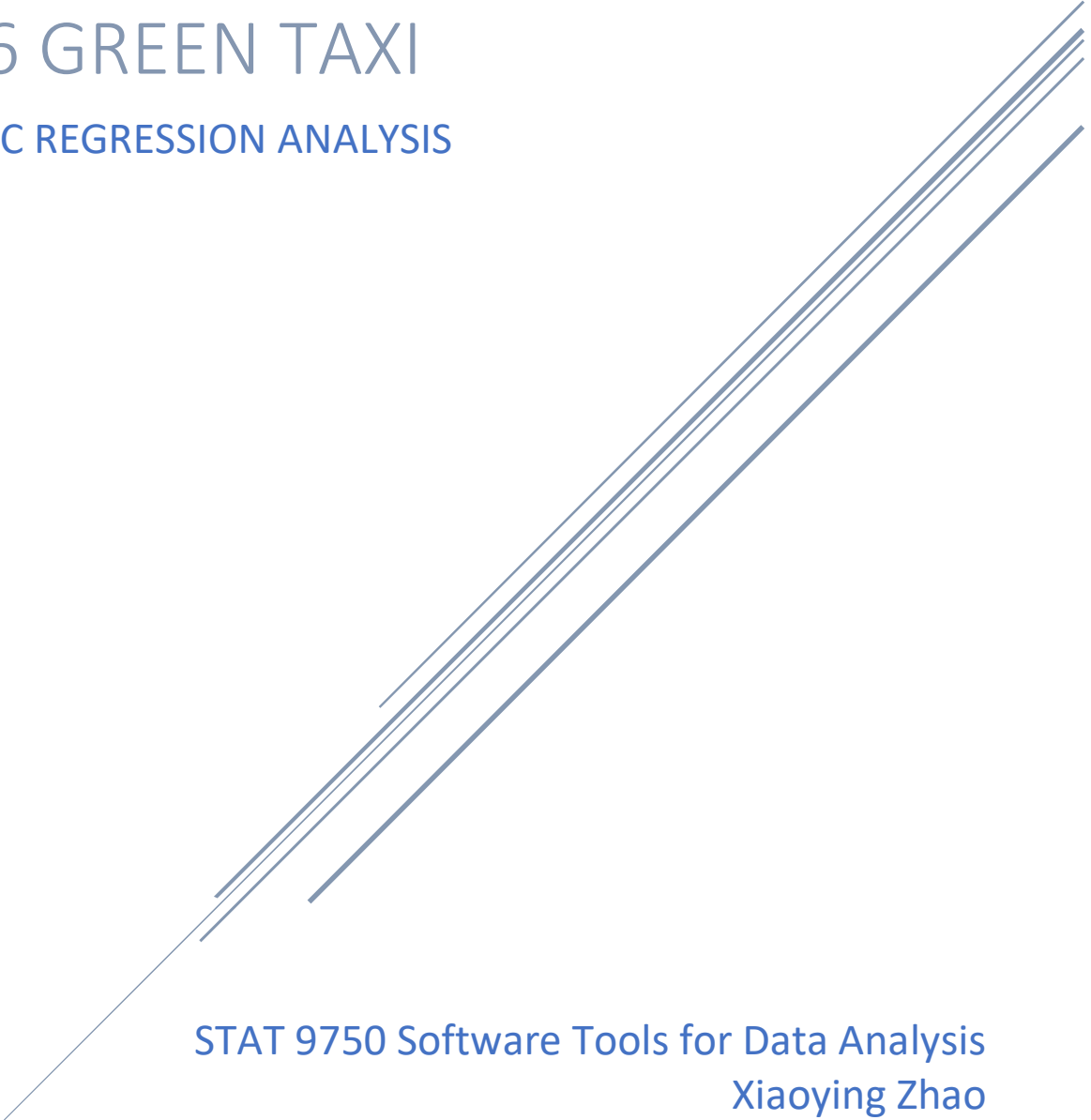


# 2016 GREEN TAXI

## LOGISTIC REGRESSION ANALYSIS



STAT 9750 Software Tools for Data Analysis  
Xiaoying Zhao

## 1. Executive Summary

Boro taxis are taxicabs can be hailed in all outer boroughs except at the airport and upper Manhattan above East 96<sup>th</sup> street and west 110<sup>th</sup> street. The color of Boro taxis are bright green in contrast to the traditional yellow taxis in New York City. There are generally two ways to get a cab, either just street hail or place the order on app. It is interesting to study the factors influencing the order type. This report presents a predict analysis under logistic regression method. The objective of this report is to describe the occurrence of order type of green taxi in terms of a linear combination of possible variables (predictors) and a constant term (intercept).

## 2. Exploratory of Data Analysis

The original data file “2016 Green Taxi Trip” is acquired from NYC Open Data and is generated by the Taxi and Limousine Commission (TLC)<sup>1</sup>. Originally, there are totally 65534 rows and 27 variables. But for this classification application, we only use 12 out of 27 variables. The dependent variable is order types with two outcomes: 1= StreetHail and 2 = Dispatch. Dispatch means the system automatically assign the job to the drive based on the distance. Because of its dichotomous nature, this variable can be the dependent variable for logistic regression.

	Variable	Definition
X	VendorID	A code indicating the LPEP provider that provided the record: 1= Creative Mobile Technologies, LLC; 2 = VeriFone Inc.
	Passenger_count	The number of passengers in the vehicle. (this is a driver-entered value)
	Trip_distance	The elapsed trip distance in miled reported by the taximeter.
	RateCodeID	The final rate code in effect at the end of the trip: 1=Standard rate; 2=JFK; 3=Newark; 4=Nassau or Westchester; 5=Negotiated fare; 6=Group ride
	Fare_amount	The time-and-distance fare calculated by the meter.
	MTA_tax	\$0.50 MTA tax that is automatically triggered based on the metered rate in use.
	Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
	Tolls_amount	Total amount of all tolls paid in trip.
	Improvement_surcharge	\$0.30 improvement surcharge assessed on hailed trips at the flag drop. The improvement surcharge began being levied in 2015.
	Total_amount	The total amount charged to passengers. Does not include cash tips.
	Payment_type	A numeric code signifying how the passenger paid for the trip: 1= Credit card; 2= Cash; 3= No charge; 4= Dispute; 5= Unknown; 6= Voided trip
Y	Trip_Type	A code indicating whether the trip was a street-hail or a dispatch that is automatically assigned based on the metered rate in use but can be altered by the driver: 1= Street-hail; 2= Dispatch.

<sup>1</sup> <https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb>

### 3. Method

#### 3.1 Model Building

The best model is selected by two different methods: stepwise selection and best selection. As for stepwise selection, both forward method and backward method are processed. But the result of forward stepwise is not ideal, since there is only intercept left. As for best selection method, two criteria are adopted: BICq and AIC. These four ways present completely different outcome. The result under AIC best selected shows the greatest number of variables, which is 8. The following procedures go with this model. The output is presented below.

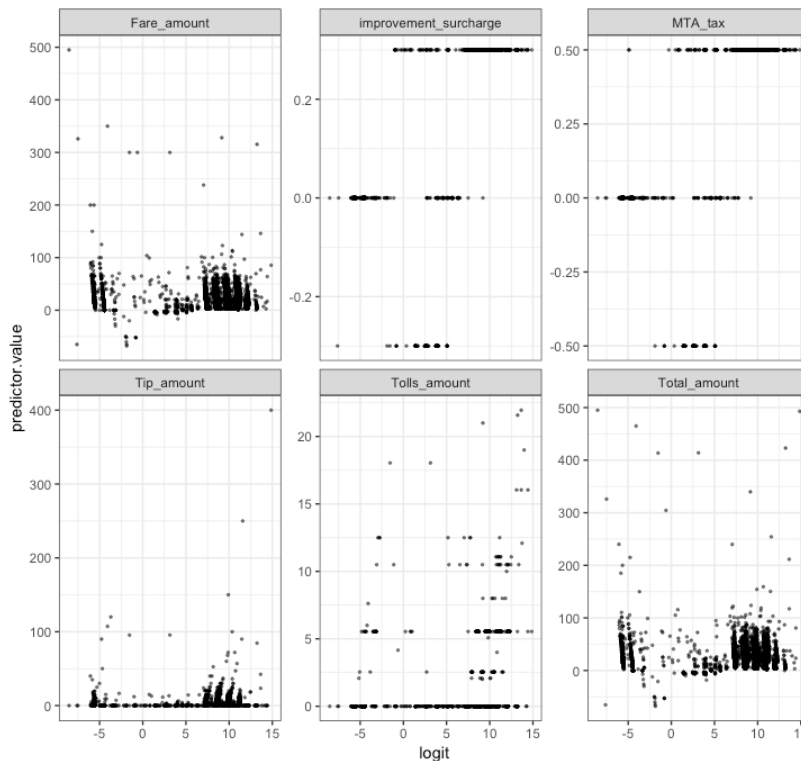
AIC  
BICq equivalent for q in (0.96669751285975, 0.994175145234698)  
Best Model:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	6.062415	0.8037309	7.542842	4.598375e-14
RateCodeID	-2.531137	0.1179861	-21.452837	4.296318e-102
Fare_amount	2.556416	0.7044770	3.628814	2.847261e-04
MTA_tax	3.956283	1.6652072	2.375850	1.750858e-02
Tip_amount	2.579201	0.7060774	3.652859	2.593363e-04
Tolls_amount	2.796398	0.7134464	3.919563	8.870962e-05
improvement_surcharge	18.136579	2.2375050	8.105716	5.243596e-16
Total_amount	-2.562528	0.7047960	-3.635843	2.770731e-04
Payment_type	1.080586	0.3208214	3.368184	7.566500e-04

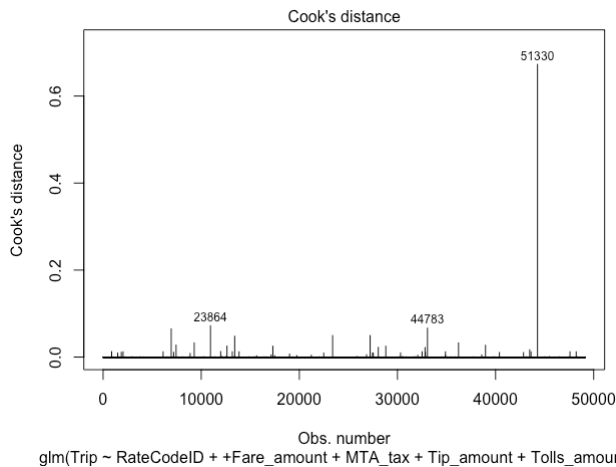
Therefore, the best model for logistic regression analysis for green taxi data is presented as below.

$$\hat{y} = \frac{e^{6.06-2.53RateCodeID+2.55Fare+3.95MTAtax+2.57Tip+2.79Toll+18.12Imp-2.56Total+1.08Pay}}{1 + e^{6.06-2.53RateCodeID+2.55Fare+3.95MTAtax+2.57Tip+2.79Toll+18.12Imp-2.56Total+1.08Pay}}$$

#### 3.2 Model Fit



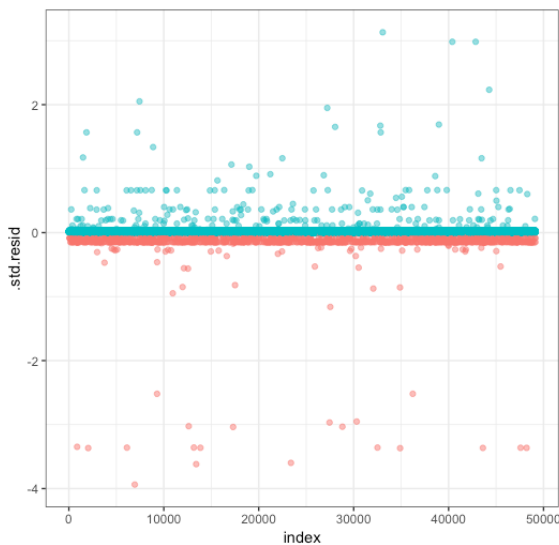
This section explores the outliers, high leverage points, and influential points. To achieve this objective, we need to check the linear relationship between continuous predictor variables and the logit of the outcome. After best selection, there are 8 predictors contained in the model. Two of them are rate code ID and payment type, which are categorical variables. Therefore, we do need to consider it in this component. The diagnostic scatterplots do not present ideal linear relationship. These variables might need some transformation. I may need to



.rownames	Trip	RateCodeID	Fare_amount	MTA_tax	Tip_amount
<chr>	<fct>	<int>	<dbl>	<dbl>	<dbl>
1	23864	Disp...	5	300	0
2	44783	Stre...	5	9.2	0.5
3	51330	Stre...	5	300	0

include 2- or 3- power term, fractional polynomials and spline function in order to achieving ideal linear relationship.

Then I would like to check the influential values, which can be examined by visualizing the Cook's distance value. We need to understand that not all outliers are influential points. We introduce standardized residual error to detect whether the data has potential influential points. Based on the standardized residuals and the Cook's distance, #23864, #44783, and #51330 observations are the top three largest values, displayed on the left side, deserving close attention. I also plot the standardized residuals.



.rownames	Trip	RateCodeID	Fare_amount	MTA_tax	Tip_amount
<chr>	<fct>	<int>	<dbl>	<dbl>	<dbl>
1	7389	Disp...	1	15.5	0
2	60400	Disp...	1	5	0
3	3849	Disp...	1	8.5	0
4	19375	Disp...	1	8.5	0
5	51666	Disp...	1	21	4.2
6	5511	Disp...	1	9.5	0
7	34643	Disp...	1	5.5	0
8	54959	Disp...	1	7.5	0
9	7048	Disp...	1	10	2.5
10	31114	Disp...	1	17	0
11	8010	Disp...	1	9	1.35
12	51817	Disp...	1	9	0
13	44783	Stre...	5	9.2	0.5
14	9986	Disp...	1	4	0
15	5623	Disp...	1	6	0
16	54679	Disp...	1	7.5	0
17	59759	Disp...	1	6.5	0

In the standardized residuals plot, we can see that there are some data points with absolute standardized residual value greater than 3, then I use filter function to present these specific observations. When we are facing outliers in the continuous predictor, we can remove the potential outliers, perform a logarithm transformation, and adapt non-parametric methods.

```
> car::vif(bestmodel)
RateCodeID      Fare_amount      MTA_tax
1.605904      27452.752544      12.447775
Tip_amount      Tolls_amount improvement_surge
1233.715881      76.861550      9.037943
Total_amount      Payment_type
37496.120411      2.568985
```

I also exam the multicollinearity by applying vif() function to the model. VIF stands for variance inflation factors. If a VIF value exceeds 5 or 10, it indicates a problematic amount of collinearity. It seems

there is sever collinearity issue. The variance inflation factor values of many predictors are extremely large. To solve the problem, I may consider removing these variables.

### 3.3 Model Interpretation

In this section, I will interpret coefficient of predictors and explore the importance of these predictors. Recall the best model I find the AIC best selection method.

$$\hat{y} = \frac{e^{6.06-2.53RateCodeID+2.55Fare+3.95MTAtax+2.57Tip+2.79Toll+18.12Imp-2.56Total+1.08Pay}}{1 + e^{6.06-2.53RateCodeID+2.55Fare+3.95MTAtax+2.57Tip+2.79Toll+18.12Imp-2.56Total+1.08Pay}}$$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	6.0624	0.8037	7.543	4.60e-14 ***
RateCodeID	-2.5311	0.1180	-21.453	< 2e-16 ***
Fare_amount	2.5564	0.7045	3.629	0.000285 ***
MTA_tax	3.9563	1.6652	2.376	0.017509 *
Tip_amount	2.5792	0.7061	3.653	0.000259 ***
Tolls_amount	2.7964	0.7134	3.920	8.87e-05 ***
improvement_surcharge	18.1366	2.2375	8.106	5.24e-16 ***
Total_amount	-2.5625	0.7048	-3.636	0.000277 ***
Payment_type	1.0806	0.3208	3.368	0.000757 ***

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			49149	10291.6	
RateCodeID	1	9734.8	49148	556.7	< 2.2e-16 ***
Fare_amount	1	15.6	49147	541.1	7.769e-05 ***
MTA_tax	1	128.3	49146	412.8	< 2.2e-16 ***
Tip_amount	1	0.6	49145	412.2	0.4415199
Tolls_amount	1	7.6	49144	404.7	0.0059460 **
improvement_surcharge	1	41.6	49143	363.0	1.099e-10 ***
Total_amount	1	15.6	49142	347.4	7.662e-05 ***
Payment_type	1	11.3	49141	336.0	0.0007551 ***

Let's take a look at the coefficient in details. We can see from the R output, under AIC best selection method, all 8 predictors are statistically significant. All these variables are with p-values less than 0.05, meaning they have a strong association with the order type, and they should be contained in the model. Each estimated coefficient is the expected change in the log odds of being in an honors class for a unit increase in the corresponding predictor

variable holding the other predictor variables constant at certain value.

Then I use anova function to create the table of deviance. We need to pay attention to the difference between the null deviance and the residual deviance. The bigger is difference the better the best model is against the null model. We can see from the output that all these predictors have visible difference, meaning adding any one of these predictors alone reduces the deviance dramatically. If the slope has a small deviance, it means this predictor does not add much to the model, and similar amount of variation is examined. I also apply the McFadden index, R<sup>2</sup>M, to explore the coefficient of determination.

```
> pR2(bestmodel)
      llh      llhNull      G2      McFadden      r2ML      r2CU
-168.0152135 -5145.7786949 9955.5269628  0.9673489  0.1833576  0.9705476
```

I also want to explore the importance of each predictor to predict the order type by dominance analysis, which is used to determining the relative importance of predictors in the regression analysis. Importance means a qualitative comparison between predictor pairs. I use dominanceAnalysis function to perform the analysis and use dominanceMatrix function to print

out the summarized result.

```
> dominanceMatrix(dapres,type="complete",fit.function="r2.m")
```

	RateCodeID	Fare_amount	MTA_tax	Tip_amount	Tolls_amount	improvement_surcharge	Total_amount	Payment_type
RateCodeID	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Fare_amount	0.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5
MTA_tax	0.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Tip_amount	0.0	0.5	0.5	0.5	0.5	0.0	0.5	0.5
Tolls_amount	0.0	0.5	0.5	0.5	0.5	0.0	0.5	0.5
improvement_surcharge	0.0	0.5	0.5	1.0	1.0	0.5	0.5	0.5
Total_amount	0.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Payment_type	0.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5

This complete dominance matrix summarizes the relationship between predictor pairs. If the value is 1, it means the predictor under the first column completely dominates the other predictors in the pair. If the value is 0.5, complete dominance cannot be established. If the value is 0, it means the predictor under the first column is completely dominated by the other predictors in the pair. We can use conditional dominance when the complete dominance cannot be established. The conditional dominance can be explored by contributionBylevel function in R.

```
Contribution by level
* Fit index: r2.m
```

level	RateCodeID	Fare_amount	MTA_tax	Tip_amount	Tolls_amount	improvement_surcharge	Total_amount	Payment_type
0	0.946	0.026	0.736	0.000	0.001	0.728	0.013	0.004
1	0.736	0.027	0.529	0.003	0.001	0.525	0.019	0.012
2	0.555	0.041	0.353	0.023	0.002	0.350	0.037	0.014
3	0.396	0.042	0.199	0.031	0.007	0.198	0.040	0.014
4	0.269	0.023	0.077	0.019	0.006	0.077	0.022	0.013
5	0.196	0.002	0.009	0.001	0.001	0.010	0.001	0.009
6	0.186	0.000	0.002	0.000	0.001	0.006	0.000	0.005
7	0.181	0.002	0.001	0.002	0.002	0.006	0.002	0.001

The average additional contribution of each predictor is calculated, under different model size. Take level 2 as an example. The average additional contribution to model of size 2 is calculated as  $(0.555+0.041+0.353+0.023+0.002+0.350+0.037+0.014)/7=0.196$ . We will compare the average additional contributions across all model sizes to set up the conditional dominance.

```
> dominanceMatrix(dapres,type="conditional",fit.function="r2.m")
```

	RateCodeID	Fare_amount	MTA_tax	Tip_amount	Tolls_amount	improvement_surcharge	Total_amount	Payment_type
RateCodeID	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Fare_amount	0.0	0.5	0.5	0.5	0.5	0.0	0.5	0.5
MTA_tax	0.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Tip_amount	0.0	0.5	0.5	0.5	0.5	0.0	0.5	0.5
Tolls_amount	0.0	0.5	0.5	0.5	0.5	0.0	0.5	0.5
improvement_surcharge	0.0	1.0	0.5	1.0	1.0	0.5	1.0	1.0
Total_amount	0.0	0.5	0.5	0.5	0.5	0.0	0.5	0.5
Payment_type	0.0	0.5	0.5	0.5	0.5	0.0	0.5	0.5

If the conditional dominance also cannot be established, we could introduce the averageContribution function to compute the mean of each predictor's conditional measures. It is used to determine the general dominance. The predictor RateCodeID has the highest value of 0.433 and generally dominates all other predictors. Then in the general dominance matrix, the predictor RateCodeID assumes a value of 1 with all other predictors.

```
Average Contribution by predictor
```

	RateCodeID	Fare_amount	MTA_tax	Tip_amount	Tolls_amount	improvement_surcharge	Total_amount	Payment_type
r2.m	0.433	0.02	0.238	0.01	0.003	0.238	0.017	0.009

```
> dominanceMatrix(dapres,type="general",fit.function="r2.m")
```

	RateCodeID	Fare_amount	MTA_tax	Tip_amount	Tolls_amount	improvement_surcharge	Total_amount	Payment_type
RateCodeID	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Fare_amount	0.0	0.5	0.0	1.0	1.0	0.0	1.0	1.0
MTA_tax	0.0	1.0	0.5	1.0	1.0	1.0	1.0	1.0
Tip_amount	0.0	0.0	0.0	0.5	1.0	0.0	0.0	1.0
Tolls_amount	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0
improvement_surcharge	0.0	1.0	0.0	1.0	1.0	0.5	1.0	1.0
Total_amount	0.0	0.0	0.0	1.0	1.0	0.0	0.5	1.0
Payment_type	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.5

If we want to exam the rank of importance among predictors, we can explore the raw values of general dominance. For green taxi case, RateCodeID is the most important predictor to explain the order type (0.433), followed by MTA tax (0.238), improvement surcharge (0.238), fare amount (0.02), total amount (0.017), tip amount (0.01), payment type (0.009), and tolls amount (0.003).

## 4. Conclusion

This report presents a logistic regression analysis of green taxi data in term of describing the occurrence of order type of green taxi in terms of a linear combination of possible variables (predictors) and a constant term (intercept). There are mainly three components to conduct the logistic regression analysis. Firstly, after briefly exploratory data analysis, it uses the best selection method under AIC criterion to build the best possible model. Then it explores the potential outliers, high leverage points, and influential point by standardized residuals and Cook's distance. Eventually, it interprets the model by examining the coefficients, importance, rank of predictors. We can conclude that rate code ID is the most important predictors of the model.