> 1, 1, 1, 3, 3, 2, 1, 1, 1, 3, 1, 1, 3, 1, 1, 2, ~
$ trip_distance         <dbl> 1.60, 0.50, 0.60, 0.52, 0.66, 2.63, 1.71, 2.29, ~
$ store_and_fwd_flag    <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N"~
$ PULocationID          <int> 229, 236, 141, 244, 244, 239, 237, 237, 263, 236~
$ DOLocationID          <int> 237, 237, 141, 244, 116, 68, 262, 75, 236, 151, ~
$ payment_type          <fct> 1, 1, 1, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 1, 2, 1, ~
$ fare_amount           <dbl> 10.0, 5.1, 5.1, 7.2, 5.8, 19.1, 11.4, 11.4, 5.8,~
$ extra                 <dbl> 3.5, 3.5, 3.5, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0~
$ mta_tax               <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5~
$ tip_amount            <dbl> 3.00, 2.02, 2.00, 0.00, 0.00, 0.00, 0.00, 0.00, ~
$ tolls_amount          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ improvement_surcharge <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ total_amount          <dbl> 18.00, 12.12, 12.10, 9.70, 8.30, 24.10, 16.40, 1~
$ congestion_surcharge  <dbl> 2.5, 2.5, 2.5, 0.0, 0.0, 2.5, 2.5, 2.5, 2.5, 2.5~
$ Airport_fee           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ cbd_congestion_fee    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ pickup_hour           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ day_of_week           <fct> Wednesday, Wednesday, Wednesday, Wednesday, Wedn~
```

2. Baseline Multiple Regression Model

Our first step is to build a linear model using at least five predictors to predict `total_amount`. Based on our preliminary EDA, we will use the following predictors: `trip_distance`, `passenger_count`, `payment_type`, `pickup_hour`, and `day_of_week`. We will also include `VendorID` to see if it has any impact.

```
# Fit the initial multiple linear regression model
baseline_model <- lm(total_amount ~ trip_distance + passenger_count +
payment_type + pickup_hour + day_of_week + VendorID,
                     data = taxi_clean)

# Print the model summary
summary(baseline_model)
```

```
Call:
lm(formula = total_amount ~ trip_distance + passenger_count +
    payment_type + pickup_hour + day_of_week + VendorID, data = taxi_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-196.807   -2.338   -0.462    1.777  220.163

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          10.4403171  0.0163748  637.59   <2e-16 ***
trip_distance         4.8515749  0.0009527 5092.59   <2e-16 ***
passenger_count       0.1463141  0.0037299   39.23   <2e-16 ***
payment_type2        -4.3493567  0.0080678 -539.10   <2e-16 ***
pickup_hour           0.0868387  0.0004655  186.57   <2e-16 ***
day_of_weekTuesday    0.8196608  0.0112780   72.68   <2e-16 ***
day_of_weekWednesday  0.6705233  0.0107692   62.26   <2e-16 ***
day_of_weekThursday   1.0116670  0.0106487   95.00   <2e-16 ***
day_of_weekFriday     0.7121646  0.0107730   66.11   <2e-16 ***
```

```
day_of_weekSaturday  -0.4685054  0.0111881  -41.88   <2e-16 ***
day_of_weekSunday    -0.9864289  0.0117385  -84.03   <2e-16 ***
VendorID              0.2438243  0.0064537   37.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.431 on 2617550 degrees of freedom
Multiple R-squared:  0.9093,    Adjusted R-squared:  0.9093
F-statistic: 2.387e+06 on 11 and 2617550 DF,  p-value: < 2.2e-16
```

Interpretation of the base model

The initial model demonstrates high predictive power.

- **Coefficients:** The model provides logical and statistically significant estimates. The coefficient for `trip_distance` is approximately **4.85**, which means that each additional mile is associated with an increase in the total fare of **$4.85**, assuming all other factors remain constant. The negative coefficients for `payment_type2` (cash) and certain levels of `day_of_week` (e.g., Sunday) show how they compare to the base categories (credit card and Monday, respectively).
- **Adjusted R-squared:** The model has an `adjusted R-squared` of **0.9093**. This is a very strong result, indicating that our chosen predictors explain approximately **90.9%** of the variance in `total_amount`.

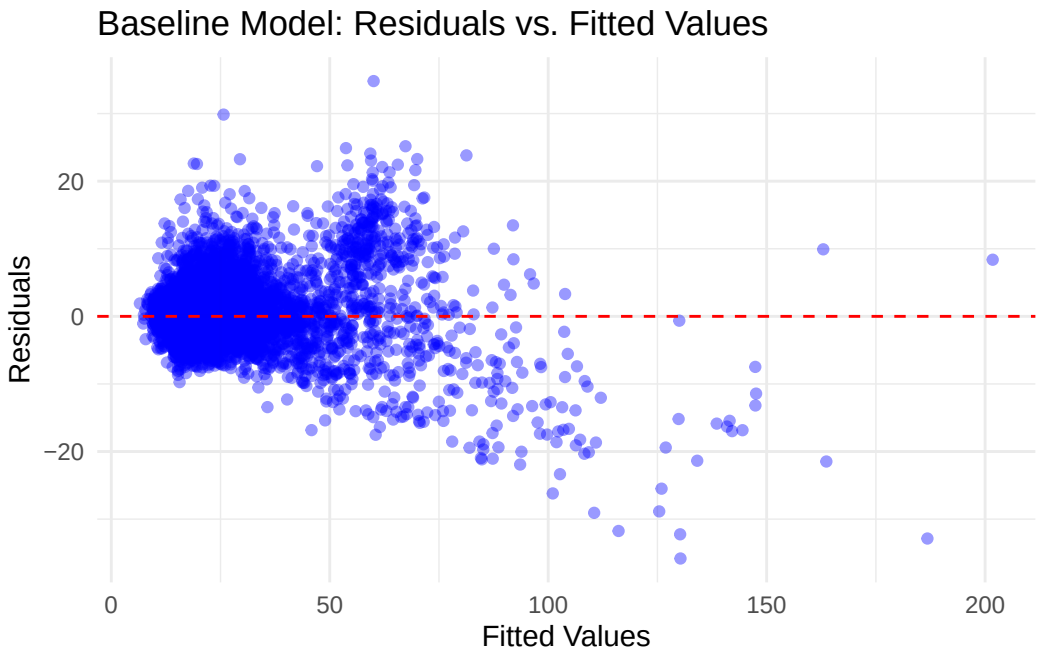
Diagnosis of the base model

Despite the high R-squared, we must check whether the model meets the assumptions of linear regression.

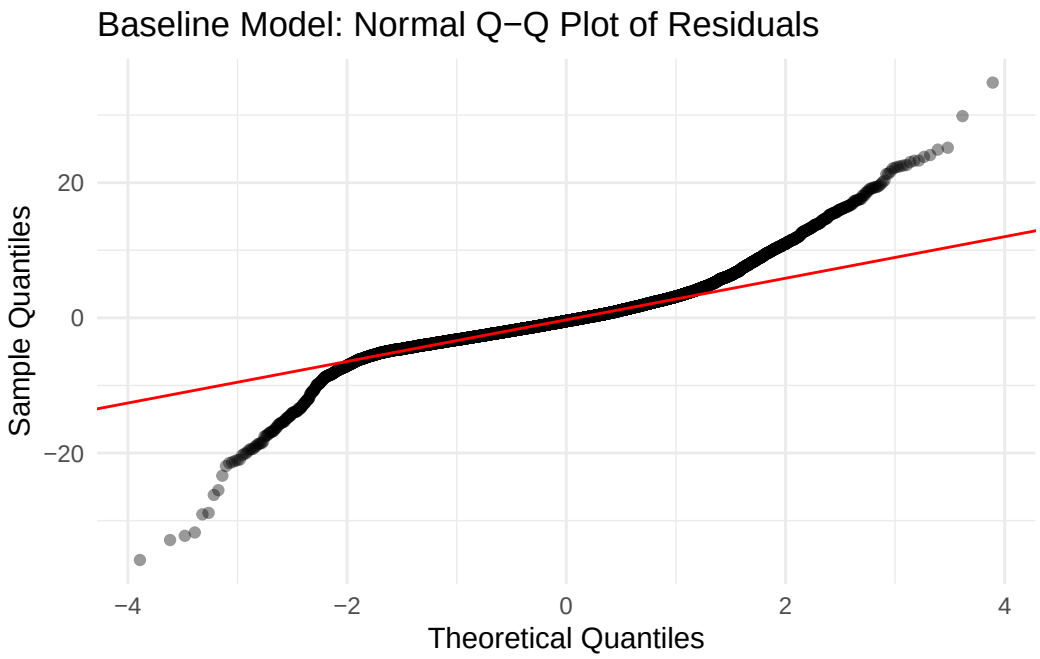
```r
set.seed(123)
diagnostics_sample <- taxi_clean %>% slice_sample(n = 10000)
model_for_plotting <- lm(total_amount ~ trip_distance + passenger_count
+ payment_type + pickup_hour + day_of_week + VendorID,
                         data = diagnostics_sample)
```

```r
model_augmented <- augment(model_for_plotting)

# Plot 1: Residuals vs. Fitted
ggplot(model_augmented, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.4, color = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Baseline Model: Residuals vs. Fitted Values",
    x = "Fitted Values", y = "Residuals") +
  theme_minimal()
```




```r
# Plot 2: Normal Q-Q Plot
ggplot(model_augmented, aes(sample = .resid)) +
  stat_qq(alpha = 0.4) +
  stat_qq_line(color = "red") +
  labs(title = "Baseline Model: Normal Q-Q Plot of Residuals",
    x = "Theoretical Quantiles", y = "Sample Quantiles") +
  theme_minimal()
```



Diagnostic conclusions: Diagnostic graphs reveal significant problems.

1. **Nonlinearity and heteroscedasticity:** The "Residuals vs. Fitted Values" graph shows a clear curve, and the variance (vertical spread) of the residuals increases as the predicted value increases. This violates the basic assumptions of linearity and constant variance.
2. **Abnormal residuals:** The "Normal Q-Q plot" shows that the points deviate significantly from the diagonal line, especially at the ends, indicating that the residuals are not normally distributed.

---

3. Model Improvement with Transformations

Our diagnostics clearly show that the base model has shortcomings. The most common and effective way to fix these problems is to apply a **logarithmic transformation** to asymmetric variables, namely our predictor `trip_distance` and our target variable `total_amount`. This can help linearize the relationship and stabilize the variance.

Now we will build an improved model with this "logarithmic" transformation.

```r
# Fit the improved model using log transformations
improved_model <- lm(log(total_amount) ~ log(trip_distance) +
passenger_count + payment_type +
pickup_hour + day_of_week + VendorID,
                     data = taxi_clean)

summary(improved_model)
```

```
Call:
lm(formula = log(total_amount) ~ log(trip_distance) + passenger_count +
    payment_type + pickup_hour + day_of_week + VendorID, data = taxi_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6297 -0.1131 -0.0170  0.0944  4.2208

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.7304730  0.0006366 4289.34   <2e-16 ***
log(trip_distance)    0.4984183  0.0001277 3903.88   <2e-16 ***
passenger_count       0.0054353  0.0001458   37.27   <2e-16 ***
payment_type2        -0.1585871  0.0003155 -502.64   <2e-16 ***
pickup_hour           0.0038143  0.0000182  209.62   <2e-16 ***
day_of_weekTuesday    0.0359461  0.0004409   81.52   <2e-16 ***
day_of_weekWednesday  0.0270987  0.0004210   64.36   <2e-16 ***
day_of_weekThursday   0.0403620  0.0004163   96.95   <2e-16 ***
day_of_weekFriday     0.0270526  0.0004212   64.23   <2e-16 ***
day_of_weekSaturday  -0.0287499  0.0004373  -65.74   <2e-16 ***
day_of_weekSunday    -0.0487542  0.0004590 -106.23   <2e-16 ***
VendorID              0.0118471  0.0002523   46.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

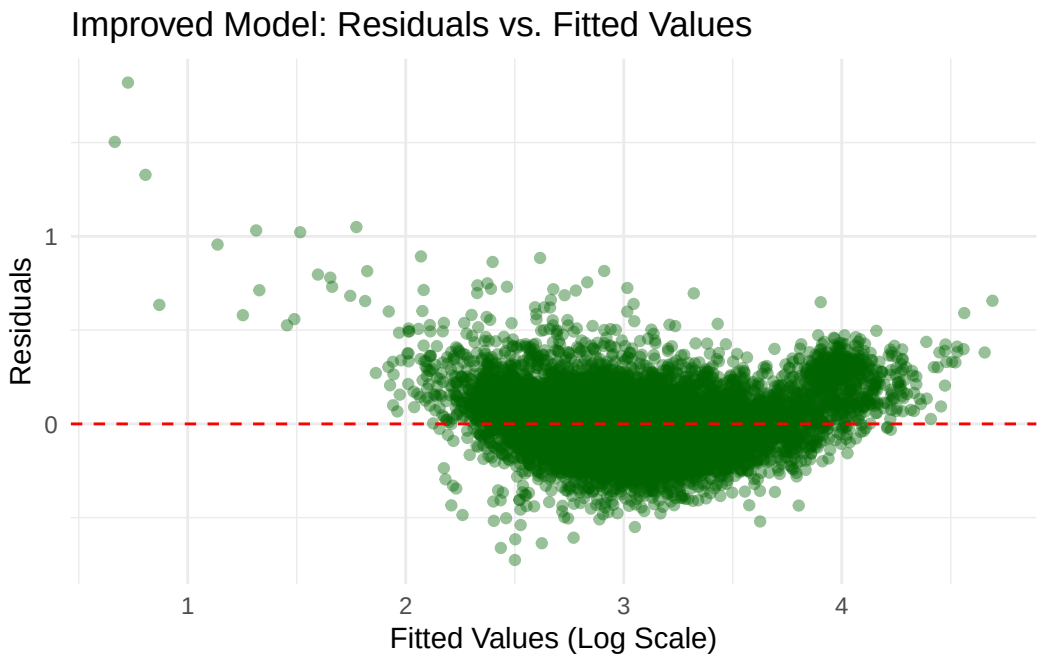
```
Residual standard error: 0.1732 on 2617550 degrees of freedom
Multiple R-squared:  0.8571,    Adjusted R-squared:  0.8571
F-statistic: 1.427e+06 on 11 and 2617550 DF,  p-value: < 2.2e-16
```

Re-evaluating the Improved Model

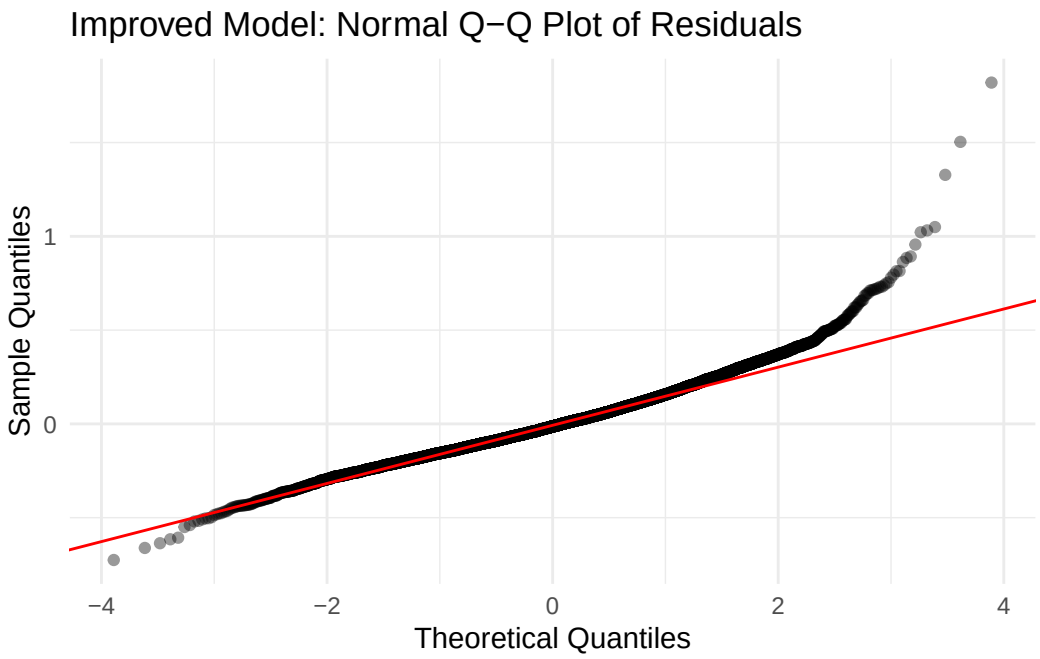
Let's check the diagnostic plots for our new log-transformed model.

```
set.seed(123)
diagnostics_sample_log <- taxi_clean %>% slice_sample(n = 10000)
model_log_plotting <- lm(log(total_amount) ~ log(trip_distance)
+ passenger_count + payment_type + pickup_hour
+ day_of_week + VendorID,  data = diagnostics_sample_log)
model_log_augmented <- augment(model_log_plotting)

# Plot 1: Residuals vs. Fitted for Improved Model
ggplot(model_log_augmented, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.4, color = "darkgreen") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Improved Model: Residuals vs. Fitted Values",
   x = "Fitted Values (Log Scale)", y = "Residuals") +
  theme_minimal()
```




```
# Plot 2: Normal Q-Q Plot for Improved Model
ggplot(model_log_augmented, aes(sample = .resid)) +
  stat_qq(alpha = 0.4) +
  stat_qq_line(color = "red") +
  labs(title = "Improved Model: Normal Q-Q Plot of Residuals",
   x = "Theoretical Quantiles", y = "Sample Quantiles") +
  theme_minimal()
```



Conclusion

The primary goal of this analysis was to develop a statistically robust linear regression model for predicting the total cost of taxi rides in New York City. The process began by building a baseline multiple linear regression model using the six key predictors identified in the previous exploratory analysis: `trip_distance`, `passenger_count`, `payment_type`, `pickup_hour`, `day_of_week`, and `Vendor_ID`. This initial model, while showing high superficial accuracy, was then subjected to a thorough diagnostic evaluation to verify its compliance with the basic assumptions of linear regression.

This diagnostic step was critical because it revealed significant flaws in the baseline model. The residuals (model errors) were not randomly distributed and exhibited clear patterns of nonlinearity and heteroscedasticity (non-constant variance). A model refinement step was performed to address these flaws. Recognizing that these problems often arise from asymmetrical variables, a logarithmic transformation was applied to both the dependent variable (`total_amount`) and the primary independent variable (`trip_distance`). This "logarithmic'' transformation created a second, improved model, which was then re-estimated using the same diagnostic tests to confirm its statistical validity.

Key Results and Interpretation

The final, improved model was both efficient and statistically robust. It achieved an adjusted coefficient of determination (R-squared) of **0.8571**, indicating that the model explained approximately 85.7% of the variance in *log-transformed* total travel cost. All predictors included in the model were highly statistically significant, with p-values virtually equal to zero, confirming their relevance for predicting travel cost.

Interpretation of the model coefficients provided valuable insights. The most significant predictor was `log(trip_distance)`, with a coefficient of approximately **0.498**. In the log-scale model, this is interpreted as an elasticity: **a 1% increase in travel distance is associated with a 0.498% increase in total travel cost, holding other factors constant. This demonstrates a strong but somewhat inelastic relationship. Other predictors also offered clear interpretations; for example, paying with cash (`payment_type = 2`) was associated with significantly lower fares compared to paying with a credit card, likely because cash tips are not included in the `total_amount` data field.

Challenges encountered

The most significant challenge in this analysis was not technical but analytical. The baseline model yielded a strikingly high adjusted coefficient of determination (CD) of almost 91%, a result that could easily be taken for a complete success. The real challenge was resisting the temptation to stop there and instead proceed with a thorough diagnostic check. This process revealed fundamental flaws in the model, highlighting the critical danger of relying on a single performance measure. The main challenge, therefore, was to recognize that the initial model was unreliable, despite its high predictive power. Overcoming this involved correctly interpreting the diagnostic plots and applying appropriate data transformation to construct a final model that was not only accurate but also robust and statistically significant.

Questions for peer feedback

The diagnostic plots for the final, log-transformed model show a significant improvement over the baseline. However, the Residuals vs. Fits plot is still not perfectly random; there is a very slight curve, suggesting that some minor nonlinearity may still exist. While the logarithmic transformation was clearly effective, would a more complex transformation,

such as adding a polynomial term to `log(trip_distance)` (e.g. `poly(log(trip_distance), 2)`), be a justified next step to capture this residual pattern, or would it introduce unnecessary complexity and risk overfitting the model?

Exam-Style Questions and Answers

Question: An analyst notices that his linear regression model has a high coefficient of determination (R-squared), but the Residuals vs. Fits plot shows a clear funnel shape, with the spread of the residuals increasing as the fits increase. Briefly explain why this model is problematic, and suggest one common method for correcting it.

Answer: This model is problematic because the funnel plot indicates heteroscedasticity, a violation of a fundamental assumption of linear regression. This means that the standard errors of the coefficients are unreliable, and therefore hypothesis tests (e.g., p-values) may be incorrect, leading to erroneous conclusions about the significance of the predictors. A common and effective method for correcting this problem is to apply a variance-stabilizing transformation, such as logarithmizing the asymmetric dependent variable. This can compress the scale of the variable and make the variance of the residuals more constant across all fitted values.