MIAMI HERBERT BUSINESS SCHOOL

REGRESSION ANALYSIS OF LYFT PRICE

Jiaqi Wang
Meng Xiao
Xin Rao

CONTENT

MIAMI HERBERT BUSINESS SCHOOL

- Overview
- Data Preparation
- Pre-Data Analysis
- Model Building
- Model Selection
- Model Interpretation
- Model Comparison
- Q&A

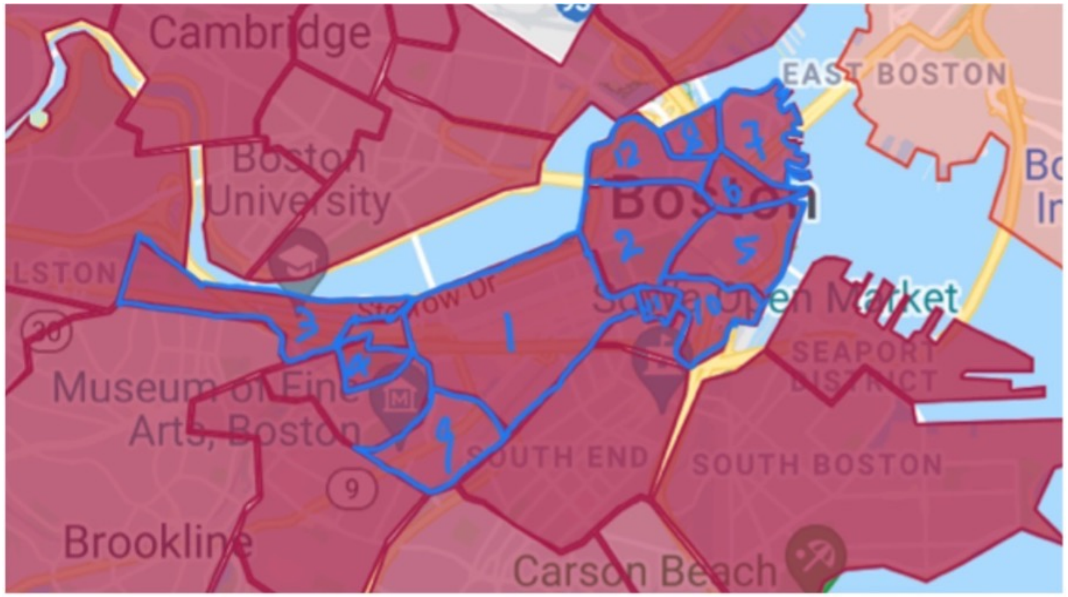
OVERVIEW--WHY DATASET

MIAMI HERBERT BUSINESS SCHOOL

- Lyft price, close to our life
- Boston, huge market, complete transportation system
- Sample size, big enough

OVERVIEW--BRIF INTRODUCTION

MIAMI HERBERT BUSINESS SCHOOL



- Kaggle
- 693071 observations
- 18 attributes
- Nov.26– Dec.18, 2018
- 12 Boston locations.

OVERVIEW--BRIF INTRODUCTION

MIAMI HERBERT BUSINESS SCHOOL

- Two datasets:
  - --price of Uber and Lyft, interval of 5 mins
  - --weather information, interval of 1 h
- Make predictions

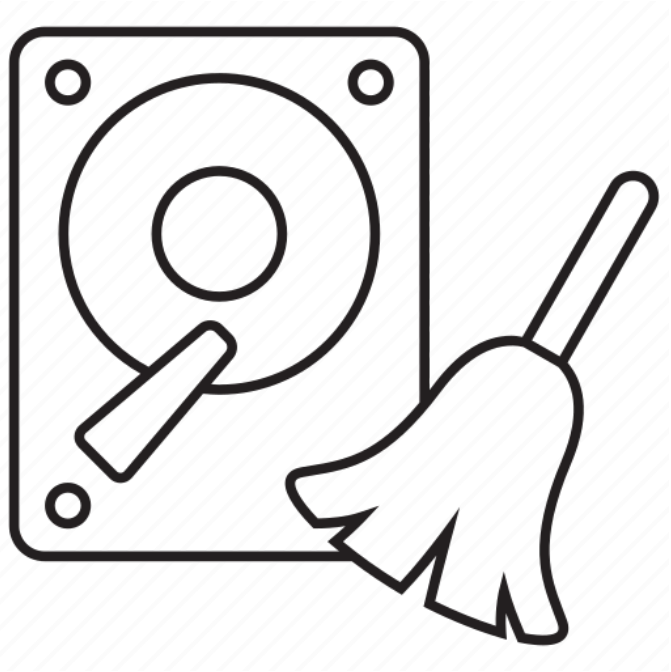
DATA PREPARATION

MIAMI HERBERT BUSINESS SCHOOL

- Subset Data
  - Lyft – 307,409 observations
  - Weather – 6,267 observations
- Combine two table
  - SQL – Cross Join

DATA PREPARATION

MIAMI HERBERT BUSINESS SCHOOL



- Time stamp
- Order time to “Time Range”
- “Rain” to “Is Rain”
- Order type based on Lyft App

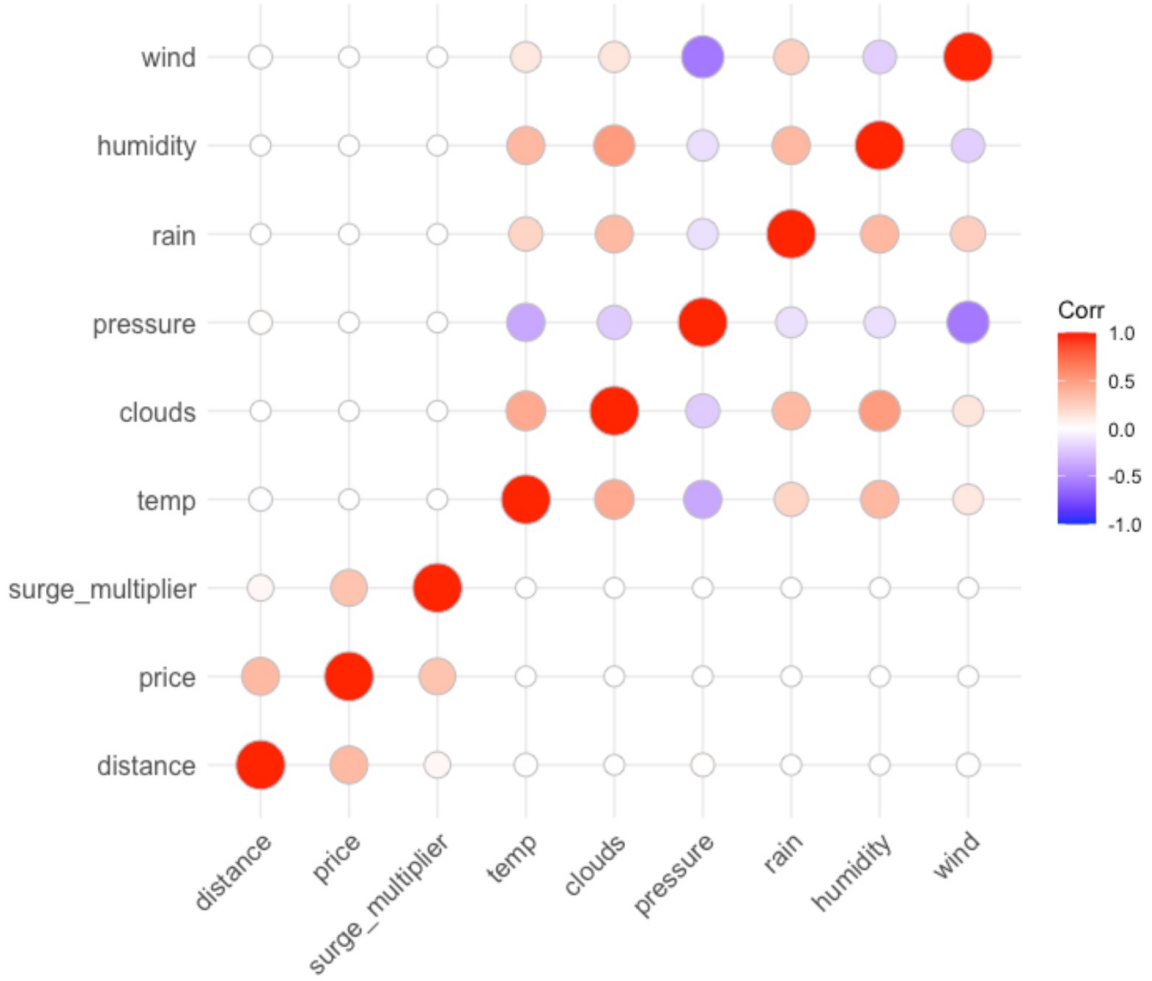
PRE-DATA ANALYSIS--ATTRIBUTES

MIAMI HERBERT BUSINESS SCHOOL

Attribute name	Type	Levels
Distance	Numeric	NA
Source	Categorical	12
Destination	Categorical	12
Order Type	Categorical	6
Surge Multiplier	Numeric	NA
Temp	Numeric	NA
Clouds	Numeric	NA
Pressure	Numeric	NA
Is Rain	Categorical	2
Humidity	Numeric	NA
Wind	Numeric	NA
Time Range	Categorical	8

PRE-DATA ANALYSIS--CORRELATION

MIAMI HERBERT BUSINESS SCHOOL



PRE-DATA ANALYSIS--RAW MODEL

MIAMI HERBERT BUSINESS SCHOOL

- Raw model:

- $$Price = \beta_0 + \beta_1 Source + \beta_2 Destination + \beta_3 Distance + \beta_4 Surge\_multiplier + \beta_5 Temp + \beta_6 Clouds + \beta_7 Pressure + \beta_8 Humidity + \beta_9 Wind + \beta_{10} Order\_type + \beta_{11} Is\_rain + \beta_{12} Time\_range$$

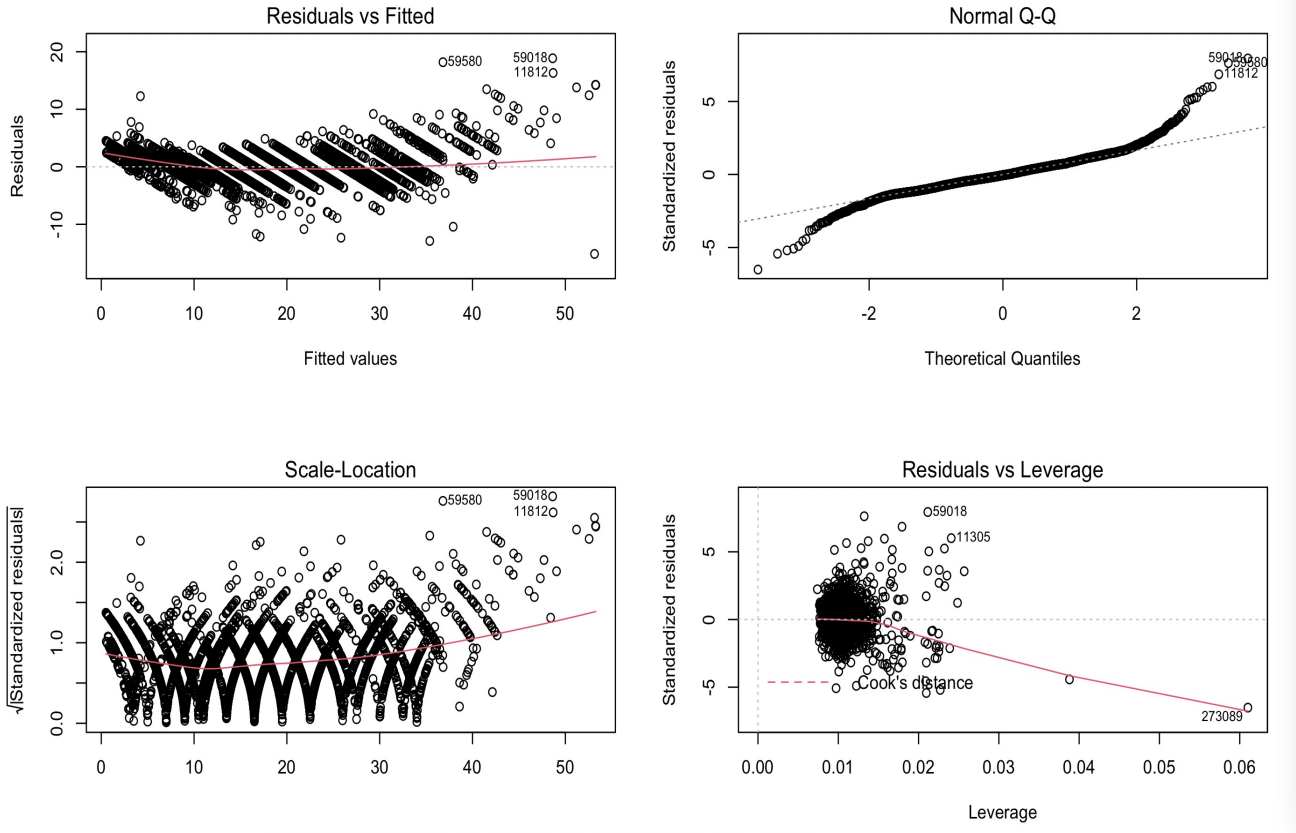
PRE-DATA ANALYSIS--SUMMARY

MIAMI HERBERT BUSINESS SCHOOL

FULL MODEL:

- $price = source + destination + distance + surge\_multiplier + temp + clouds + pressure + humidity + wind + order\_type + is\_rain + time\_range$

- Adjusted R^2: 0.9397



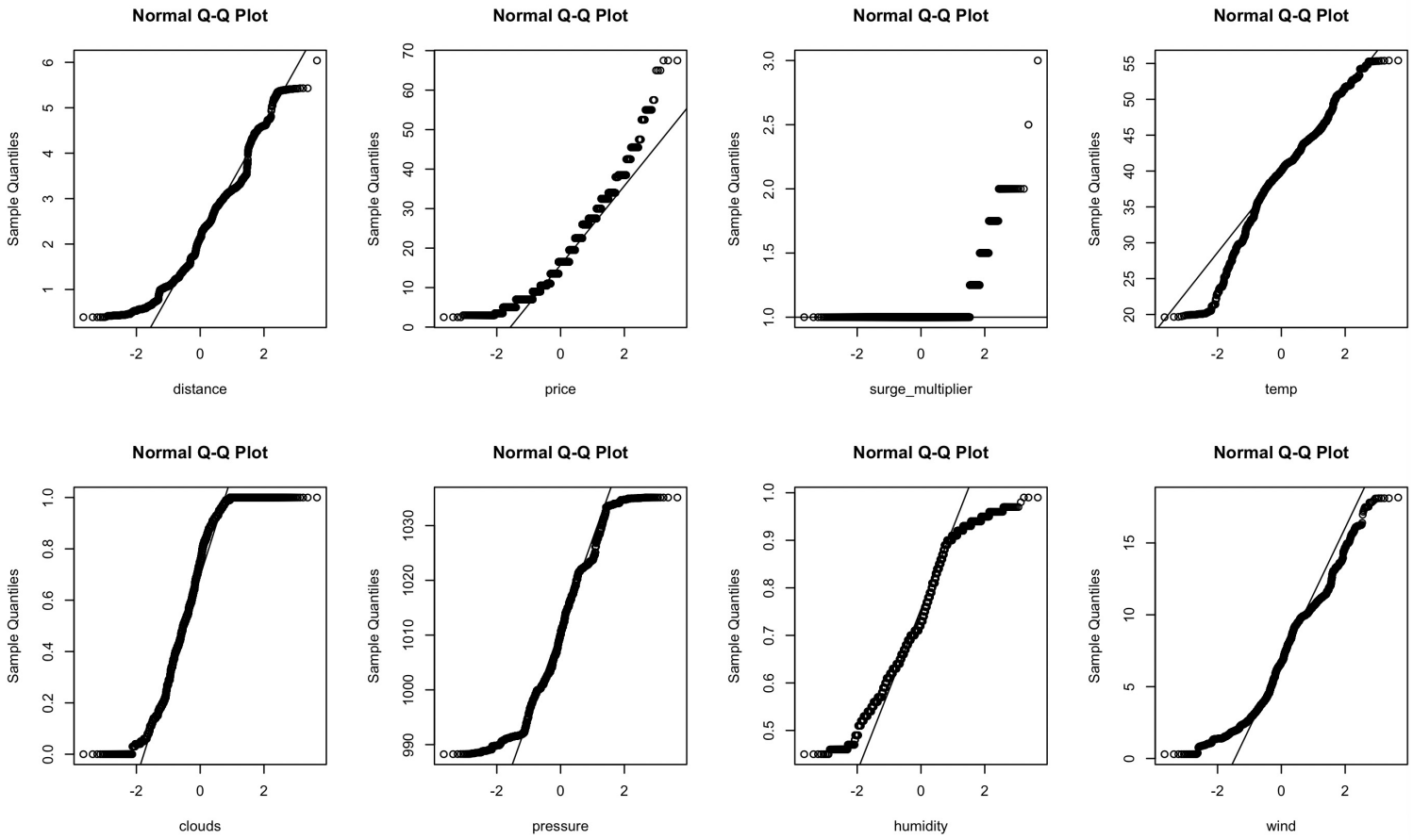

```
Shapiro-Wilk normality test

data:  residuals(lmod_full)
W = 0.94401, p-value < 2.2e-16
```

PRE-DATA ANALYSIS--SUMMARY

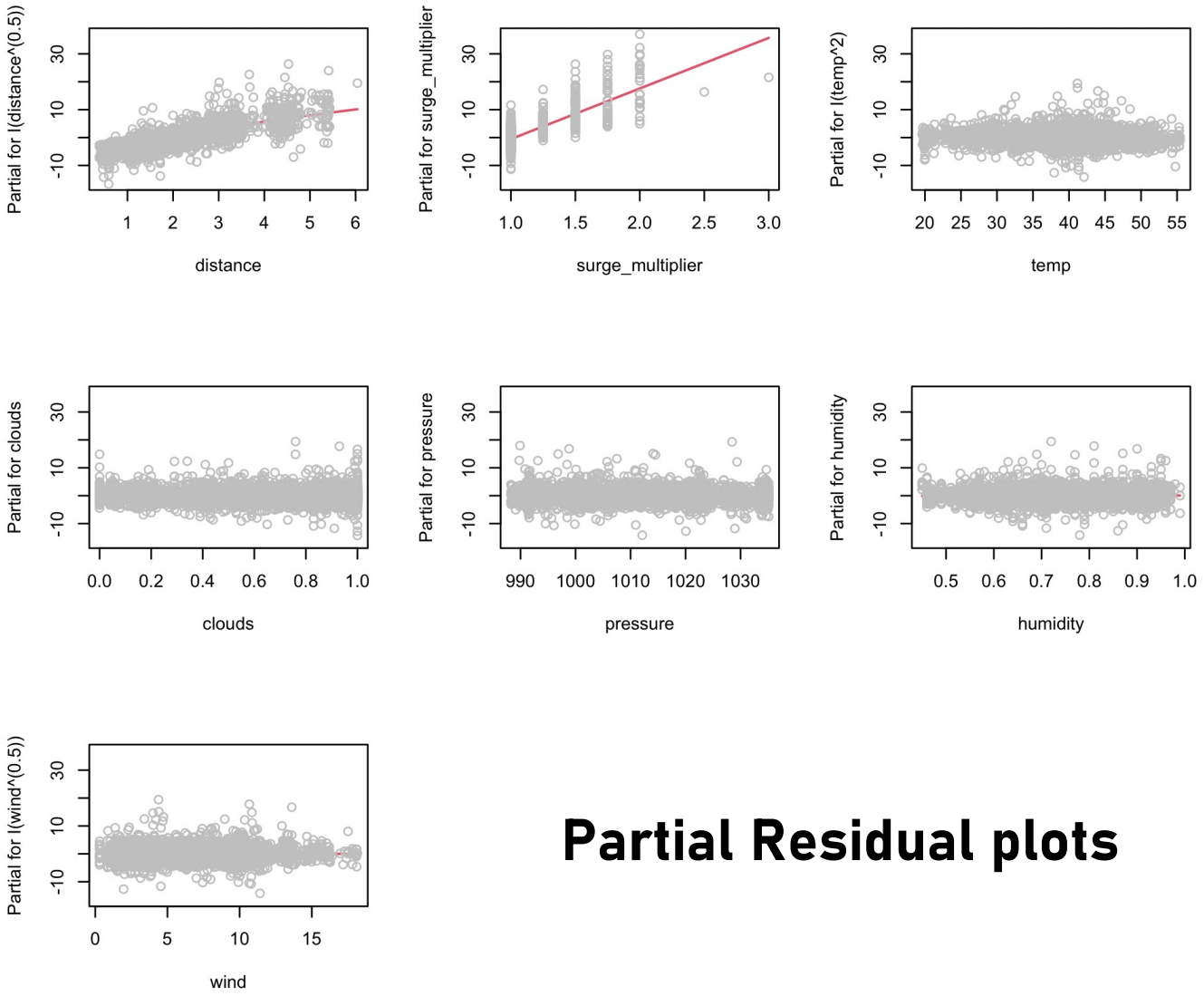
MIAMI HERBERT BUSINESS SCHOOL

Check normality of variables by QQ plot



PRE-DATA ANALYSIS--SUMMARY

MIAMI HERBERT BUSINESS SCHOOL

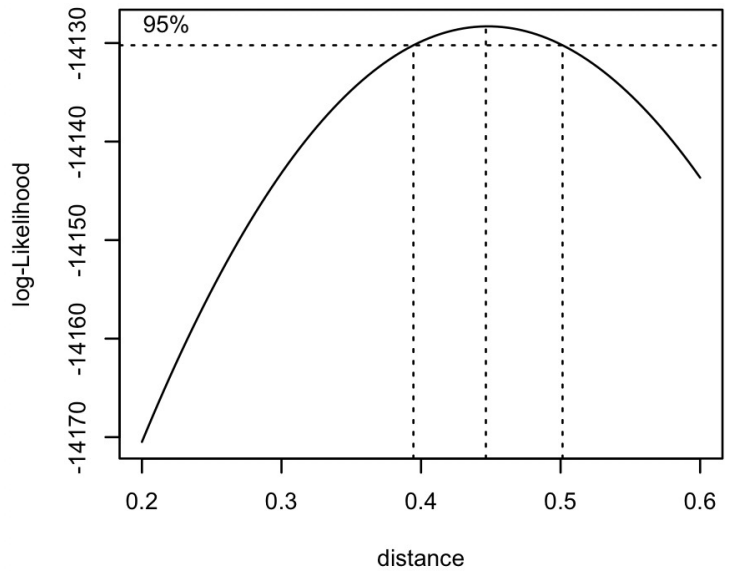


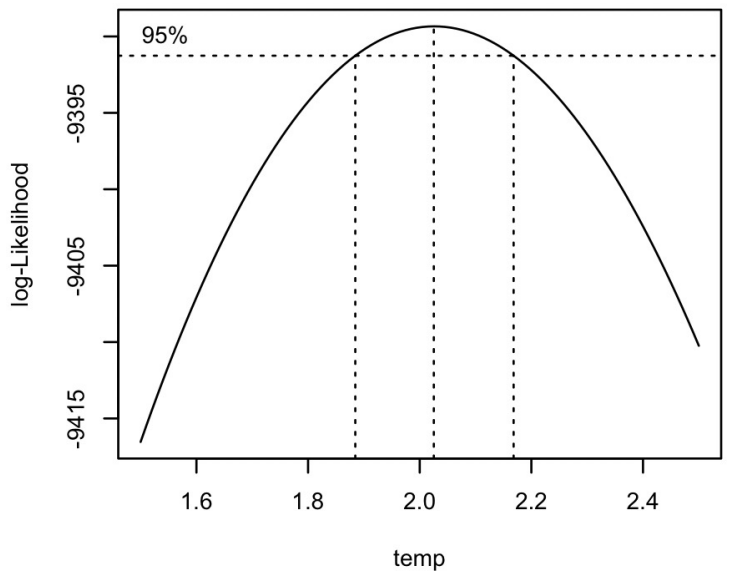
Partial Residual plots

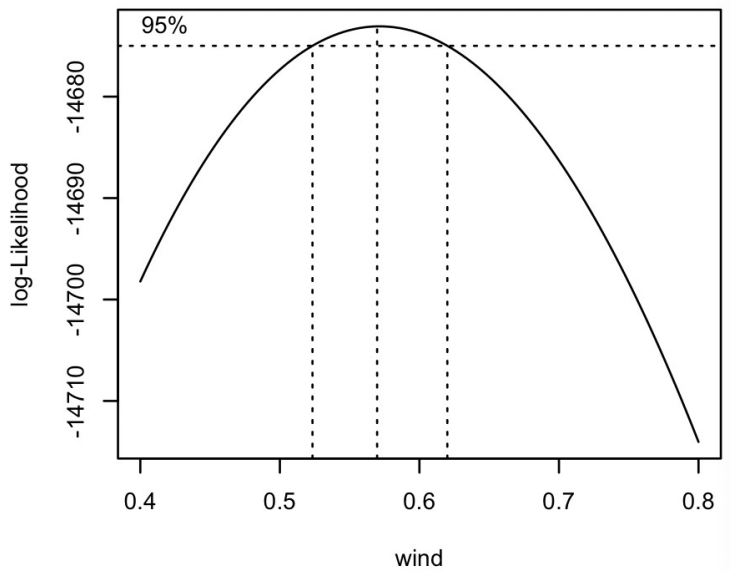
MODEL BUILDING

--PREDICTORS TRANSFORMATION

MIAMI HERBERT BUSINESS SCHOOL







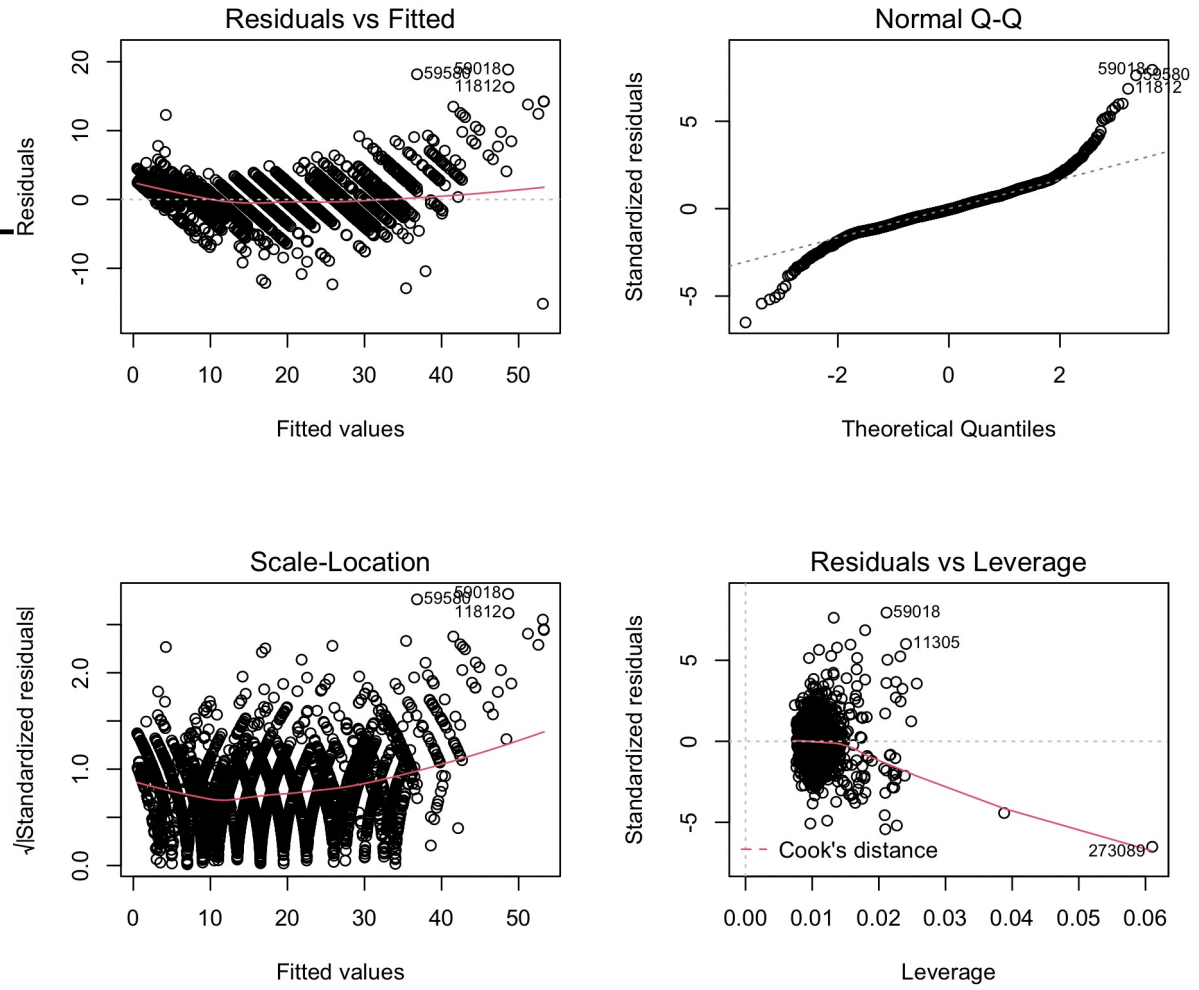
$distance$ to $\sqrt{distance}$ $temp$ to $temp^2$ $wind$ to $\sqrt{wind}$

MODEL BUILDING

--PREDICTORS TRANSFORMATION

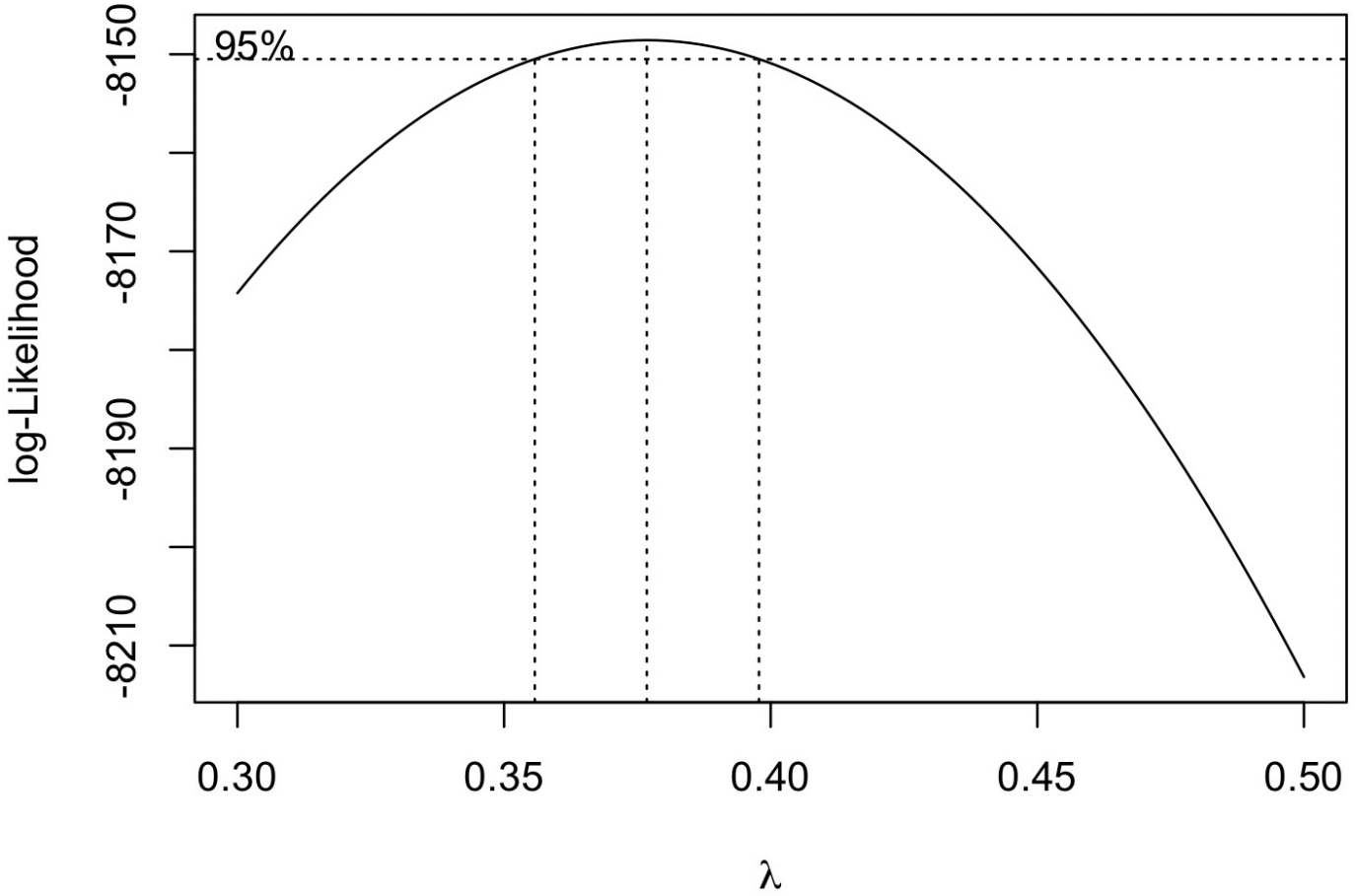
- $price = source + destination + \sqrt{\boldsymbol{distance}} + surge\_multiplier + \boldsymbol{temp^2} + clouds + pressure + humidity + \sqrt{\boldsymbol{wind}} + order\_type + is\_rain + time\_range$

- Adjusted R-squared: 0.9392



MODEL BUILDING

--RESPONCE TRANSFORMATION



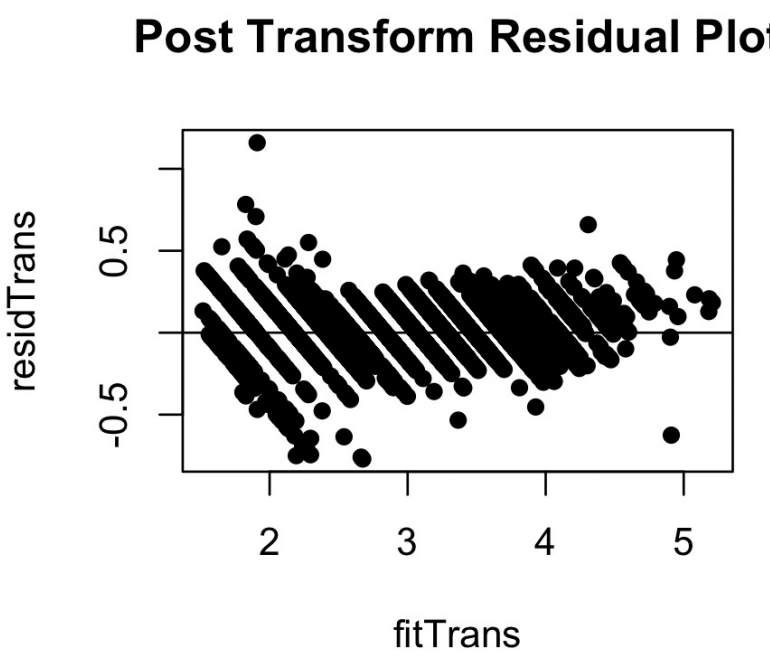
$$price \text{ to } price^{0.4}$$

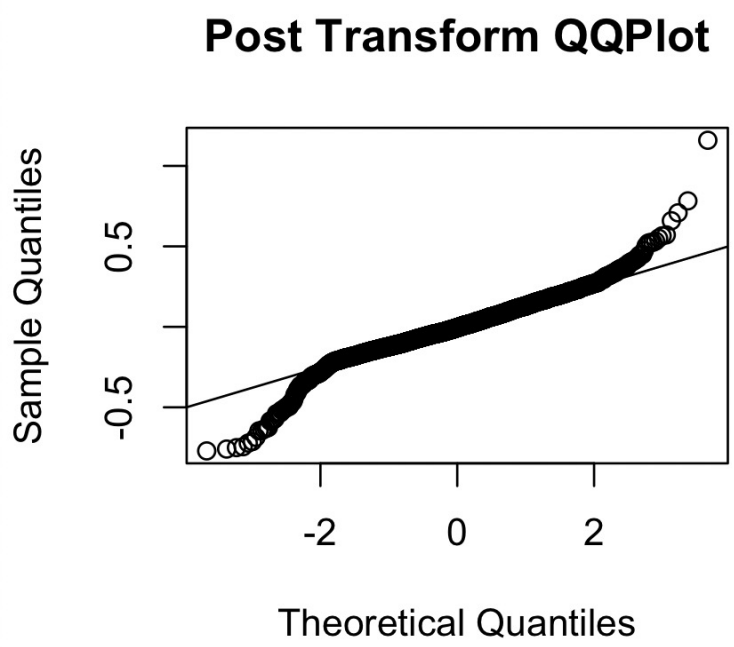
MODEL BUILDING

--RESPONCE TRANSFORMATION

MIAMI HERBERT BUSINESS SCHOOL

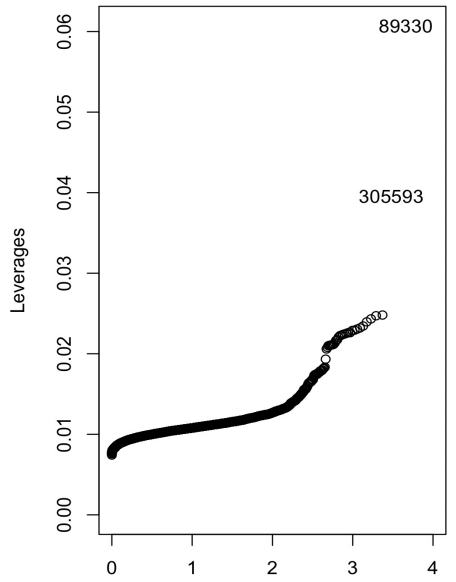
- $\boldsymbol{price}^{\mathbf{0.4}} = source + destination + \sqrt{\boldsymbol{distance}} + surge\_multiplier + \boldsymbol{temp}^{\mathbf{2}} + clouds + pressure + humidity + \sqrt{\boldsymbol{wind}} + order\_type + is\_rain + time\_range$
- Adjusted R^2: 0.961



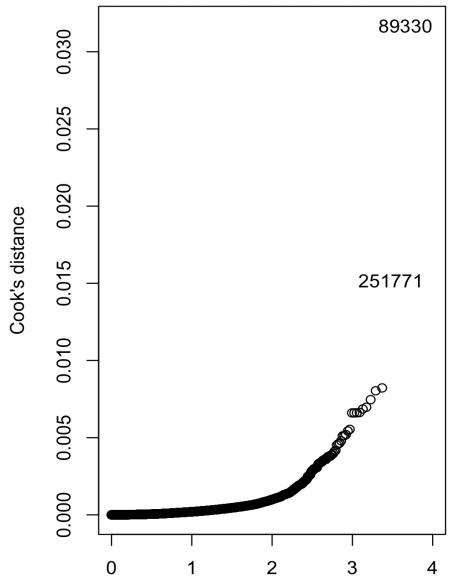


MODEL BUILDING
--OUTLIER DELETION

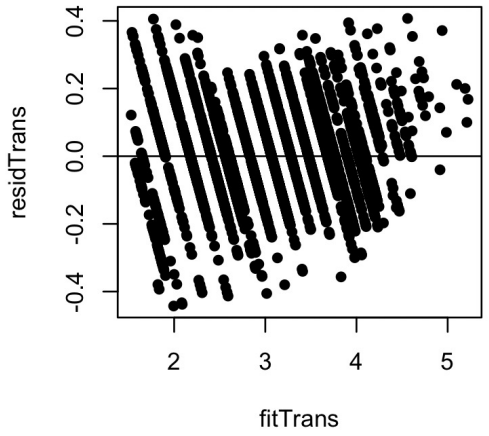
- Delete outliers based on studentized residual, leverage values and cook's distances
- Higher R-squared: 0.9689

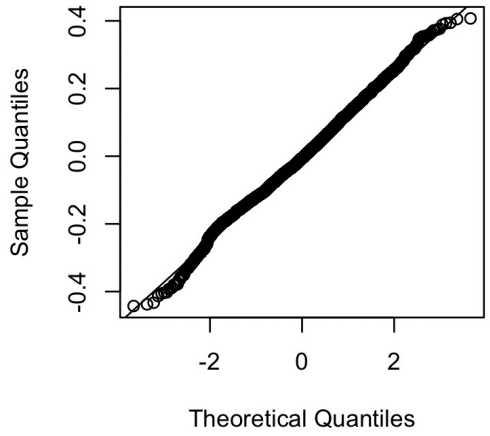


Post Transform Residual Plot



Post Transform QQPlot





MODEL SELECTION

--STEPWISE FORWARD SELECTION(AIC)

MIAMI HERBERT BUSINESS SCHOOL

MODEL (AIC):

$$\boldsymbol{price}^{\mathbf{0.4}} = source + destination + \sqrt{\boldsymbol{distance}} + surge\_multiplier + order\_type$$

	AIC Score
AIC Model	**-5169.35**
Last Model	**-5150.16**

MODEL SELECTION

--INTERACTION TERM

MODEL (with interaction term):

$$price^{0.4}$$

$$= distance * surge\_multiplier + source + destination + \sqrt{distance}$$

$$+ surge\_multiplier + order\_type$$

```
Analysis of Variance Table

Model 1: price^(0.4) ~ order_type + I(distance^(0.5)) + surge_multiplier +
    source + destination
Model 2: price^(0.4) ~ distance * surge_multiplier + order_type + I(distance^(0.5)) +
    surge_multiplier + source + destination
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3918 61.433
2   3916 58.183  2    3.2492 109.34 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	AIC Score
Int_Model	-5379.83
Last Model	-5169.35

MODEL INTERPRETATION

- $distance * surge\_multiplier$
- $source, destination$

MIAMI HERBERT BUSINESS SCHOOL

```
Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.651287   0.055148  11.810  < 2e-16 ***
distance                        0.120742   0.023601   5.116 3.27e-07 ***
surge_multiplier                0.864821   0.042051  20.566  < 2e-16 ***
order_typeLux                   1.032503   0.006785 152.184  < 2e-16 ***
order_typeLux Black             1.375024   0.006813 201.834  < 2e-16 ***
order_typeLux Black XL          1.887140   0.006705 281.456  < 2e-16 ***
order_typeLyft                  0.355852   0.006767  52.583  < 2e-16 ***
order_typeLyft XL               0.843049   0.006762 124.665  < 2e-16 ***
I(distance^(0.5))               0.077883   0.047612   1.636 0.101964
sourceBeacon Hill              -0.035413   0.009641  -3.673 0.000243 ***
sourceBoston University        -0.090223   0.013756  -6.559 6.13e-11 ***
sourceFenway                   -0.041064   0.013766  -2.983 0.002871 **
sourceFinancial District       -0.044476   0.010885  -4.086 4.47e-05 ***
sourceHaymarket Square         -0.037532   0.012874  -2.915 0.003573 **
sourceNorth End                -0.006671   0.013054  -0.511 0.609372
sourceNorth Station            -0.037309   0.009679  -3.854 0.000118 ***
sourceNortheastern University  -0.040784   0.013458  -3.030 0.002458 **
sourceSouth Station            -0.003769   0.013080  -0.288 0.773243
sourceTheatre District          0.027473   0.009778   2.810 0.004985 **
sourceWest End                 -0.024141   0.009431  -2.560 0.010516 *
destinationBeacon Hill         -0.010130   0.009657  -1.049 0.294233
destinationBoston University   -0.052722   0.010573  -4.986 6.42e-07 ***
destinationFenway              -0.059712   0.010088  -5.919 3.52e-09 ***
destinationFinancial District  -0.002583   0.010562  -0.245 0.806815
destinationHaymarket Square    -0.013347   0.009290  -1.437 0.150880
destinationNorth End           -0.004712   0.009441  -0.499 0.617704
destinationNorth Station        0.004360   0.009881   0.441 0.659073
destinationNortheastern University -0.015010   0.010187  -1.473 0.140697
destinationSouth Station              NA         NA      NA       NA
destinationTheatre District     0.024031   0.009585   2.507 0.012216 *
destinationWest End             0.006572   0.009643   0.681 0.495598
distance:surge_multiplier       0.095262   0.016402   5.808 6.83e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1219 on 3919 degrees of freedom
Multiple R-squared:  0.9715,    Adjusted R-squared:  0.9713
F-statistic:  4459 on 30 and 3919 DF,  p-value: < 2e-16
```

MODEL INTERPRETATION

MIAMI HERBERT BUSINESS SCHOOL

- *order_type*

Order type	Beta rate
Lyft	0.356
Lyft XL	0.843
Lux	1.033
Lux Black	1.375
Lux Black XL	1.887
Line**	1

MODEL COMPARISON--MODEL_2

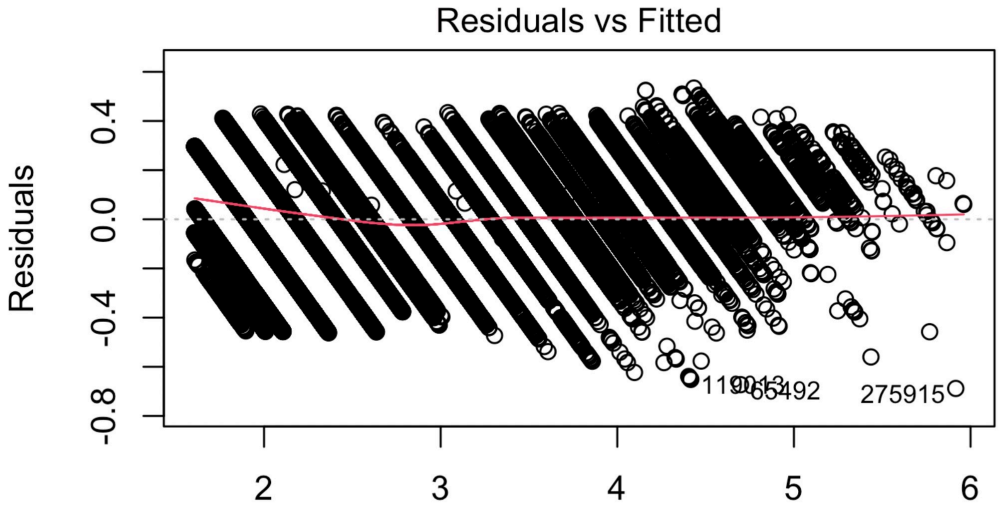
MIAMI HERBERT BUSINESS SCHOOL

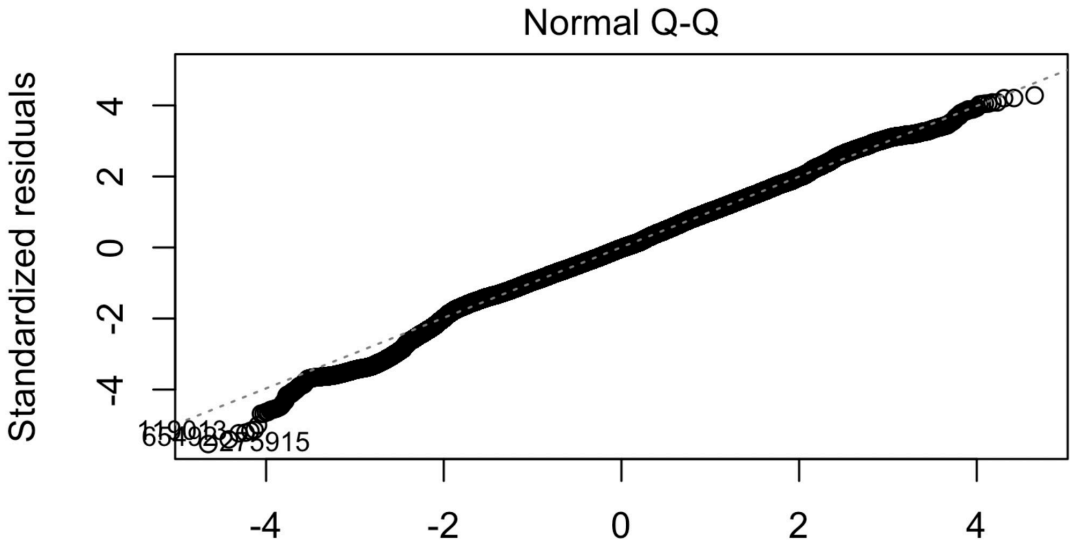
- Model:
  - $\boldsymbol{price}^{\mathbf{0.4}} = source + destination + Distance +$

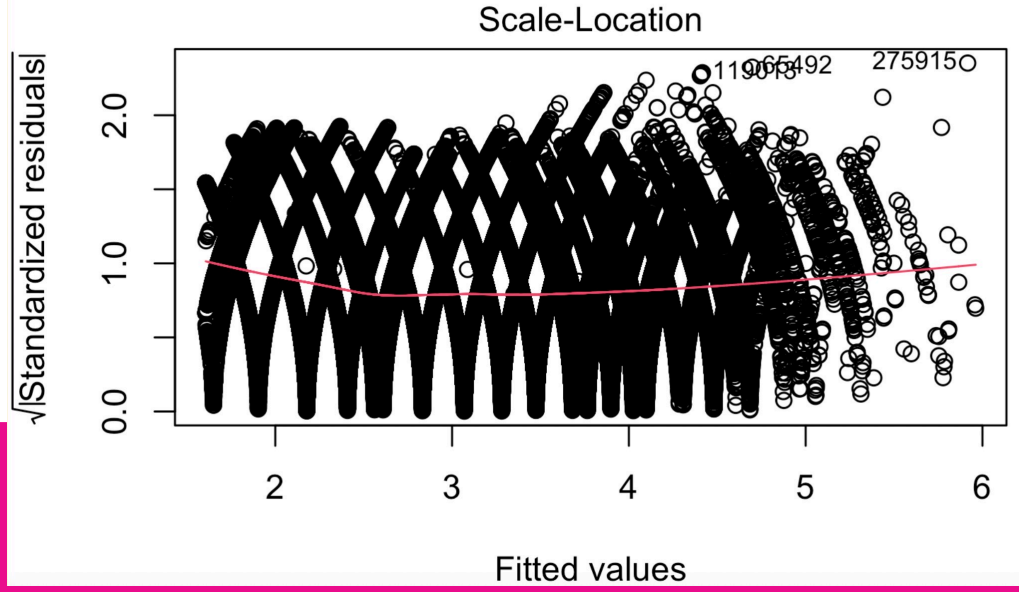
    $surge\_multiplier + temp + clouds + order\_type + is\_rain$
- Difference:
  - Whole dataset
  - Add weather variables

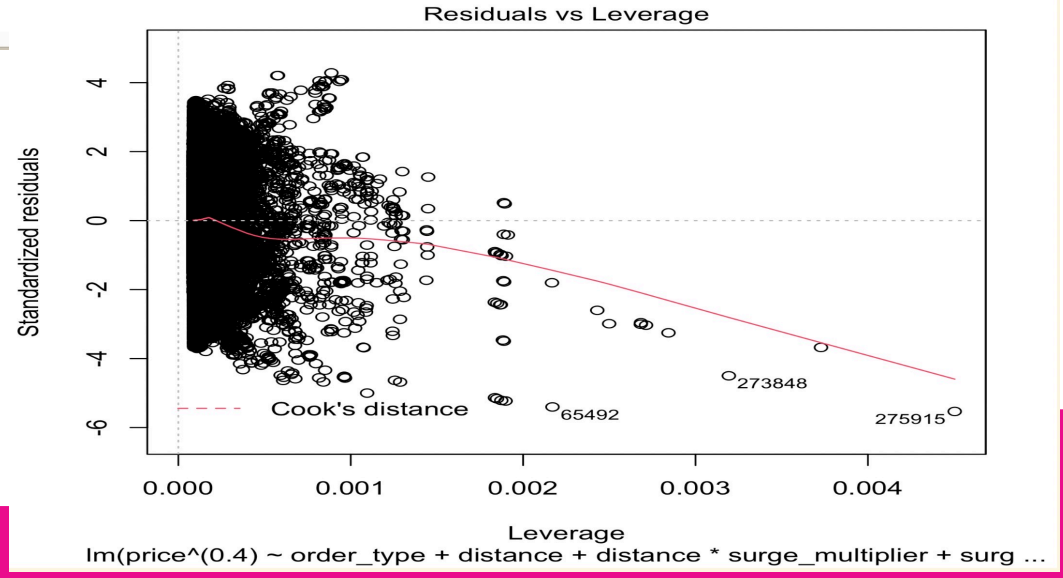
MODEL COMPARISON--MODEL_2

MIAMI HERBERT BUSINESS SCHOOL









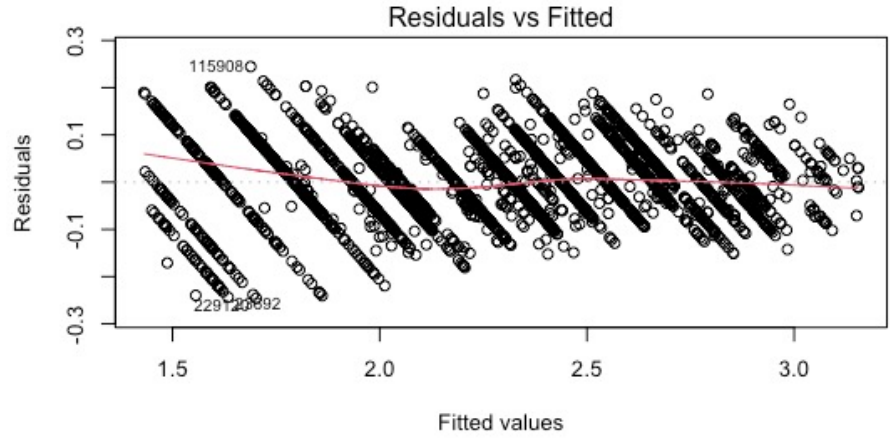
MODEL COMPARISON--MODEL_3

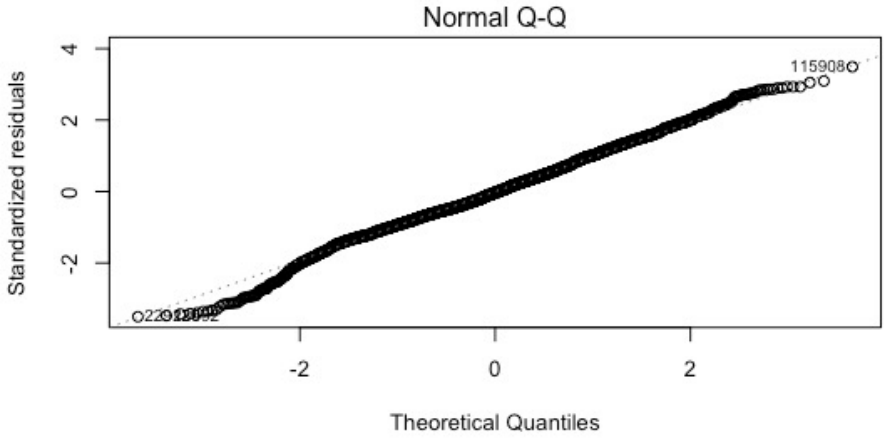
MIAMI HERBERT BUSINESS SCHOOL

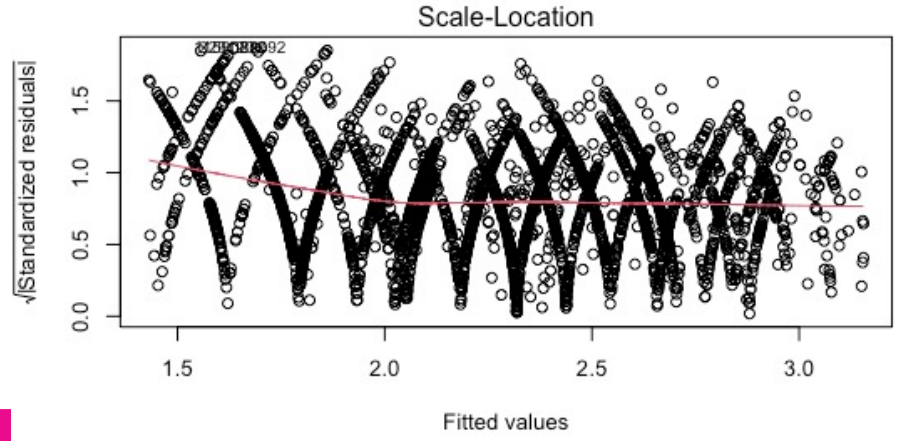
- Model:
  - $\boldsymbol{price\_mult}^{\mathbf{0.3}} = source + destination + \sqrt{\boldsymbol{distance}} + order\_type$
- Difference
  - Price_mult -- price / surge_multiplier
  - Drop surge_multiplier

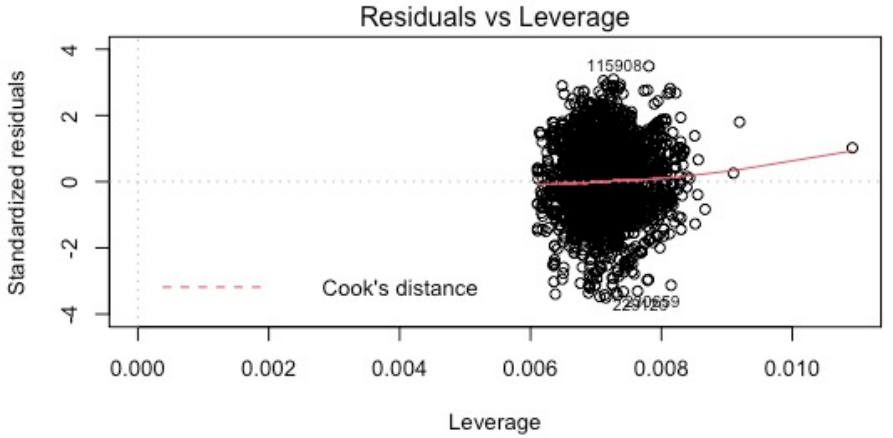
MODEL COMPARISON--MODEL_3

MIAMI HERBERT BUSINESS SCHOOL









MODEL COMPARSION

MIAMI HERBERT BUSINESS SCHOOL

	Adjusted R^2	AIC score
MODEL_1	0.9713	-5379.83
MODEL_2	0.9692	-402101.7
MODEL_3	0.9681	-9743.738

MIAMI HERBERT
BUSINESS SCHOOL

Q & A