New York uber taxi price analysis

Zishan Cheng

16/12/2020

Abstract

Uber is raised quickly in recent year due to its efficiency. Unlike the traditional taxi economy, online reservation is provided. In this project, we would like to analysis the average price of uber in different time of a day in New York city by ratio and regression estimation methods. Based on our analysis, we can tell that the average price per mile is approximate $7.03/mile; the greatest varibility of the price happens at rush hour, especially 9 am.

Keywords: uber, stratified, ratio, regression

Github: https://github.com/carol-png/304project

Introduction

Uber, is an American multinational ride-hailing company providing services that include peer-to-peer ride sharing, food delivery (Uber Eats), and etc. The company is based in San Francisco and has operations in over 785 metropolitan areas worldwide. Its platforms can be accessed via its websites and mobile apps (Dara 2020). Compared with the traditional taxi economy, reservation can be set online to make the , additionally, the price of uber is fluid via the demand.

In this project, we would like to investigate unit uber prices (price per mile) in New York at different time of a day. This research could give a guidance that at which hour, uber drivers are more likely to earn above the average; at which hour, uber drivers are less likely to earn above the average. To uncover it, ratio and regression estimation under stratified sampling (each hour as), with focuses on point estimation and confidence intervals is applied for this analysis.

Methodology

Data

Our data "uber" is collected via `dashshader` website (Rougier 2013). It involves all background data source in New York city (mainly Manhattan Island), on Jan 2015. The variables our data containing are 'pickup location x', 'pickup location y', 'total miles', 'total payment' and

etc. The **target population** of our study is all activated uber taxi drivers in New York in the past 11 years (uber is founded in 2009) and further years. The **sample population** is activated uber taxi drivers in New York in Jan, 2015 and the sample size, $N = 10679307$ (however, such data set is around 750 MB which is almost impossible to upload on *Quercus*. To accommodate the size of file less than 25 MB, we randomly picked 1e6 observations).

The target variable is 'price per mile', defined as

$$\text{price per mile} = \frac{\text{total payment}}{\text{total miles}}$$

Since all variables are numerical, we can remove observations containing any `NA`, `NaN`, and infinite values. Then, let us visualize the histograms of variable 'price per mile' and 'total miles'.

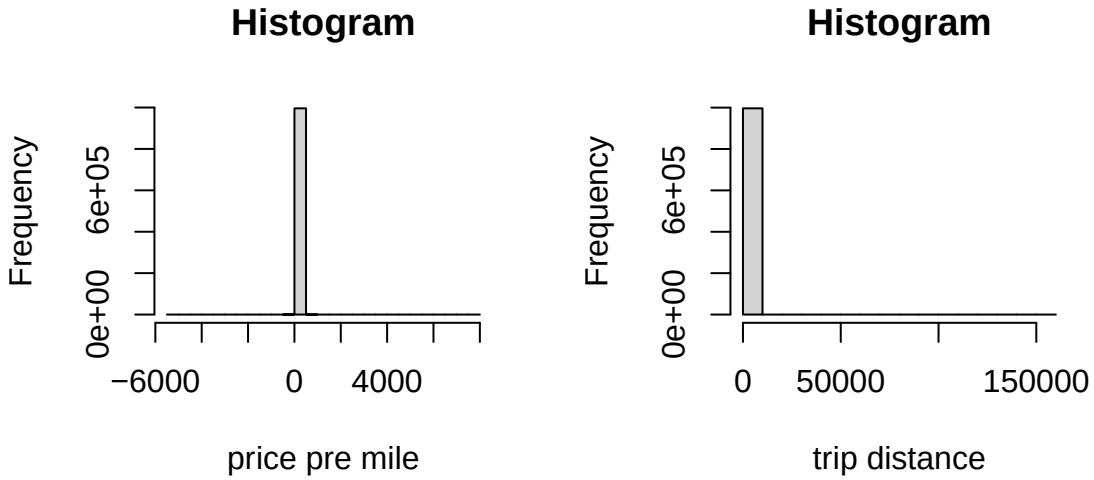

Figure 1: Price and Trip Histograms

Through this Figure, we would tell that there are extremely large values, even close to `6e5` and some negative values in these two variables. We suppose it is caused by incorrect recordings of uber system. To better investigate the relationship, such outliers are removed. In our case, we set the boundary of 'price per mile' and 'total miles' as $(0, 20]$ and $(0, 30]$, respectively.

The Figure 2 illustrates the overall 'price per mile' in Manhattan Island. Coordinate x and y represent latitude and longitude, respectively. The shade of color represents the 'price per mile'. If the color in one region is bright which means that the 'price per mile' at this pick up location is higher. Obviously, passages at the left bottom corner of Manhattan Island are less willing to pay tips than other regions.

In our analysis, we will choose trip distance as our auxiliary variable. However, the left one of Figure 3 shows that the correlation between 'total miles' and 'price per mile' is negative and they are non-linear. In both ratio and regression stratified sampling, we assume the

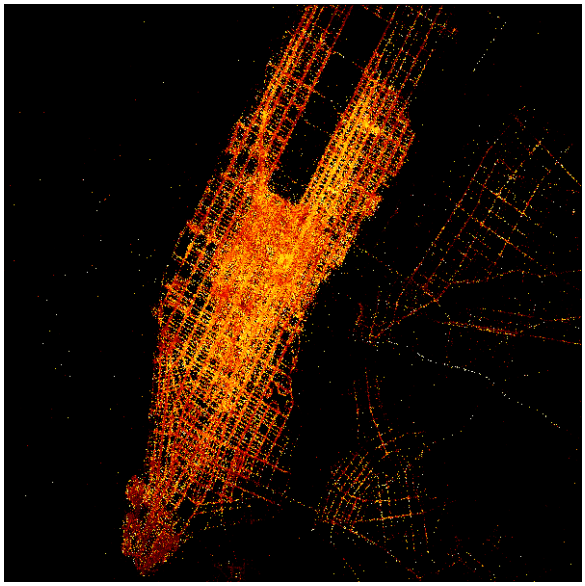

Figure 2: Manhattan Heat Map (colour represents the unit price)

relationship is linear. Thus, we would like to perform power transformation (Box and Cox 1964) on these two variables. Based on the shape, we can decrease the power of x and y,

$$x^* = log(x)$$

$$y^* = log(y)$$

The right one of Figure 3 shows that the relationship of 'total miles' and 'price per mile' is almost linear.

Model

Ratio estimation and regression estimation (Singh and Mangat 2013) are often performed when the target variable is hard to obtain but the auxiliary variable is relatively easier to obtain. Additionally, the relationship between target variable and auxiliary variable should be strong.

In our case, the 'total miles' can be observed by uber drivers before the start of the trip, while, the 'price per mile' largely depends on traffic and the generosity of passages. Due to the strong relationship between 'price per mile' and 'total miles', it is suitable to use ratio estimation and regression estimation in our case.

Ratio estimation under stratified sampling

- Separate Ratio Estimator Method:

  Ratio is defined as

$$\hat{R_{SR}} = \frac{\bar{y_h}}{\bar{x_h}}$$

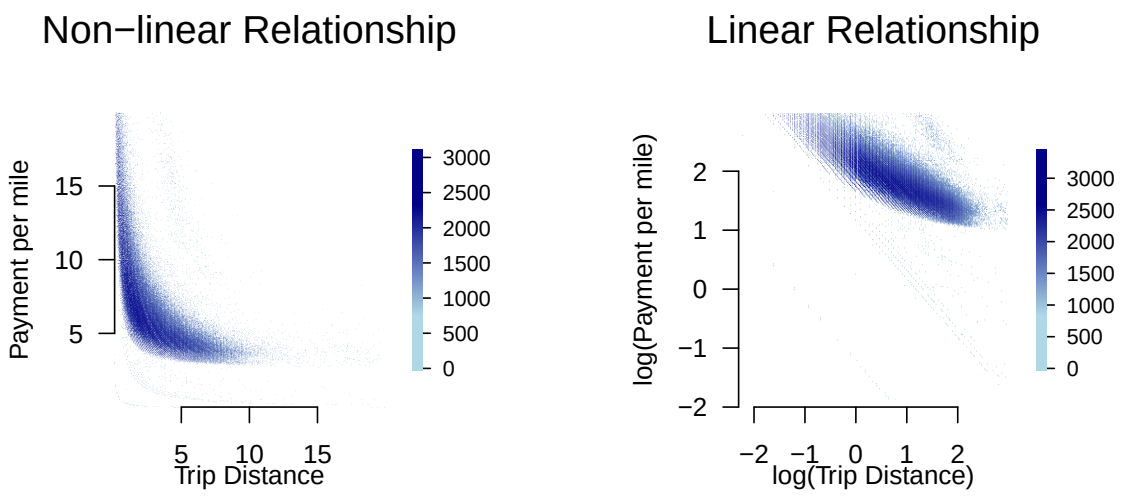

Figure 3: Relationship between Trip Distance and Payment per Mile

where $\bar{y_h}$ represents the average price per mile, $\bar{x_h}$ represents the average total miles.

The strata mean is

$$\hat{y_{Uh}} = \hat{R_{SR}} x_{Uh}$$

where x_{Uh} represents average total miles at given hour h

The separate ratio estimate of $\hat{y_U}$ is

$$\hat{y_U} = \sum_{h=1}^{H} w_h \hat{y_{Uh}}$$

where $w_h = \frac{N_h}{N}$ represents weights of such strata; N_h is the sample population size of hour h and N is the sample population size; y_{Uh} represents average price per mile at given hour h.

The variance of y_u is defined as

$$V(y_u) = \sum_{h=1}^{H} [w_h^2 \frac{N_h - n_h}{N_h n_h (n_h - 1)} \sum_{i=1}^{n_h} (y_{hi} - \hat{R_{SR}} x_{hi})^2]$$

where n_h is the resample size at hour h.

- Combined Ratio Estimator Method:

  Based on weighted least square

$$\hat{R_{CR}} = \frac{y_{st}}{x_{st}}$$

where $y_{st} = \sum \frac{N_h}{N} \bar{y}_h$ and $x_{st} = \sum \frac{N_h}{N} \bar{x}_h$ thus

$$\hat{\mu_y} = \hat{R_{CR}} \mu_x$$

where μ_x is the estimate the population mean of `total miles`

The variance is

$$V(y_u) = \sum_{h=1}^{H} [w_h^2 \frac{N_h - n_h}{N_h n_h (n_h - 1)} \sum_{i=1}^{n_h} (y_{hi} - \hat{R_{CR}} x_{hi})^2]$$

Regression estimation under stratified sampling

- Separate Regression Estimator Method:

  Based on origin least square, $\hat{a}_h$ and $\hat{b}_h$ can be obtained, thus

  $$\hat{y_{Uh}} = \hat{a}_h + \hat{b}_h x_{Uh}$$

  estimate of $\hat{y_U}$ is

  $$\hat{y_U} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{y_{Uh}}$$

  The variance is

  $$Var = \sum_{h=1}^{H} [w_h^2 \frac{1 - w_h}{n_h (n_h - 2)} SSE_h^2]$$

  where SSE_h is the error sum of squares at hour h.

- Combined Regression Estimator Method:

  The $\hat{y_U}$ can be obtained by

  $$\hat{y_U} = \hat{y_{st}} + b_c (\hat{x_U} - \hat{x_{st}})$$

  where $b_c = \frac{\sum_{h=1}^{H} c_h b_h}{\sum_{h=1}^{H} c_h}$ and $c_h = w_h^2 \frac{(1 - w_h)}{n_h} s_{xh}^2$; $\hat{y_{st}} = \sum_{i=1}^{H} w_h \bar{y}_h$ and $\hat{x_{st}} = \sum_{i=1}^{H} w_h \bar{x}_h$.

  The variance is

  $$Var = \sum_{h=1}^{H} [(\frac{N_h}{N})^2 \frac{1 - f_h}{n_h (n_h - 2)} \sigma_h^2]$$

  where σ_h^2 is

  $$\sigma_h^2 = \sum_{i=1}^{n_h} [(y_{hi} - y_h) - b_c (x_{hi} - x_h)^2]$$

Results

Table 1: Statistical Summary (log)

	Mean	Standard.Variance
Separate Ratio	1.9660	0.0106
Combined Ratio	1.9525	0.0102
Separate Regression	1.9512	0.0005
Combined Regression	1.9512	0.0005

Table 2: Statistical Summary

	Mean	Standard.Variance
Separate Ratio	7.14	13.42
Combined Ratio	7.05	11.72
Separate Regression	7.04	0.61
Combined Regression	7.04	0.61

Table 1 illustrates the statistical summary (log) of those four models and Table 2 shows the transformed statistical summary, we can tell that

- the mean of all four estimates is around 7.03 which means that in general, the unit price of uber is around \$**7.03** (exp(1.95)) per mile in Manhattan, in Jan, 2015.
- the standard variance of regression estimators is much smaller than that of the ratio estimators. Thus, compared with the ratio estimation, regression estimation is more robust.
- the difference of the `separate` and `combined` strategy is negligible.

Since the difference of strategy `separate` and `combined` is not obvious, we would only talk about one of each (i.e. Combined Ratio Estimation and Combined Regression Estimation)

In Figure 4, the color of the bars represents the number of observations in each strata (N_h). The dark colour (i.e. purple) represents less N_h and the bright colour (i.e. yellow) represents high N_h. The length of the bar represents the confidence interval $\mu \pm 1.96$sd. The horizontal solid line represents the estimate mean of the overall 'price per mile' (both methods give very similar intervals). We could tell that

1. Combined Ratio estimator gives way larger variance than the Combined Regression Model.
2. The variance at daily time (from 8 to 19) is larger than that at night (from 19 to 8).
3. The number of observations from 6 to 8 is the most and passages in the morning (5) are less activated.
4. In the day time, the payment per mile at 5 or 6 has lowest variance. Conversely, from 8

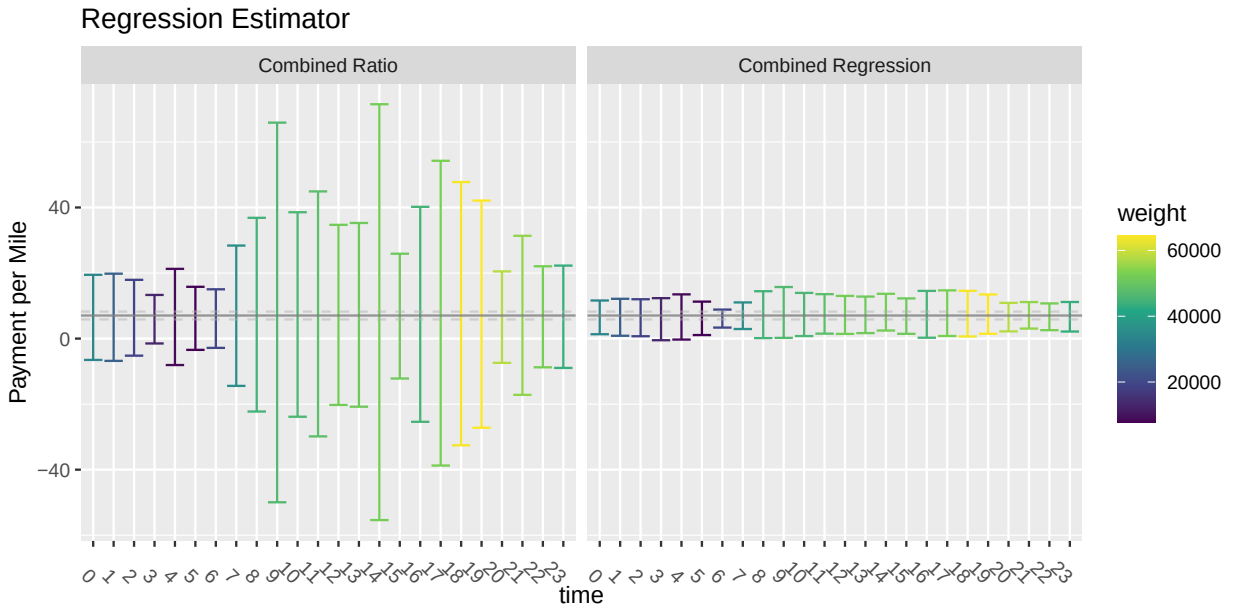

Figure 4: Regession and Ratio Estimation Comparison

to 11, the variances are the largest which makes sense since in the morning rush hour, uber users are more likely to increase the rate to be on time to work.

Discussion

Conclusion

In conclusion, based on the New York uber data, the mean price is around $7.03 per mile. We construct four models, separate/combined ratio estimation and separate/combined regression estimation based on the stratified sampling. The difference between separate and combined model is negligible, however, the difference between ratio and regression estimation is great, especially in standard deviation. The standard deviation of ratio estimation is around 12, however, in regression estimation, the standard deviation is only 0.6. We may conclude that compared with the ratio estimation, the regression estimation is more stable.

In addition, we found that, in general, people at 18 to 20 are more likely to call uber than other time. From 8 to 11, the price pre mile varies a lot which shows that people tend to pay higher price (give more tips) at this time. In contrast, from 0am to 6am, the price varies little (the number of passages is less as well) showing that the uber drives are more likely to get less tips than the morning time.

Weekness

Our model has some weakness as well

- **Data**: In our case, the **study error** is mainly combined by two parts, *technique error*

and *seasonal impact*.

  - Through the data, we can find some tips are extremely large, even close to 4 million which is obviously impossible. It may be caused by uber system incorrect recording or some other technique issues. We need to set a boundary of the tip to remove such bugs. The choices are not arbitrary and reasonable at some manner. Since *Manhattan Island is 22.7 square miles in area, 13.4 miles (21.6 km) long and 2.3 miles (3.7 km) wide, at its widest (near 14th Street).*, in general, a possible trip distance in Manhattan Island should be no more than 30 miles. However, there could be some unusual trips larger than 30 miles (i.e. travel between the Manhattan Island and its surroundings). All these recordings are omitted manually.
  - The *seasonal impact* would be that since the collected data is only in Jan (presumably winter), the tips may also vary via seasons (i.e. in summer, people are less likely to call uber, etc) which we cannot obtain.

- **Model**: we only picked two main models (ratio and regression). However, these two models have their own drawbacks.
  - Ratio estimation: ratio estimates are biased and corrections must be made. In our case, we do not adjust the biased issue.
  - Regression estimation: it has assumptions such that the residuals are normally distributed, the residuals should be independent, etc. In this project, we do not really check the model adequacy which should be done in the next move.

The Next Steps

After fitting the regression model, we should do some residual checks, such as Augmented Dickey-Fuller test (Dickey and Fuller 1979) for stationary, Box-Pierce test (Box and Pierce 1970) for the independence, Jarque-Bera test (Jarque and Bera 1980) for the normality, etc.

Additionally, we should try to use some other models, such as `GLMM` (generalized linear mixed model) (McCulloch and Neuhaus 2014) to fit grouped data (set `hour` as the random effect).

References

Box, George EP, and David R Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society: Series B (Methodological)* 26 (2): 211–43.

Box, George EP, and David A Pierce. 1970. "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models." *Journal of the American Statistical Association* 65 (332): 1509–26.

Dara. 2020. *Uber Company*. https://www.uber.com/ca/en/community/.

Dickey, David A, and Wayne A Fuller. 1979. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association* 74 (366a): 427–31.

Jarque, Carlos M, and Anil K Bera. 1980. "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals." *Economics Letters* 6 (3): 255–59.

McCulloch, Charles E, and John M Neuhaus. 2014. "Generalized Linear Mixed Models." *Wiley StatsRef: Statistics Reference Online*.

Rougier, Nicolas. 2013. "Shader-Based Antialiased Dashed Stroked Polylines."

Singh, Ravindra, and Naurang Singh Mangat. 2013. *Elements of Survey Sampling*. Vol. 15. Springer Science & Business Media.