Statistical Analysis Report

Tip Frequency In New York City Yellow Taxi Services

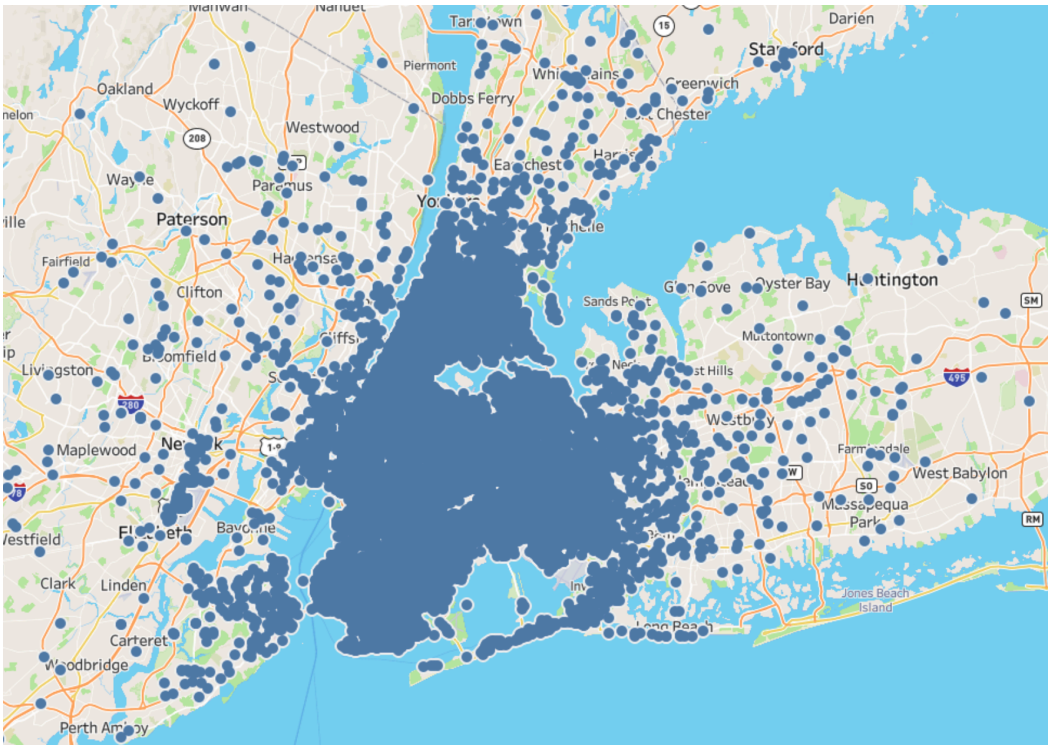

Table of Contents

1. Introduction

Adequate methods to evaluate high tip amounts has been an urgent need for people who participated in taxi industries. This report aims to discover the correlations between tips frequency and the other attributes including external data source such as weather and temperature.

The expected outcome in this report would be a generalized linear regression attempt regards the tip occurrence of a trip. That is, the number of trips ends up with a tip amount above 0 within 24 hours. Trade-offs and developments between the two models will be discussed in detail. To develop a valid model, we introduce a range of data preprocessing and analytical steps in this project.

2. Methodology

Adopted tools: R, Python

2.1 Data Selection

To start with, we conduct researches regard the taxi industries to gain insightful background knowledge.

The adopted data source in this project is New York City Taxi and Limousine Commission (also known as TLC), a government-owned constitution who regulates taxi industries. According to Todd (2016), for-hire vehicles such as Uber has already become the taxi genre with the largest market share since 2016. FHV utilize its main advantages such as inexpensiveness and convenience to steal a large proportion of customers from the traditional taxi industry. There is no doubt that FHV is replacing yellow and green taxis to become the mainstream taxi service genre. Yellow taxis have been an expensive choice; however, people are stilling willing to pay tips for a yellow taxi ride. Hence, we are interested in tips frequency in the yellow taxi industries after a huge loss in the number of customers. Since the rapid expansion of the FHV industry takes place in recent years, the determination regards the data period is selected to be from 2017 to 2018. The detail of the sampling steps is introduced later in the text.

Nevertheless, there are potential concerns about taxi data. The most concerning factor is data integrity. Sometimes the data is not recorded properly as technological issues happen. Also due to the on-developing privacy concerns, there is no guarantee of completeness and accuracy from TLC regards taxi data. Therefore, there is a risk of developing biased and inaccurate data during the investigation. With this in mind, adequate data preprocessing and data sampling has been conducted to avoid handling biased data.

2.2 Data Sampling and Subsampling

There are two sets of data gathered for analysis, which are the whole year records in 2017 and 2018. The plan is to use the data in 2017 as the training set, whereas records in 2018 are used as development and test data. For the training set, data are randomly sampled among the whole years. In this project, we aim to select a large scale of data and we expect a generalized but powerful model. Therefore, we randomly sample 2,000 records from each month in 2017. There is a total amount of 24,000 records in the training set. For developing and testing data, we randomly sample 300 records from each month in 2018. There is a total amount of 3,600 records in testing and developing set. Overall, the expected ratio between training, developing and testing set is approximately 70:15:15.

2.3 External Data Source

The external data source is also considered in this project. In this case, it is believed that weather and temperature may potentially influence the tip frequency. To examine the climate factors, we adopted the data from "holiday-weather.com". The annual weather records can be found from the website. In this website, we choose to use daily weather record throughout 2017.

To extract data from the website, the web crawler techniques are applied. In this case, the adopted tool is "import.io". After filtering the raw climate data, there are four remaining components in the dataset in which we are interested in.
day: The date of that month.
tempC: The daily average temperature in the Celsius unit.
tempF: The daily average temperature in the Fahrenheit unit.
weather: A written description regards the weather. There are four discrete values as "Cloudy", "Sunny", "Rain", "Snow".

The taxi data and climate data are joined by applying inner join on the date. It has to be stated that several dates in June and July did not contain valid climate records. Eventually, it results in a minor reduction in the number of records.

2.4 Data Cleaning and outlier handling

The data cleaning start with handling missing values. Although the target attribute is the tip amount in this project, we are not expecting to see any missing value in every column. It is because the missing value is considered to hurt the fit or result in biased outcome when applying statistical models. There are two options considered to deal with missing values. The first is to apply mean imputation. This method is easy to implement, however, this method may risk of losing statistical power. We prefer the data that are not implemented by the investigator. Hence, the records with missing values are dropped.

The second step of data cleaning is to regard the outliers. We detect the outliers from both the upper bounds and lower bounds. The aim is to remove unrealistic values. For instance, the upper bound for a trip duration does not exceed 12 hours as TLC stated. Hence, the records with a trip duration above 12 hours are not taken into consideration. Furthermore, except for Celsius's temperature, all negative values from any arbitrary columns seem to be unrealistic and need to be removed. In addition, records with a passenger count of 0 are also considered to be unreasonable and need to be dropped.

2.5 Data Preprocessing

Data preprocessing regards filtering information from unexaminable attributes are considered. For instance, the attribute "pickup DateTime" seems to be essential or relevance when developing a model to explain tips amount. However, the "Data time" type is neither continuous nor discrete. In other words, it is mathematically uninterpretable. However, drop this attribute may result in a loss regards critical information. After consideration, we decide to transform the data into a discrete attribute. We assign an integer from range 1 to 4 for each record. The integer represents the season where the rides take place. "1" representing Spring, start from March to May; "2 "

representing Summer, start from June to August; "3" representing Autumn, start from September to November; "4" representing Winter, start from December to January.

There are some attributes only taking discrete values in this dataset. For instance, "mta_tax" and "improvement surcharge" only take two possible values, 0 or 0.5. In other words, it can be transformed into binary attributes. Furthermore, data preprocessing also include converting discrete string variables into numerical values. For instance, the "weather" attribute contains four discrete strings as "Cloudy", "Sunny", "Rain", "Snow". We convert each of them into numerics to generate an interpretable categorical variable. The resulting transformation areas below "Cloudy"=1, "Sunny"=2, "Rain"=3, "Snow"=4.

For tempC, this attribute records Celsius's temperature, which may include negative values. Since negative values are not capable for log transformation. There are potential disadvantage regards this attribute. To overcome the issue, we add a constant to this attribute to let the minimum value become a positive integer.

Note that at this stage, the data is ready for linear regression modelling. However, we are interested in the tip frequency within a given period. That is, our response variable is count data. It is believed that the Poisson distribution may satisfied the assumptions better than the linear regression, which is discussed in detail later in text. Hence, further data preprocessing is performed.

To make the data suitable for Poisson regression model, our mission is to transfer the 'tip_amount' field into count data. The first operation is to change the 'tip_amount' from continuous variable to binary attribute. Let the record with a tip amount of 0 be 0, records with any value above 0 be 1. Note that at this stage, the data is capable for a generalized linear model with binomial link function. Binomial glm is a reasonable option for determining whether a taxi ride ends up with a tip income, however, not a good option for determining ride count. The reason is that the expect outcome for binomial glm is usually a probability or binary result (True or False), whereas we are seeking for count data in this project.

Next, aggregate the data by a specific period. The chosen period in this project is 24 hours. That is, count the number of records that ends up with a tip income within 24 hours. For the rest attributes, we take average of the values. Note that, this method is reasonable when handling continuous data but not categorical data. For instance, the aggregated "Vender_ID" become a continuous variable. We consider this type of transformation make the attribute loses it characteristic and information. If we choose to assign the instance with a value with the largest frequency, the resulting data could become biased. Hence, we dropped the categorical attributes that cannot be merged. However, categorical attributes such as daily weather and season are still categorical after merging. We keep those mergeable attributes since merging does not reduce its information.

3. Trade-offs Between Linear Regression and Poisson Regression

The following discussion is conducted regards modelling and distribution selection for the data.

The research theme for this assignment is "Yellow Taxi Tip Frequency within 24 hours in NYC". From the topic, we can conclude fowling facts.

- The expected outcome is integer.
- The expected outcome is non-continuous.
- The expected outcome is the number of occurred event within a fixed time series.

- The events (taxi rides with tips) occur with a specific rate.
- The events (taxi rides with tips) are independent of each other.

The above facts indicate that the tip frequency satisfied the assumptions of Poisson distribution. If we want to model the data assuming Poisson distribution, generalized linear model with a Poisson link is a popular choice. However, it is still considered to be appropriate to use a linear regression. Hence, let us determine the assumptions for linear model and generalized linear model for comparison.

In general cases, there are following four assumptions we need to satisfy when applying linear regression models.

- Weak exogeneity

In other words, the predictors are correctly observed. In this case, we consider TLC to be a reliable source. Hence, this assumption is satisfied.

- Linear relationship between predictors and response variable

In general cases, this assumption is satisfied as long as the outcome model is considered to be reasonable. The usual method for checking this assumption is produce diagnostic plot after we come up with a decent model.

- Constant variance & independent errors

In general cases, this assumption is satisfied as long as the outcome model is considered to be reasonable. The usual method for checking this assumption is produce diagnostic plot after we come up with a decent model.

- X is of full rank

This is assumption is satisfied since there is no grouping occur at this stage.

As shown above, we consider the data satisfied the linear regression assumptions. For generalized linear models, there are three more components in addition to general linear models (Tony, 2008).

- The probability distribution comes from the exponential family

As discussed before, Poisson distribution satisfied the assumptions and it is an exponential distribution.

- Valid link function g suggesting that $g^{-1}(\mu) = \mu = E(Y)$
- Linear relationship between η and predictors, that is a linear predictor $\eta = X\beta$

In general cases, this assumption is satisfied as long as the outcome model is considered to be reasonable. The usual method for checking this assumption is produce diagnostic plot after we come up with a decent model.

At the current stage, it seems that a generalized linear model with Poisson link is a more specific approach in comparison to the traditional linear regression. The reason is that Poisson distribution is usually the perfect suit for event frequency. Nevertheless, there exists a huge disadvantage for Poisson distribution. That is, Poisson regression usually assume that the variance and the mean are equal. In reality, however, glm with Poisson link usually suffers from an issue called overdispersion. This issue indicates that Poisson regression no longer provide good fit on the data. The potential alternative solution including fitting a negative-binomial regression instead or apply quasi-likelihood.

Negative binomial distribution is a relative distribution to Poisson distribution. Negative binomial distribution is in general expected to converge to a Poisson distribution. In short, assume the probability of an event occur is p, and the probability of not occurring is (1-p). Until r failures happens, the number of success X follows a negative binomial distribution as $X \sim NB(r, p)$. Negative binomial distribution is also exponential family. Hence, all assumptions and properties from linear regression and glm hold for negative binomial link glm. The most important note regards negative binomial distribution is that we assume inequality between conditional mean and variance.

Overall, the linear regression is considered to be appropriate for general cases. Poisson link glm is suitable for the scenario, however, it may suffer from overdispersion. If overdispersion happens, the resulting fit for the Poisson model may be less accurate than linear model. To overcome this issue, we may consider a negative binomial

approach. It is hard for us to determine the goodness of a model fits without a model. Hence, the next step is constructing models by applying different techniques for further comparison.

4. Exploratory Analysis and Key Findings

First of all, we determine the data types. That is, the background information for each attribute.

N	Season	fare_amount	tempC	tempF
Min. :20.00	Min. :1.000	Min. : 9.439	Min. : 1.00	Min. :19.00
1st Qu.:57.00	1st Qu.:1.000	1st Qu.:11.968	1st Qu.:14.00	1st Qu.:42.00
Median :65.00	Median :3.000	Median :12.689	Median :20.00	Median :54.00
Mean :64.21	Mean :2.498	Mean :12.898	Mean :21.07	Mean :55.48
3rd Qu.:73.00	3rd Qu.:4.000	3rd Qu.:13.693	3rd Qu.:30.00	3rd Qu.:71.00
Max. :93.00	Max. :4.000	Max. :17.035	Max. :38.00	Max. :86.00

tip_amount	tolls_amount	total_amount	trip_distance	weather
Min. :1.042	Min. :0.0000	Min. :12.13	Min. :1.963	Min. :1.000
1st Qu.:1.660	1st Qu.:0.2174	1st Qu.:15.09	1st Qu.:2.600	1st Qu.:1.000
Median :1.818	Median :0.3058	Median :16.03	Median :2.910	Median :2.000
Mean :1.830	Mean :0.3158	Mean :16.17	Mean :2.915	Mean :1.888
3rd Qu.:2.027	3rd Qu.:0.4064	3rd Qu.:17.29	3rd Qu.:3.156	3rd Qu.:2.000
Max. :2.637	Max. :0.8743	Max. :21.08	Max. :4.263	Max. :4.000

Figure 1 Descriptive statistics regards the attributes

N: The response variable, representing the number of taxis with tips within24 hours.
Season: Categorical attribute, four discrete representing each season. We treat it as factors.
fare_amount: The average of pure time-distance fare amount within 24 hours.
tempC: The average temperature in the Celsius unit within 24 hours.
tempF: The average temperature in the Fahrenheit unit within 24 hours.
tip_amount: The average tip rate within 24 hours.
tolls_amount: The average tolls amount within 24 hours.
total_amount: The average total cost for the rides within 24 hours.
trip_distance: The average trip distance within 24 hours.
weather: Categorical attribute with 4 levels, each representing the weather within 24 hours. We treat it as factors

From the above descriptive statistics, we can see that there is no strong evidence of outliers. Besides that, the distribution of each continuous seems approximately normal. The only concern is that the distribution spread of temperature is much wider in comparison to the rest attributes. We may need to apply log transformation on temperature.

Furthermore, since we have added the seasonal factor and weather factor to the dataset, we expect the categorical attributes have impact on the distribution of N. Thus, the following plots are produced.

As presented in figure 2, there are two plots indicating the distribution of N under climate factors. To begin with, the plot on the left side demonstrates the distributions of N under season groups. We can see that in Spring, there seems to exits two groups, which are from 45 to 60 and 60 to 80. Summer and Autumn seems to be normally distributed. In Winter, however, there are evidences of left skewness. In sum, we can see different distribution characteristics from

each group. This fact may suggest that season is a significant attribute while fitting.

The plot of the right side, indicate weather factors. We can see that the distribution of N under cloudy and sunny weather are appears to be normal. The problem is the rain and snow weather. There is a strong sign of insufficient data under rainy weather. This can result in a biased mislead. Whereas in snow weather, the data volume is also insufficient. Furthermore, there are two peaks indicating a probability of two populations. Overall, weather attribute may result in biased results since the data volume are not evenly distributed among each group.

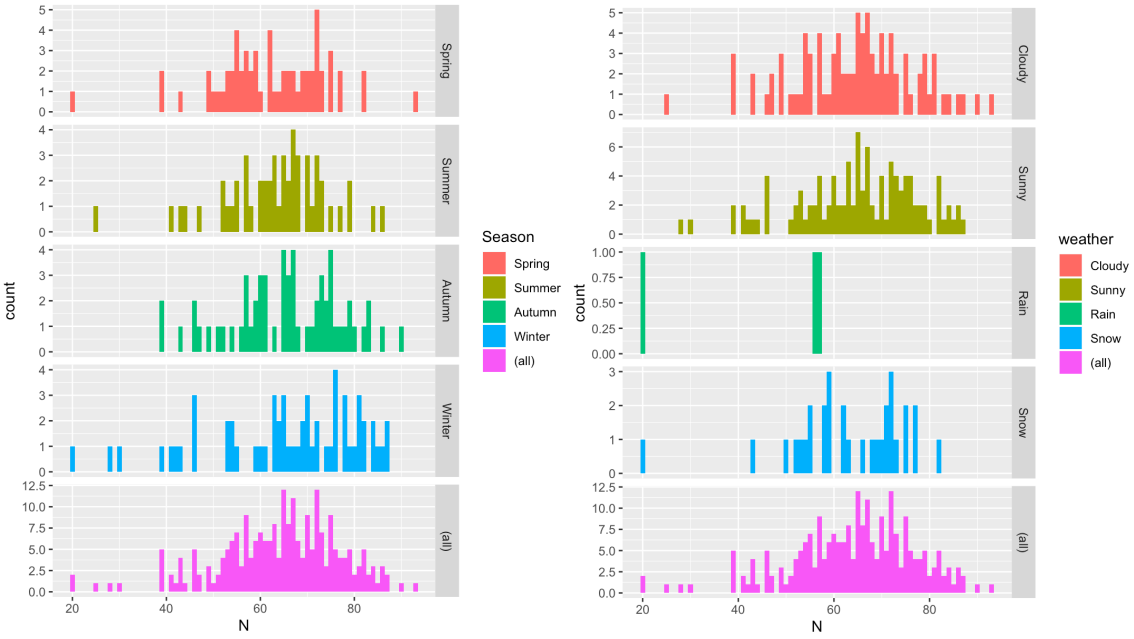

Figure 2 Distribution of N under seasonal groups and weather groups

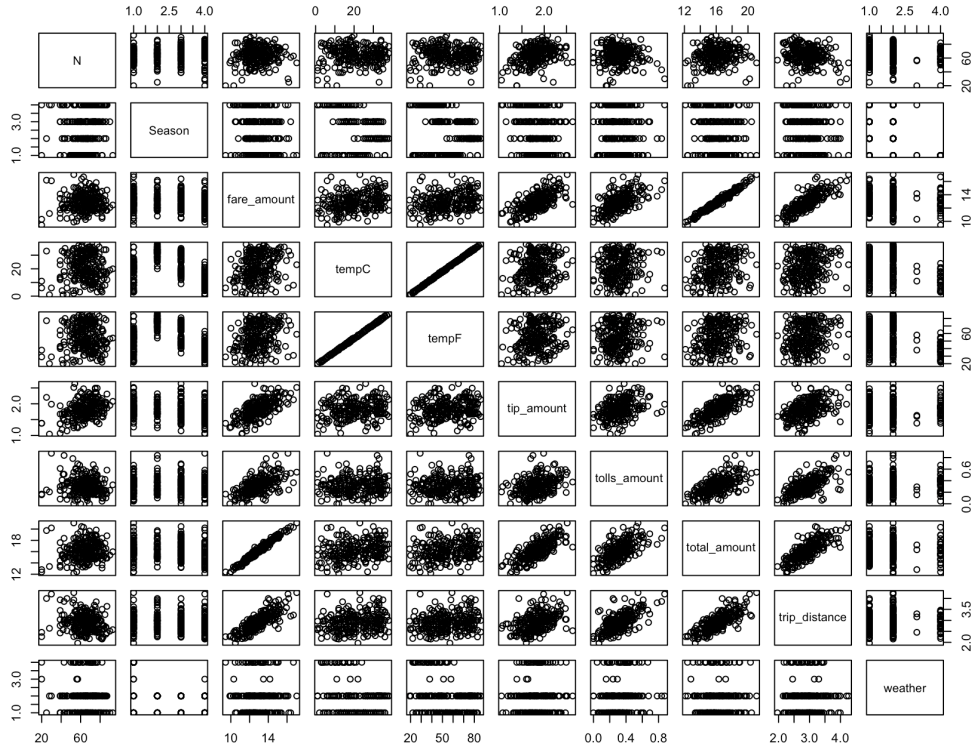

Figure 3 Scatterplot matrix

Next, we produce a scatter plot matrix to catch correlations between the attributes. As figure 3 demonstrated, there exist two groups of data contain strong linear underneath correlation. The first group is trip distance, fare amount and total amount. This fact is easy to understand since higher trip duration usually means a higher fare paid. We

suspect these attributes may have underneath interaction while fitting the model.
The second group is regard temperatures. tempC and tempF have a perfect positive correlation. This is because the only difference between the two attributes is the measurement unit. Hence, the two attributes are indeed telling the same information. In this case, we adopt tempF for fitting the model since tempC is once transformed. We also log transform tempF since the plot shows strong evidence of non-linearity between N and tempF.

To begin with building up a model, we first construct a base model using all the predictors. Then we apply model selection to drop the insignificant attributes. Figure 4 indicates the summary of the base model. We are interested in the Pr(>|z|) column. In short, the p-value here demonstrates the probability of an attribute is not correlated to the response variable. Hence, if the p-value is below a certain significance level, we conclude this attribute is highly likely to be a relevant variable in regards to the response variable. In this case, the significance level is 0.05. For instance, the trip_distance ends up with an extremely low p-value, which means this attribute is highly relevant.

```
> model<-glm(N~.,family = poisson,data)
> summary(model)

Call:
glm(formula = N ~ ., family = poisson, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3877  -0.7912   0.0823   0.8500   2.8726

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.6293491  0.1938406  18.723  < 2e-16 ***
SeasonSummer  -0.0007749  0.0308552  -0.025  0.97996
SeasonAutumn   0.0288381  0.0287069   1.005  0.31510
SeasonWinter   0.1045182  0.0337381   3.098  0.00195 **
fare_amount    0.1063813  0.0886550   1.200  0.23016
tempF          0.0758140  0.0394663   1.921  0.05473 .
tip_amount     0.4051571  0.1012049   4.003 6.25e-05 ***
tolls_amount   0.0262597  0.1250439   0.210  0.83366
total_amount  -0.0692873  0.0856061  -0.809  0.41830
trip_distance -0.2785464  0.0413088  -6.743 1.55e-11 ***
weatherSunny   0.0015119  0.0180881   0.084  0.93339
weatherRain   -0.2467289  0.0894405  -2.759  0.00581 **
weatherSnow    0.0336481  0.0386186   0.871  0.38359
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 635.84  on 232  degrees of freedom
Residual deviance: 425.07  on 220  degrees of freedom
AIC: 1844.1

Number of Fisher Scoring iterations: 4
```

Figure 4 Base model

Regards model selection, we use Akaike Information Criteria (AIC) as the goodness of fit criteria. AIC is a widely-applied selection criteria for glm fitting. In short, it approximates the maximum likelihood and parameter amount to determine the quality of the model. The advantage of AIC is that AIC help reduce overfitting since the risk of overfitting always increase as the number of instance increase. The disadvantage is that AIC cannot determine the quality of the fit. That is, the resulting model may still be a poor model if the input model is not good enough.

```
N ~ Season + fare_amount + tempF + tip_amount + trip_distance +
    weather

               Df Deviance    AIC
<none>              426.22 1841.3
- tempF          1   429.51 1842.6
- weather        3   435.24 1844.3
- fare_amount    1   433.02 1846.1
- Season         3   437.09 1846.2
- trip_distance  1   485.47 1898.5
- tip_amount     1   493.58 1906.7
```

Figure 5 Model after selection

Figure 5 indicates the remaining attributes after model selection. Each turn, we choose transformation which would result in the minimum AIC. We keep repeating the same process until the model converge. Thus, we have the converged model.

```
Call:
glm(formula = N ~ Season + fare_amount + tempF + tip_amount +
    trip_distance + weather + Season:fare_amount + Season:tempF +
    fare_amount:tempF + fare_amount:tip_amount + fare_amount:trip_distance +
    fare_amount:weather + tempF:tip_amount + tempF:weather +
    tip_amount:trip_distance + trip_distance:weather, family = poisson,
    data = data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1730  -0.6698   0.0085   0.7334   2.6726

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 3.2945199  2.1731388   1.516  0.12951
SeasonSummer                3.0763866  1.1207917   2.745  0.00605 **
SeasonAutumn                1.2078697  0.8434140   1.432  0.15211
SeasonWinter                0.3612692  0.9268340   0.390  0.69669
fare_amount                 0.0271660  0.1736070   0.156  0.87565
tempF                      -0.6818961  0.5160984  -1.321  0.18642
tip_amount                  3.1478031  0.5415765   5.812 6.16e-09 ***
trip_distance              -0.9174785  0.2828966  -3.243  0.00118 **
weatherSunny               -0.8724521  0.2788479  -3.129  0.00176 **
weatherRain                13.4681842  8.5607280   1.573  0.11566
weatherSnow                 0.4706953  1.0255816   0.459  0.64627
SeasonSummer:fare_amount   -0.0441132  0.0242834  -1.817  0.06928 .
SeasonAutumn:fare_amount    0.0230257  0.0223269   1.031  0.30240
SeasonWinter:fare_amount    0.0450271  0.0275815   1.633  0.10257
SeasonSummer:tempF         -0.5974965  0.2713641  -2.202  0.02768 *
SeasonAutumn:tempF         -0.3588183  0.2120996  -1.692  0.09069 .
SeasonWinter:tempF         -0.1931830  0.2112565  -0.914  0.36048
fare_amount:tempF           0.1446190  0.0384670   3.760  0.00017 ***
fare_amount:tip_amount     -0.2195808  0.0386838  -5.676 1.38e-08 ***
fare_amount:trip_distance  -0.0557411  0.0203175  -2.744  0.00608 **
fare_amount:weatherSunny   -0.0005787  0.0262498  -0.022  0.98241
fare_amount:weatherRain     0.9779659  0.3677086   2.660  0.00782 **
fare_amount:weatherSnow     0.1136499  0.0474717   2.394  0.01666 *
tempF:tip_amount           -0.5261413  0.1205412  -4.365 1.27e-05 ***
tempF:weatherSunny          0.1850545  0.0612826   3.020  0.00253 **
tempF:weatherRain          -6.7411106  3.3760914  -1.997  0.04586 *
tempF:weatherSnow          -0.3175177  0.2259685  -1.405  0.15998
tip_amount:trip_distance    0.7097907  0.1430101   4.963 6.93e-07 ***
trip_distance:weatherSunny  0.0558859  0.0802967   0.696  0.48643
trip_distance:weatherRain          NA         NA      NA       NA
trip_distance:weatherSnow  -0.2290115  0.1442252  -1.588  0.11231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 635.84  on 232  degrees of freedom
Residual deviance: 277.02  on 203  degrees of freedom
AIC: 1730.1

Number of Fisher Scoring iterations: 4
```

Figure 6 Reduced interaction model

As stated above, we suspect there are undergoing interactions between attribute. In general cases, interaction term

help explain the data. Therefore, we fit the remaining attributes once again but this time with interaction terms. After that, we apply AIC model selection to the interaction model. The resulting attributes are shown in figure 6.

To interpret this result, we first discover the reason of row "trip_distance:weatherRain" being undefined. This factor is consistent with the distribution plot. That is, the insufficiency of data record under rainy weather results in no interactions between attributes. Furthermore, there are significant interactions between fare amount, trip distance and tip amount, which is consistent with the correlation plot. However, the most surprising findings from this summary are the interactions between temperature, fare amount and tip amount. As the p-value in the figure is being low, we have confidence to believe that there are interactions between these attributes and it help explain the data. There are also significant interactions between weather and fare amount. Nevertheless, as stated above, the insufficient data regards raining weather reduced the statistical power of this interpretation.

Next, we check whether this model satisfied the Poisson generalized linear regression assumption. What' s more, we determine whether this model suffers from overdispersion issue.

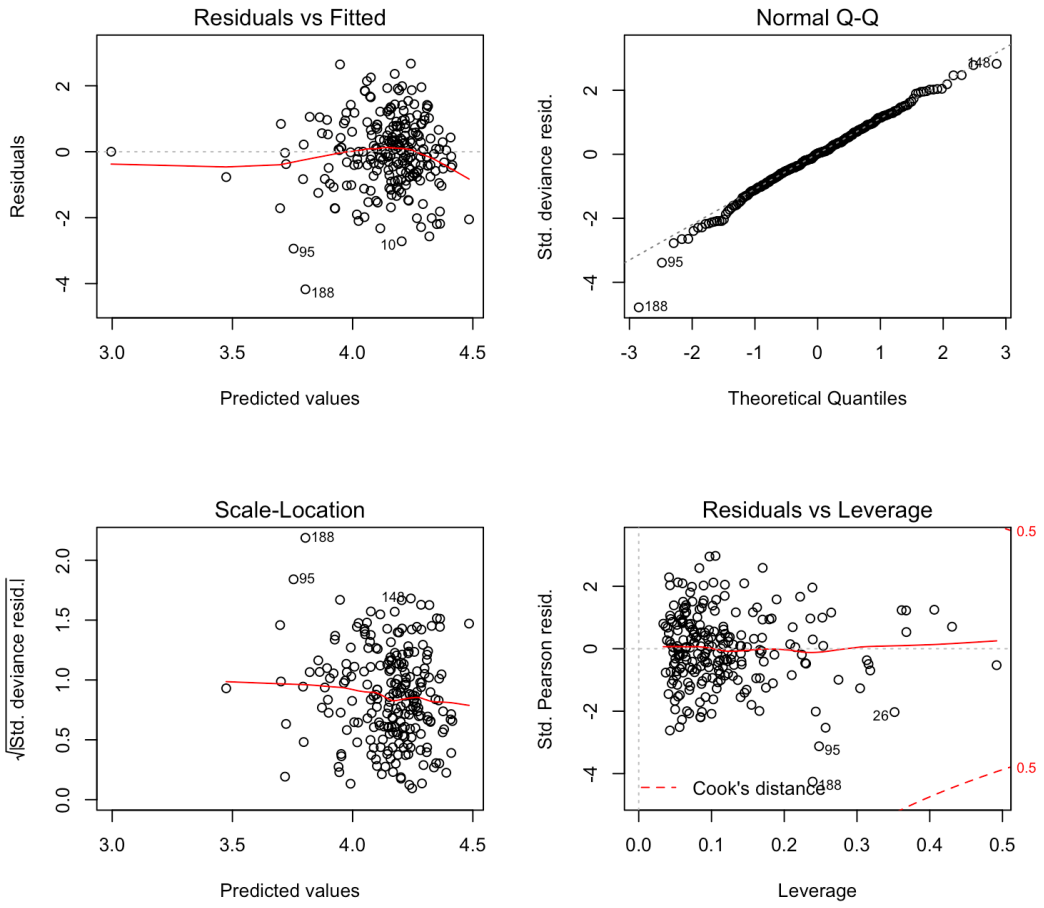

Figure 7 Diagnostic plots for the reduced interaction model

According to figure 7, we can discover some existing issues. Among the four diagnostic plots, the only plot seems normal is the leverage plot at the bottom-right. This plot indicates that there are no potential outliers having large impact on the fit. The reason normal QQ plot concerns us is that there are evidences of outliers on both sides. However, there is no severe issue regards this plot. The rest plots seem problematic. The residual vs fitted concerns us since the points are converged. There are evidences of high fitted values resulting in high residuals. Scale location plot is the direct evidence of this Poisson model having overdispersion issue. In general, we expect the absolute Pearson residual to be approximately 1. In this plot, however, there are signs of fitted value exceeding our expectation. Although the dispersion level is not too strong, indicating the model is not terrible, we still want to classify the exact overdispersion

level.

In short, overdispersion happens when the model variance is much higher than what we expect. Particularly in Poisson cases, we usually expect mean equals to the variance. Nevertheless, Poisson regression always suffers from overdispersion issue in reality applications. Even if we apply quasi-likelihood method or negative binomial distribution, it is still difficult for us to completely avoid overdispersion (Rodríguez, 2019). Figure 8 shows the evidence, we use the same attributes develop a negative binomial data. However, the diagnostic plots are even worse. As we can see, the left side of the QQ plot seems terrible since there is clear sign of abnormality. Leverage plot is worse than the Poisson model, indicating potential influential points on the plot. Scale-location plot still show evidence of overdispersion and the existence of outliers.

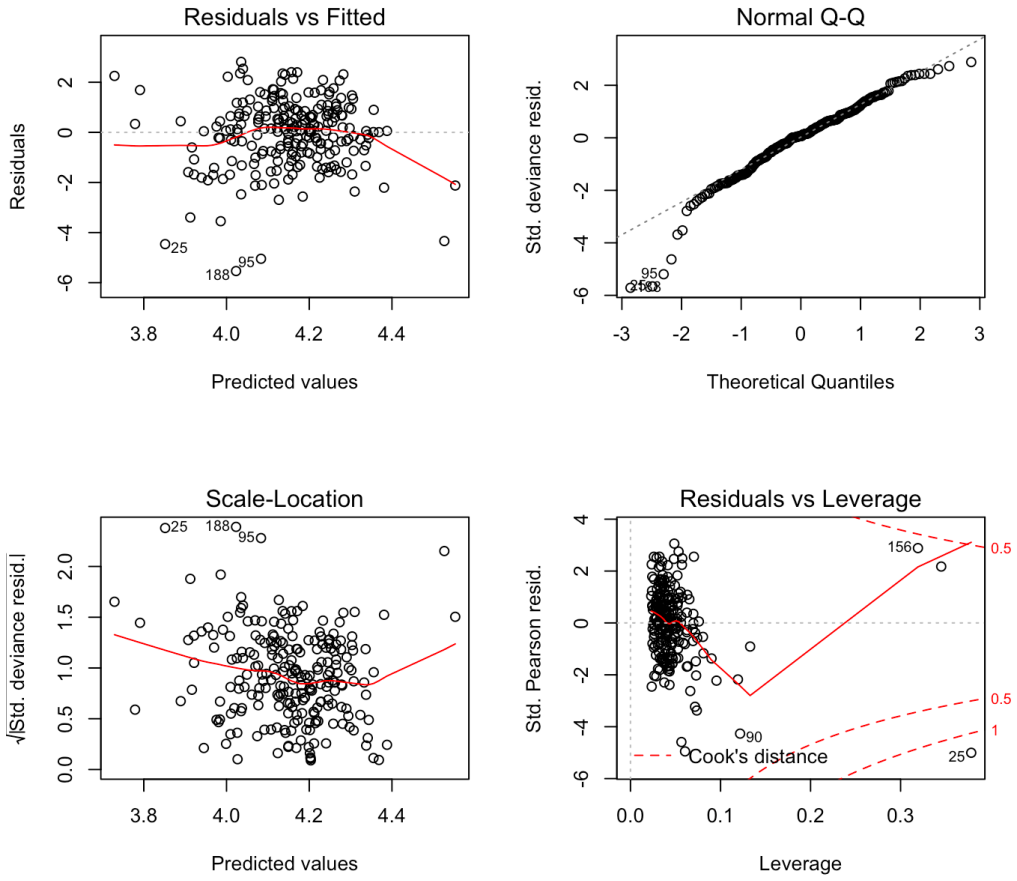

Figure 8 Diagnostic plots for negative binomial glm

Hence, we consider the quality of the negative binomial model is worse than the Poisson model. In purpose of determining the level of the overdispersion regards the Poisson model, we also generated additional diagnostic plots.

There are certain cases where the model fit is heavily influenced by a group of outliers. In which case the dispersion is affected consequentially. As indicated in figure 9, observation 188 and observation 95 are clearly potential influential points. We conclude these two points have a larger impact on the fit than other observations. Nevertheless, as long as the majority observations are normal, we suggest the influence from these two points are limited.

In figure 10, we discuss the overdispersion level by investigating the residual distribution. The left plot demonstrates the Pearson residuals. Theoretically, we expect the absolute value of Pearson' s standardized residuals to variate around 1. As shown in the plot, the majority observations are in between the two red lines. We consider overdispersion does not have severe impact on the model if and only if the residuals variate at an acceptable level. From figure 10, we suggest that the overdispersion level seems acceptable.

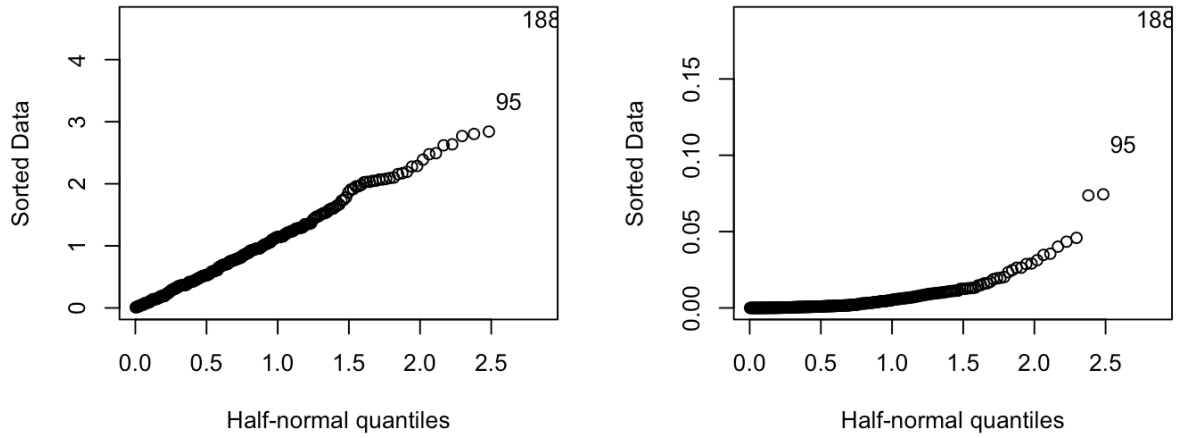

Figure 9 Jack-knife residual and Cook's distance

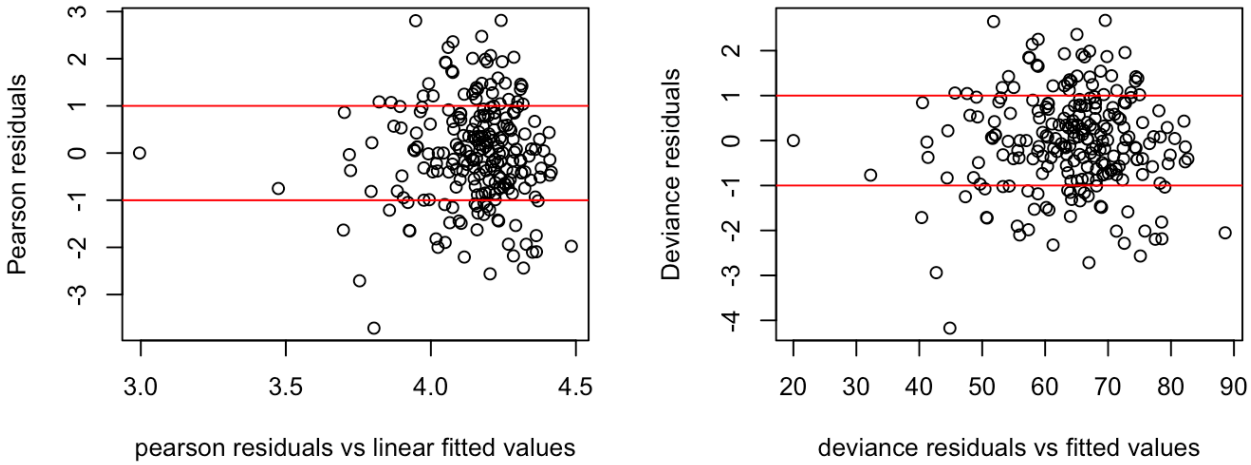

Figure 10 residual plots

To totally accept the model, we also need numerical evidence. There are three value in figure 11. The first value is the chi-square statistic at 0.05 level with a degree of freedom the same as the Poisson model. The second statistic is the model deviance and the third value are Pearson' s chi-square. We expect Pearson' s chi-square of the model and the deviance of the model to be close to chi-square. In this case, we can see the difference is around 40, which is not a large value as the difference is less than 15%. Hence, we consider the overdispersion level of this model is acceptable.

```
> qchisq(0.95, df.residual(model3))
[1] 237.2404
> deviance(model3)
[1] 277.0202
> pr <- residuals(model3,"pearson")
> sum(pr^2)
[1] 271.6941
```

Figure 11 Chi-square deviance and Pearson's statistics

5. Cross Validation and Model Quality

Since we suggest that overdispersion is not a severe issue regards this model, we proceed to develop model and determine model quality by applying cross validation on the development data.

Figure 12 represents the result of 10-fold cross validation using the model from figure 6 on development data. There are three values that we pay attention to, which are root mean square error (RMSE), R-squared, and mean absolute error (MAE). We interpret RMSE and MAE first since both of them measure error magnitude. MAE is measured without counting the directions of the errors, whereas RMSE uses root of the squared errors. Hence, we usually consider RMSE to be superior to MAE. From figure 12, we can see that the value of RMSE is 17.9 and MAE is 14.7, which in both cases are high. In other words, the residual spreads are high. According to the scenario of the data, however, we cannot conclude that the model itself is poor because of high the residual spread. For instance, the response variable is count data. If the response is in large scale, the resulting residual without transformation should be high as well. To give an adequate justification on the model, we need a residual plot.

```
> data$tempF<-log(data$tempF)
> tc <- trainControl("cv",10,savePred=T)
> (fit <- train(N~Season + fare_amount + tempF + tip_amount +
+                 trip_distance + weather + Season:fare_amount + Season:tempF +
+                 fare_amount:tempF + fare_amount:tip_amount + fare_amount:trip_distance +
+                 fare_amount:weather + tempF:tip_amount + tempF:weather +
+                 tip_amount:trip_distance + trip_distance:weather,data=data,method="glm",trControl=tc,family=poisson(link = "log")))
Generalized Linear Model

52 samples
 6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 47, 45, 47, 47, 47, 47, ...
Resampling results:

  RMSE      Rsquared   MAE
  17.99024  0.1724335  14.75914
```

Figure 12 Cross Validation

In figure 13, a fitted vs actual plot is generated for visualization. We can see that the majority data points are distributed around the normal line. In which case looks fine. However, we can see 3~4 outliers from the plot. Those outliers may dramatically affect the measurements of descriptive statistics. Overall, we consider the residual level are reasonable.

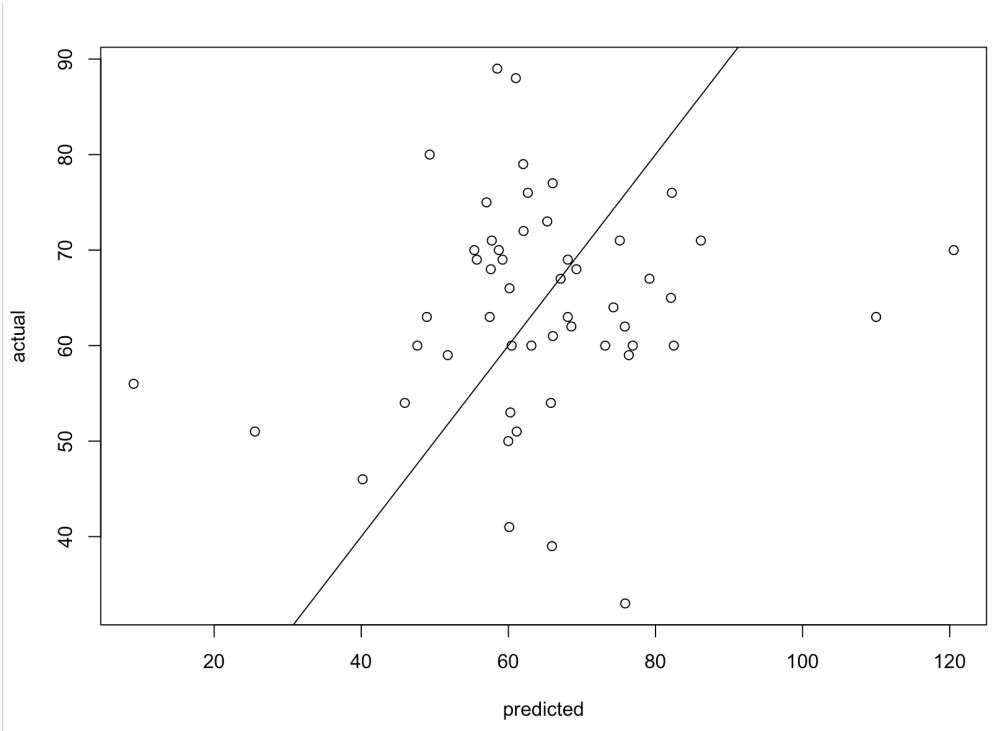

Figure 13 Residual Plot of the predictions

Now back to figure 12 to interpret R-squared. R-squared take a value from 0 to 1, indicating the proportion of

unexplained variance. We consider this statistic measurement is somehow relative to overdispersion issue. In other words, this value reflects on the trace of overdispersion that we discussed above. As we can see, R-square in this case is 0.17. We consider this to be an acceptable level.

To fully comprehend the meaning of this statistic, we consider the assumptions of the generalized linear regression model with Poisson link.

$$E(y_i) = g(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n) + \varepsilon_{ii}$$
$$y_1 \sim Poisson(\lambda_1), \qquad y_2 \sim Poisson(\lambda_2), \ldots\ldots$$

In glm, we assume the population of the response variable y follows a Poisson distribution with a Poisson parameter λ. In most cases, we assume the error parameter ε_{ii} to be independent and identically distributed. Once this assumption is not satisfied, we conclude that there are undergoing dependent variables that are not explained. For instance, we may need an extra attribute or a term $\beta_{n+1}x_{n+1}$ to reduce ε_{ii}. In the worst cases, different y may come from different Poisson distribution. For instance, dramatic difference between λ_1 and λ_2. Note that there is a trade-off between overfitting and low r-squared. The more the data explained, the riskier we have overfitting issue. The best outcome is to leave r-squared at a balanced level that controls overfitting risk as well as explaining the majority of the data. That is the reason we suggest 0.17 in this model is acceptable.

To further improve our model and solve the above issue, there are two options for us to consider. The first option is finding significant external resources that can help explain the response variable. The second option is to apply random effect models (Penny & Holmes, 2003).

Now the final model is generated, we can interpret the parameters in English. Despite of the interaction terms, we can conclude several clues that may cause a high tip frequency. Firstly, people are highly unlikely to pay tips in Winter in comparison to other seasons. Then, high basic fare amount usually will not result in tips income. Also, customers prefer to pay every low tips regularly but not paying high tips seldomly. Furthermore, high trip disrtances usually results in tips income. Reagrds the weather factor, as stated above, there are data insufficiency issues. However, in this case, raining weather usually results in high tip frequency in comparison to other weathers.

```
Call:  NULL

Coefficients:
                (Intercept)                SeasonSummer                SeasonAutumn
                   6.762687                    2.688129                    2.140216
               SeasonWinter                 fare_amount                       tempF
                  -1.412075                    0.007388                   -2.514679
                 tip_amount               trip_distance                weatherSunny
                  -3.155727                    3.220892                   -0.691812
                weatherRain                weatherSnow `SeasonSummer:fare_amount`
                  13.263992                          NA                   -0.057328
```

Figure 14 Coefficients of the final model without interaction terms

6. Conclusion

Overall, generalized linear model with Poisson link is applied to model the tip frequency of the NYC yellow taxi data. To start with, we applied a set of methods to preprocess the data, which including handling missing value and data aggregation. Furthermore, we applied log transformation to some of the attributes to improve the modelling quality. As a result, the data is capable for Poisson regression modelling.

During model development phrase, we first fit a base model using all the attributes that we considered relevant. However, it is found that there are certain attributes not being significant. We choose AIC as the goodness of fit criteria to remove attributes. At the end of model selection, we gain a decent Poisson model.

Nevertheless, we concern about the overdispersion issue. We determine the overdispersion level by visualizing diagnostic plots. Eventually, we found that the glm assumptions are satisfied and the overdispersion level is acceptable. What' s more, during cross validation, we found that the R-square statistic is linked to the overdispersion issue. The reason of overdispersion may due to the proportion of unexplained variance. Nevertheless, we suggest that the level of dispersion do not damage the model quality severely according to the trade-off between overfitting and r-squared. At the end, we discovered several clues which can explain the tip frequency. However, the insufficient data volume in weather attribute affect the model accuracy and interpretation. To further improve the model, we consider to apply random effect models or find more significant external data source.

7. Related Works

Todd W.. 2019. Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. [ONLINE] Available at: https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/. [Accessed 30 August 2019].

Penny, W. and Holmes, A. (2003). Random-Effects Analysis. [online] Fil.ion.ucl.ac.uk. Available at: https://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/pdfs/Ch12.pdf [Accessed 7 Sep. 2019].

grumble10, V. (2019). Checking (G)LM model assumptions in R. [online] biologyforfun. Available at: https://biologyforfun.wordpress.com/2014/04/16/checking-glm-model-assumptions-in-r/ [Accessed 7 Sep. 2019].

Tony, M. (2008). EVALUATION OF GENERALIZED LINEAR MODEL ASSUMPTIONS USING RANDOMIZATION. [online] Mun.ca. Available at: https://www.mun.ca/biology/dschneider/b7932/B7932Final10Dec2008.pdf [Accessed 7 Sep. 2019].

statmath (2019). Introduction to Generalized Linear Models. [online] Statmath.wu.ac.at. Available at: http://statmath.wu.ac.at/courses/heather_turner/index.html [Accessed 7 Sep. 2019].

UCLA (2019). Negative Binomial Regression | R Data Analysis Examples. [online] Stats.idre.ucla.edu. Available at: https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/ [Accessed 7 Sep. 2019].

Rodríguez, G. (2019). Models for Over-Dispersed Count Data. [online] Data.princeton.edu. Available at: https://data.princeton.edu/wws509/r/overdispersion [Accessed 7 Sep. 2019].

Lillis, D. (2019). Generalized Linear Models in R, Part 6: Poisson Regression for Count Variables - The Analysis Factor. [online] The Analysis Factor. Available at: https://www.theanalysisfactor.com/generalized-linear-models-in-r-part-6-poisson-regression-count-variables/ [Accessed 7 Sep. 2019].