

Anubhav Hazra

Anubhav_Hazra_Dissertation.pdf

 St. Xavier's College (Autonomous) Kolkata

Document Details

Submission ID

trn:oid:::3618:89091237

Submission Date

Apr 1, 2025, 11:54 PM GMT+5:30

Download Date

Apr 1, 2025, 11:56 PM GMT+5:30

File Name

Anubhav_Hazra_Dissertation.pdf

File Size

564.1 KB

23 Pages

4,247 Words

20,570 Characters

0% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report





- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 15 words)
- Abstract
- Methods and Materials

Custom Section Exclusions




1 Section Titles, 4 Keywords

Section title	No. of Section Starters	Section Starters
"Acknowledgements"	4	<div>Acknowledgements</div> <div>Acknowledgement</div> <div>Acknowledgment</div> <div>Acknowledgments</div>

Match Groups

-  **1** Not Cited or Quoted 0%
Matches with neither in-text citation nor quotation marks
-  **0** Missing Quotations 0%
Matches that are still very similar to source material
-  **0** Missing Citation 0%
Matches that have quotation marks, but no in-text citation
-  **0** Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

- 0%  Internet sources
- 0%  Publications
- 0%  Submitted works (Student Papers)

Integrity Flags





0 Integrity Flags for Review

No suspicious text manipulations found.




Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

-  **1** Not Cited or Quoted 0%
Matches with neither in-text citation nor quotation marks
-  **0** Missing Quotations 0%
Matches that are still very similar to source material
-  **0** Missing Citation 0%
Matches that have quotation marks, but no in-text citation
-  **0** Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

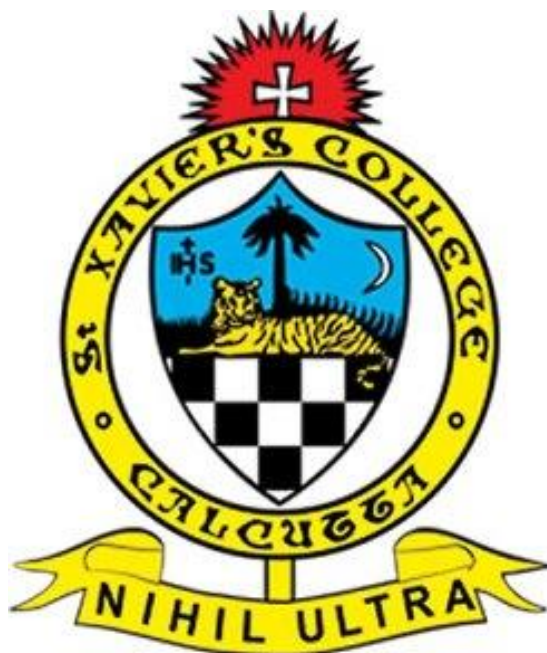
- 0%  Internet sources
- 0%  Publications
- 0%  Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- 1** Internet

www.coursehero.com <1%



DEPARTMENT OF STATISTICS

ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA.

Beyond 'Expectation': analyzing cab fares through a
Quantile regression approach.

Name: Anubhav Hazra

Roll no.: 466

Registration no.: A01-1112-0736-22

Supervisor: Prof. Rahul Roy

Session: 2022-2025

DECLARATION

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Anubhav Hazra

Anubhav Hazra

April 2025

ACKNOWLEDGEMENT

I would like to take this opportunity to thank my dissertation supervisor Prof. Rahul Roy, without whose guidance, suggestions and encouragement I would not have been able to complete this project.

I would like to thank all the other professors at the Department of Statistics, St. Xavier's College, Kolkata for helping me develop a passion and interest for Statistics which further motivated me immensely in doing my research work.

I would also like to thank my family and friends for their constant support and encouragement towards all my endeavors.

CONTENTS

	Page
Introduction	5
Data Description	7
Exploratory Data Analysis	8
Methodology	12
Results and Discussion	15
Goodness of Fit	20
Limitations	21
Conclusion	22
References	23

INTRODUCTION

With the day to day hassle of work and travel in cities, something that has prominently risen into importance are cab services, namely Uber, Rapido and Ola, especially as far as at least India is concerned. Their aid of a quick, private and convenient travel often compels us to avail such a service as compared to a possibly tiring experience of public transportation. However, the word “convenient” is pretty subjective in this comparison. Whatever be the disadvantages of public transportation, it would always be the more affordable and thus the more preferred option of the two for most people. The sky-soaring prices of cab ride fares are sometimes about 30 times the total fare of public transport for the same distance to travel even on the same day and time.

That being said, what is actually more interesting and strange is the variation in these fares. Some of them, though would make complete sense to anyone, considering the various other factors which cause the fare to vary. For example, the same distance of travel would cost much less at 8 am in the morning than 12 am at night, when there is obviously a scarcity of available vehicles. Also, the cab fare could spike up in the middle of the ride due to heavy traffic, poor weather conditions, or a sudden change in the route of travel. Often, festive occasions influence the prices highly, and so on.

This dissertation mainly aims at figuring out the importance of the different factors that influence a cab ride’s fare, over different values of the fare, i.e. when the fares are usually low, the relative importance of some predictor might be more, as compared to when the fares would be generally high. As a result of which, a lot of factors may arise due to which the classical linear regression model might not give us a full picture of the analysis, possibly resulting in an unsatisfactory fit, some of the reasons being presence of multiple outlier values or a change in the variation of the cab fares conditioning on values of the different predictors, i.e. heteroscedasticity. The classical model gives us the average effect, or more specifically the conditional mean of the response, given the predictor values. We might want to get an idea of the corresponding conditional quantiles, thereby helping us to understand the effect of the different covariates on varying quantiles of the cab fare (response). This would help us sketch out a better picture of the varying distributions, alongside relaxing the classical model’s assumption of homoscedasticity for providing a good fit. This procedure would mean that we would have separate regression lines corresponding to each quantile, the parameter estimates for which are to be obtained by minimizing the sum of skewed (weighted) absolute deviations from the response values, where the skewness/weights of the deviations are set based on the particular quantile whose regression line we intend to fit.

Following a brief theoretical introduction and explaining the intuition behind quantile regression and its various advantages over the classical model, this dissertation analyses a dataset of New York City cab trip data from the year 2018 (sourced from Kaggle) by:

1. Fitting of decile (10%) lines on the response values, i.e. a regression line representing each decile.

2. Monitoring how the model parameters (coefficients of the predictors) vary over the deciles, and thus compare them with the parameter estimates yielded by the classical model.
3. Concluding on the importance of the predictors, thereby giving us a better and concise idea of how the cab fare would be varying in different conditions.

DATA DESCRIPTION

We have sampled data on cab rides for New York City of the year 2018 (sourced from kaggle.com) ,consisting of 3120 rows and 7 columns, namely– “distance”, “duration”, “extra”, “tolltax”, “day”, “time” and “fare”. We are interested in studying the relation between the cab ride fare (i.e. “fare”, our study variable) and 6 other covariates.

Source: https://www.kaggle.com/datasets/neilclack/nyc-taxi-trip-data-google-public-data?resource=download&select=original_cleaned_nyc_taxi_data_2018

Explanations:

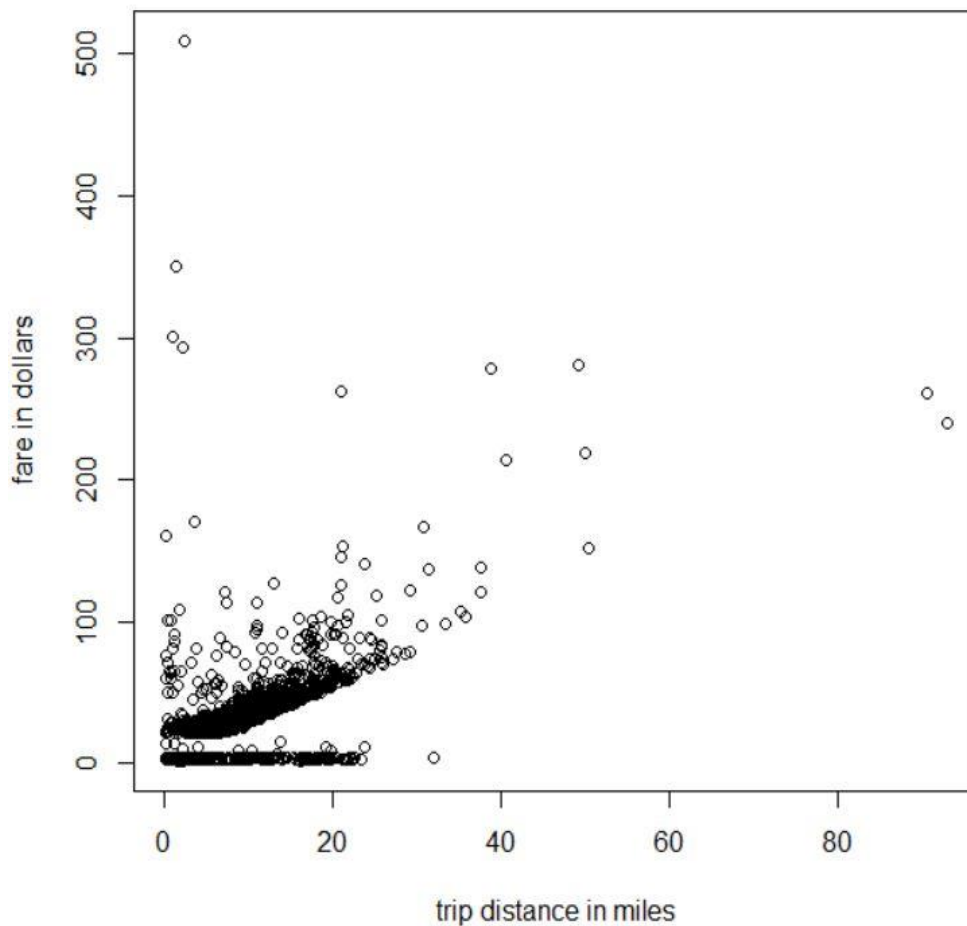
- 1."distance" =The elapsed trip distance in miles reported by the taximeter.
- 2."duration"= duration in minutes
- 3."extra" =Miscellaneous extras and surcharges. This only includes the rush hour and overnight charges, or spike in the fare due to poor weather or heavy traffic.
- 4."tolltax" =Total amount of all tolls paid in the trip.
5. "day": Dummy variable. 1 denotes weekday, 0 denotes weekend. Weekend refers to sat and sun
6. "time": Dummy predictor represented by 1 or 0.
6am to 6pm: 1 (day)
6pm to 6am: 0 (night)
7. "fare"=The total amount charged to passengers (in USD)

Reasons why these predictors are suitable:

Distance and duration of the ride make absolute sense as longer cab rides are expected to have a higher fare. Pricing mechanisms may vary based on whether a day is a weekend or not, as this influences the number of other cab rides being booked in a certain region. Similarly, for the time of the day, it might be easier i.e. more availability of cabs at some point of the day, as compared to other times. A lot of other predictors from the initial source data have not been considered, due to them being comparatively inappropriate with respect to our interest.

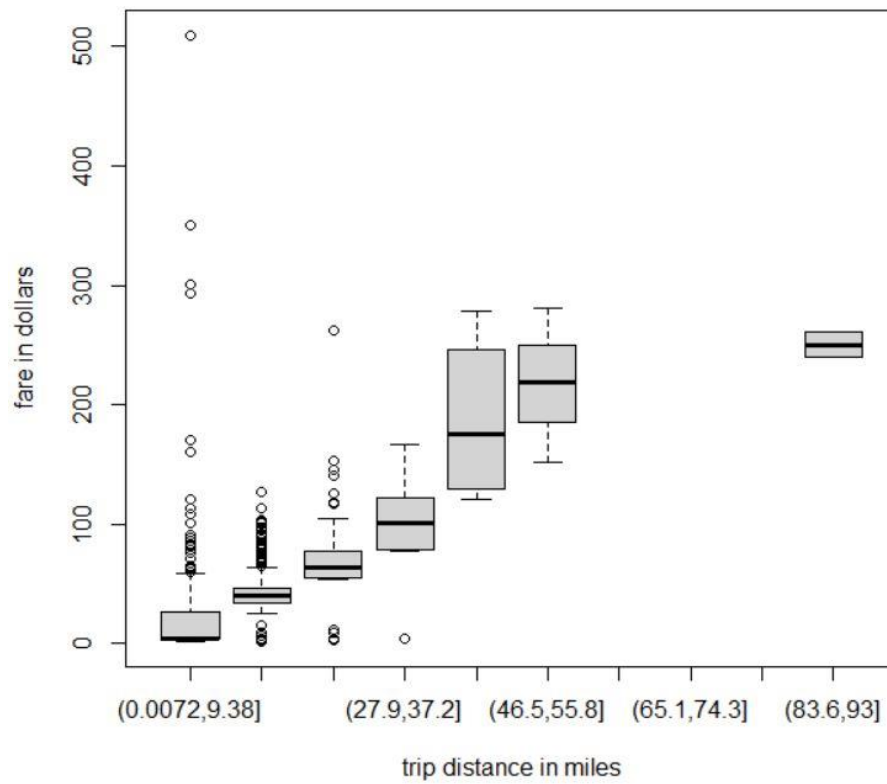
EXPLORATORY DATA ANALYSIS

Before diving into any statistical analysis, it is better if we get a crude picture of the data we are dealing with. This would help us understand better, the generic trends and patterns, important statistical features like the central tendency, dispersion, skewness or kurtosis. We first look at the relation between our response, i.e. the cab fare and some of the predictor variables individually. Let us start with a scatterplot between cab fare and the trip distance.

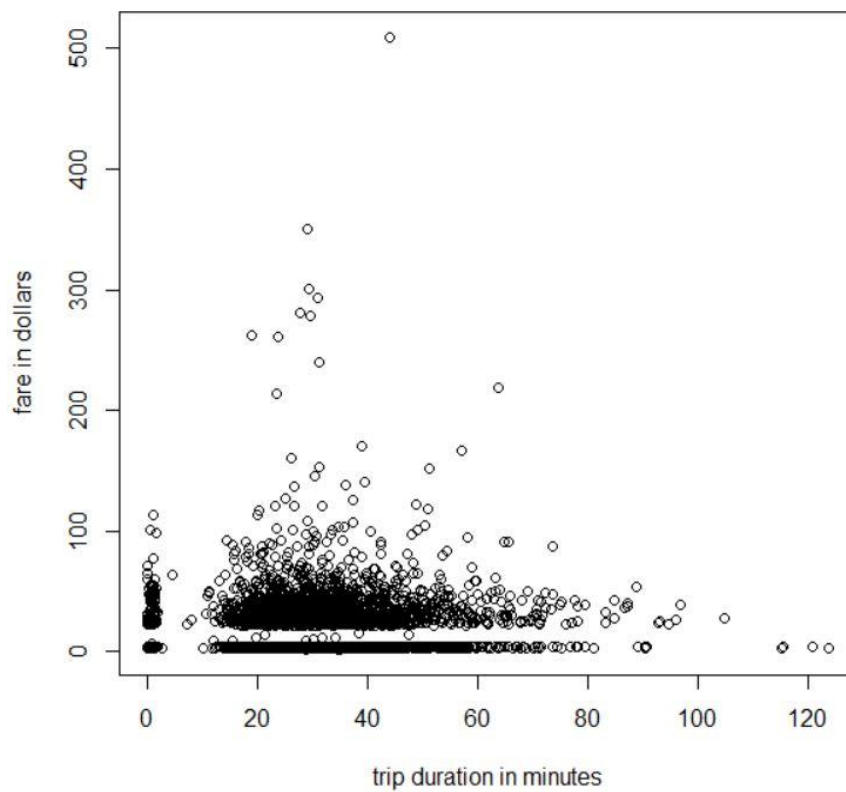


From the plot, it seems that, as the trip distance increases, the cab fare too increases on an average. We look at a boxplot between the fare and few groups (class intervals) of the distance.

Also, presence of multiple potential outliers are observed.



From the boxplot above, it seems like heteroscedasticity is present.



The above plot is a scatterplot between cab fare and trip duration.

Thus, broadly speaking, across all plots, we do happen to find numerous potential outliers and potential leverage points, along with a possible presence of heteroscedasticity. What is more unusual is the relations between the variables. For example, in the first plot, we see a lower fare for the smaller trip distances, but looking at the plot of fare against duration, we observe the fare to be more or less the same for a wide range of trip duration values.

We regress the cab fare on the other predictors by means of a classical multiple linear regression model, given by:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + e_i, \quad i=1(1)n$$

Where y_i is the response corresponding to the i th observation and n is the total number of observations.

$\beta_0, \beta_1, \dots, \beta_6$ are model parameters, e_i is the model error. x_{ki} 's are the values of the k th predictor corresponding to the i th observation. $k=1(1)6$.

By ordinary least squares regression,

We minimize $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \sum_{k=1}^6 \beta_k x_{ki})^2$ with respect to $\beta_0, \beta_1, \dots, \beta_6$ and obtain their estimates $\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_6$ respectively.

Thus we would obtain a regression line:

$$\widehat{y}_i = \widehat{\beta}_0 + \sum_{k=1}^6 \widehat{\beta}_k x_{ki}$$

This is nothing but an estimate of the conditional expectation of y given x .

$$\text{or, } \widehat{E}(Y | X_1 = x_1, \dots, X_6 = x_6) = \widehat{\beta}_0 + \sum_{k=1}^6 \widehat{\beta}_k x_{ki}$$

We then execute these calculations with the help of R to obtain the parameter estimates.

```
Coefficients:
              Estimate
(Intercept)    6.07182
distance        2.55286
duration        0.02242
extra          -2.00454
tolltax         1.22896
as.factor(day)1 -1.01732
as.factor(time)1 -0.26122
```

The coefficient estimates are to be interpreted as the average change in the value of the response, when the corresponding predictor increases by 1 unit, keeping all other covariates fixed.

However, we must look at the goodness of fit for this chosen model, for which we would refer the multiple R square measure, which is the proportion of the variance in the response, explained by the chosen model.

Residual standard error: 18.93 on 3053 degrees of freedom
Multiple R-squared: 0.5205, Adjusted R-squared: 0.5195

From R, Multiple R-squared=0.5205. Thus, approximately 52.05% of the total variation of the response can be explained by the above mentioned multiple linear regression model, which may not be satisfactory enough.

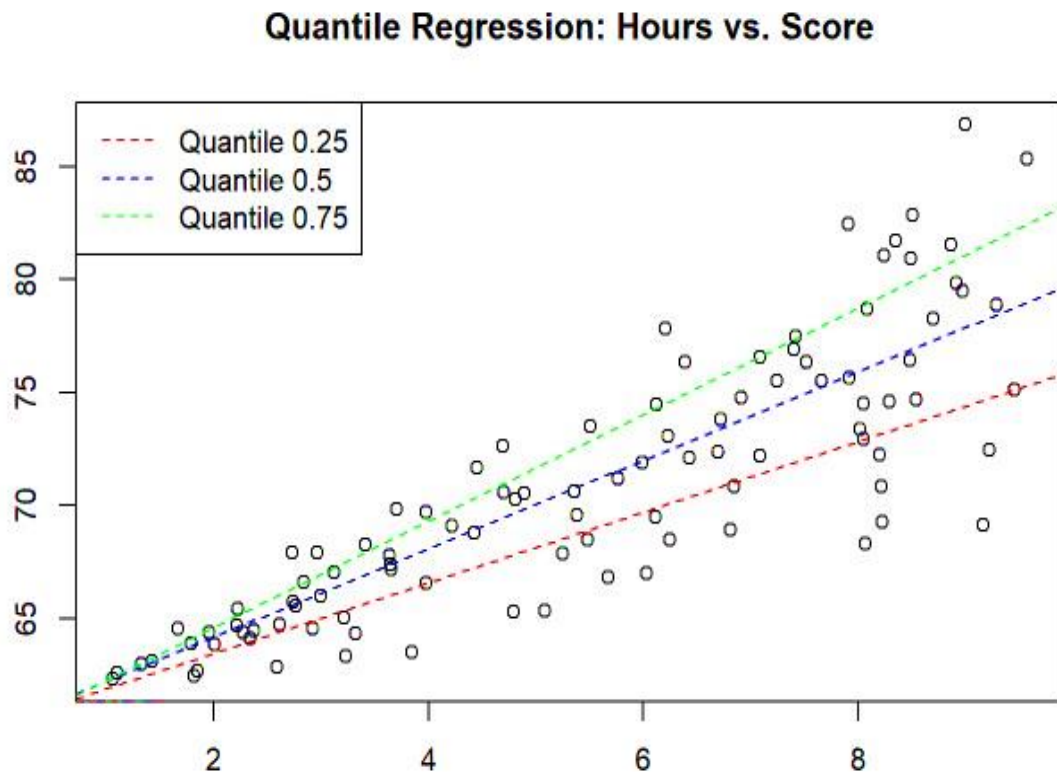
Our motive, however is to find a way to tackle the reasons which affect the classical model's efficiency. For this, we need to understand the possible reasons for it. As mentioned earlier, ordinary least squares regression helps us obtain an estimate of the conditional mean of the response variable given the predictors. Mean being a measure of central tendency that is more affected by outliers, results in this model being more affected by outliers. Also, the classical linear regression model is based on a lot of assumptions, which are important for the model to be efficient. Some of these being homoscedasticity of the error term, the error term being normally distributed etc. Thus, it is quite fair to say that the presence of many potential outliers and changing conditional variance of the response given predictors may have been the leading cause for the poor fit. Unlike this model which only gives the mean effect, we might want to get an idea of the corresponding conditional quantiles, thereby helping us to understand the effect of the different covariates on varying quantiles of the cab fare (response). For example, the coefficient for duration has been estimated to be 0.02242, or, an increase in the trip duration by 1 minute would cause the fare to go up by 0.02242 dollars. But this again, is an average estimate. The effect of trip duration of the response may be higher for certain values of the fare and lower for some other range of values.

METHODOLOGY AND THEORITICAL BACKGROUND

Quantile regression, unlike ordinary least squares regression, is a kind of regression, which fits a line for each conditional quantile, by minimizing a weighted sum of absolute residuals, also known as the quantile loss function. However, here there is a separate loss function for each conditional quantile which we want to fit. That in itself, is a kind of backtracking procedure as the loss function we are going to minimize is devised on the basis of the quantile that we would get, or “want to get” as the answer after minimizing. Some important advantages of quantile regression would be:

1. It fits conditional quantile lines of the response variable given the predictor values, thus offering a better coverage and a clearer picture of the varying distributions as we go along the values of the response, instead of just giving a picture of the mean line.
2. It is more robust to the presence of multiple outliers as compared to classical OLS regression, owing to the fact that quantile measures are less affected by outliers than the mean.
3. It is suitable for data exhibiting heteroscedasticity, as the varying levels of dispersion can be possibly tackled.

Consider, for example, (outside the context of our dataset) a plot of 2 variables: exam score and hours of study.



The above is a plot of heteroscedastic data, i.e. varying levels of dispersion. The lines help us to sketch out an idea of the conditional quantiles of the response given predictor.

Quantile Loss Function:

Let $q \in [0,1]$; q th quantile of y is that value of y below which q proportion of the observations lie, and above which $(1-q)$ proportion of the observations lie.

We know that, median is the value that minimizes the sum of absolute residuals,

i.e. minimizes $L = \sum_{i=1}^n |e_i|$,

Now L can be rewritten as $L = \sum_{i:e_i < 0} (-e_i) + \sum_{i:e_i > 0} (e_i)$

Our objective is to devise a function such that the positive residuals and negative residuals have an equal weightage. This constraint would help us to decide the factors/weights needed to be put for the positive and negative residuals. Now, proportion of positive and negative residuals isn't equal for the q th quantile (except for $q=0.5$, as the weightages are already equal in this case). We multiply the negative residuals by the factor $k*(1-q)$ and the positive residuals by the factor $k*q$, where k is any positive real number. This equalizes the weightage for both.

Thus, the loss function which is minimized by the q th quantile can be written as:

$$L_q = \sum_{i:e_i < 0} [k * (1 - q) * |e_i|] + \sum_{i:e_i > 0} [k * q * |e_i|]$$

Note, here $|e_i| = |y_i - \beta_0 - \sum_{k=1}^6 \beta_k x_{ki}|$

Proof:

Let $\beta_0 + \sum_{k=1}^6 \beta_k x_{ki} = F$ (say)

Let, out of n observations, f be $< F$, thus $(n-f) > F$

We write the loss function as:

$$L_q = k*n*[\sum_{i:e_i < 0} [(1 - q) * |y_i - F|/n] + \sum_{i:e_i > 0} [q * |y_i - F|/n]]$$

Differentiating this w.r.t F and equating to zero, we have

$$k*n*[\sum_{i:e_i < 0} [(1 - q) * -(y_i - F)/|y_i - F|] + \sum_{i:e_i > 0} [-q * (y_i - F)/|y_i - F|]] / n = 0$$

$$\Rightarrow [\sum_{i:e_i < 0} [(1 - q) * -(y_i - F)/|y_i - F|] + \sum_{i:e_i > 0} [q * -(y_i - F)/|y_i - F|]] / n = 0$$

$$\text{Now, } (y_i - F)/|y_i - F| = \begin{cases} 1, & \text{if } y_i > F \\ -1, & \text{if } y_i < F \end{cases}$$

$$\Rightarrow (1-q)*f/n + q*-(n-f)/n = 0$$

$$\Rightarrow (1-q)*f/n = q*(n-f)/n$$

$$\Rightarrow f*(1-q) = (n-f)*q$$

$$\Rightarrow f = nq$$

$$\Rightarrow f/n = q$$

Thus an estimate of the proportion of observations less than $F = f/n = q$

Therefore, F is such a value below which q th proportion of the observations lie, which we're using to estimate the q th conditional quantile of the response values, given the predictor values.

Hence proved, F minimizes L_q .

The regression model, would remain the same though.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + e_i \quad , i=1(1)n$$

The only thing that changes is the way we are obtaining our model parameter estimates.

We minimize the skewed loss function L_q w.r.t the parameters, and accordingly obtain an estimate for the q th conditional decile.

We would obtain a set of $(\widehat{\beta}_{0q}, \dots, \widehat{\beta}_{6q})$ for each decile. $q = 0.1, 0.2 \dots 0.9$.

And the q th conditional quantile is to be fitted by the line:

$$\widehat{y}_{iq} = \widehat{\beta}_{0q} + \sum_{k=1}^6 \widehat{\beta}_{kq} x_{ki}$$

RESULTS AND DISCUSSION

We fit the quantile regression model on our data using the `rq()` function of the `quantreg` library in R and obtain 9(corresponding to each decile) fitted lines.

Table of Parameter Estimates:

	<i>Intercept</i>	<i>Distance</i>	<i>Duration</i>	<i>Extras</i>	<i>Toll Tax</i>	<i>Day of Week</i>	<i>Time of Day</i>
10%	2.8000000	0.0000000	0.0000000	1.0000000	1.363573	0.000000	0.0000000
20%	-0.2459285	2.747055	0.020423	0.7286983	1.071564	-0.731221	0.1911833
30%	1.9159820	2.713966	0.010324	0.4314099	1.067516	-0.462311	0.08633286
40%	2.7480519	2.759740	0.000000	1.0000000	1.067077	-0.224026	-0.2240260
50% <i>(median)</i>	2.8951580	2.856965	0.002447	1.0407089	1.075739	-0.287122	-0.4050839
60%	3.0023810	2.976190	0.0000000	1.0952381	1.061326	-0.250000	-0.3452381
70%	3.8684932	3.082192	0.0000000	1.3835616	1.168855	-0.500000	-1.068493
80%	6.5082996	3.162279	0.01075	1.0720785	1.287787	-0.897270	-3.270887
90%	13.2649731	2.919055	-0.001810	0.5922786	1.657640	-1.4879	-4.812808
OLS estimate	6.0718202	2.552856	0.022421	-2.004540	1.228963	-1.0173	-0.2612177

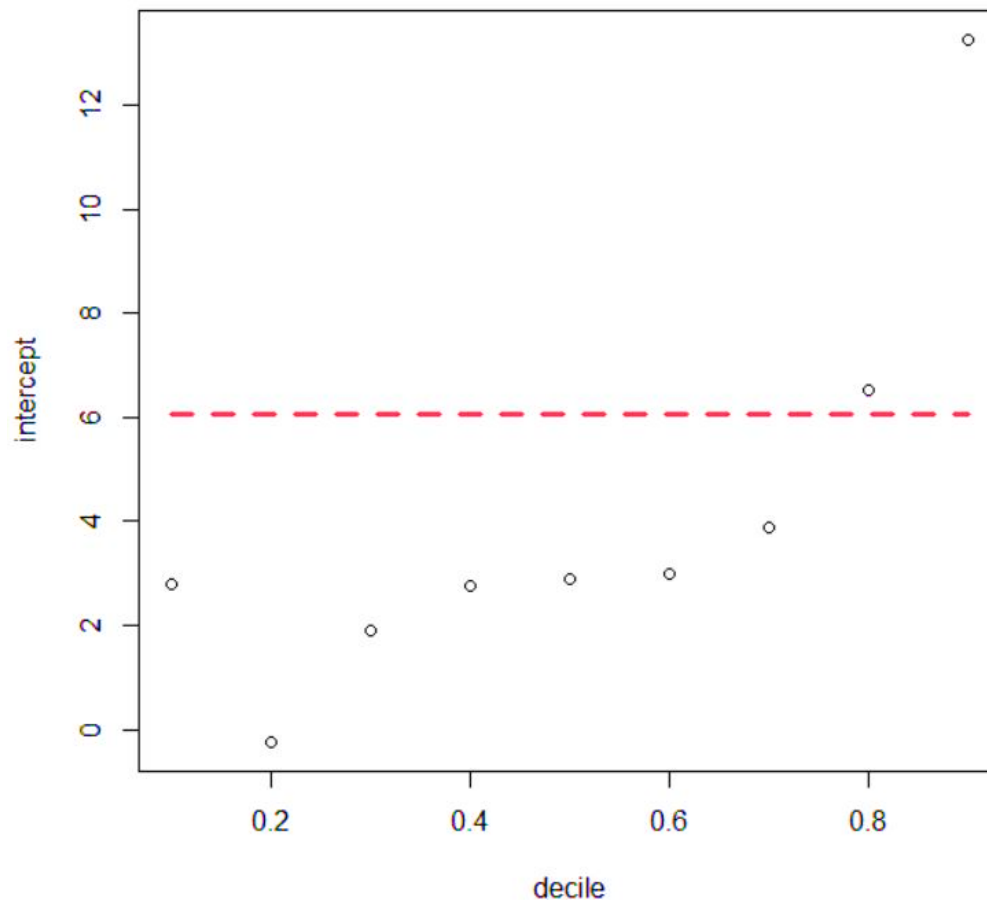
Table 1

The columns in Table 1 denote the coefficients of the respective variables and the rows denote the decile fitted. The last row denotes estimates of least squares regression. Note that, some of the estimates which are magnitudinally negligible such as 1.69×10^{-16} have been rounded off to zero.

Diagrammatic understanding of some of the results:

The plots below show how the parameter estimates of a particular variable vary over the lines fitted for each decile. The red dotted line denotes the least squares estimate.

The estimates are of intercept, trip distance, day (of week) and hour (of day) respectively.

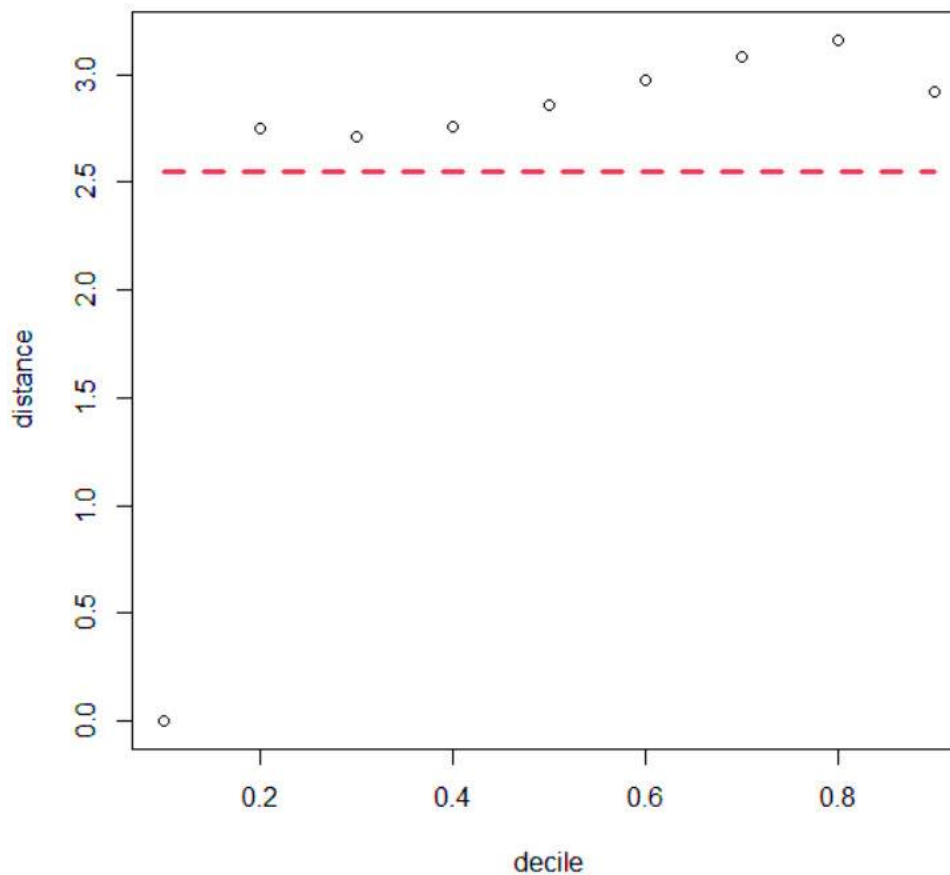


(Fig 1)

Interpretations from the graphs:

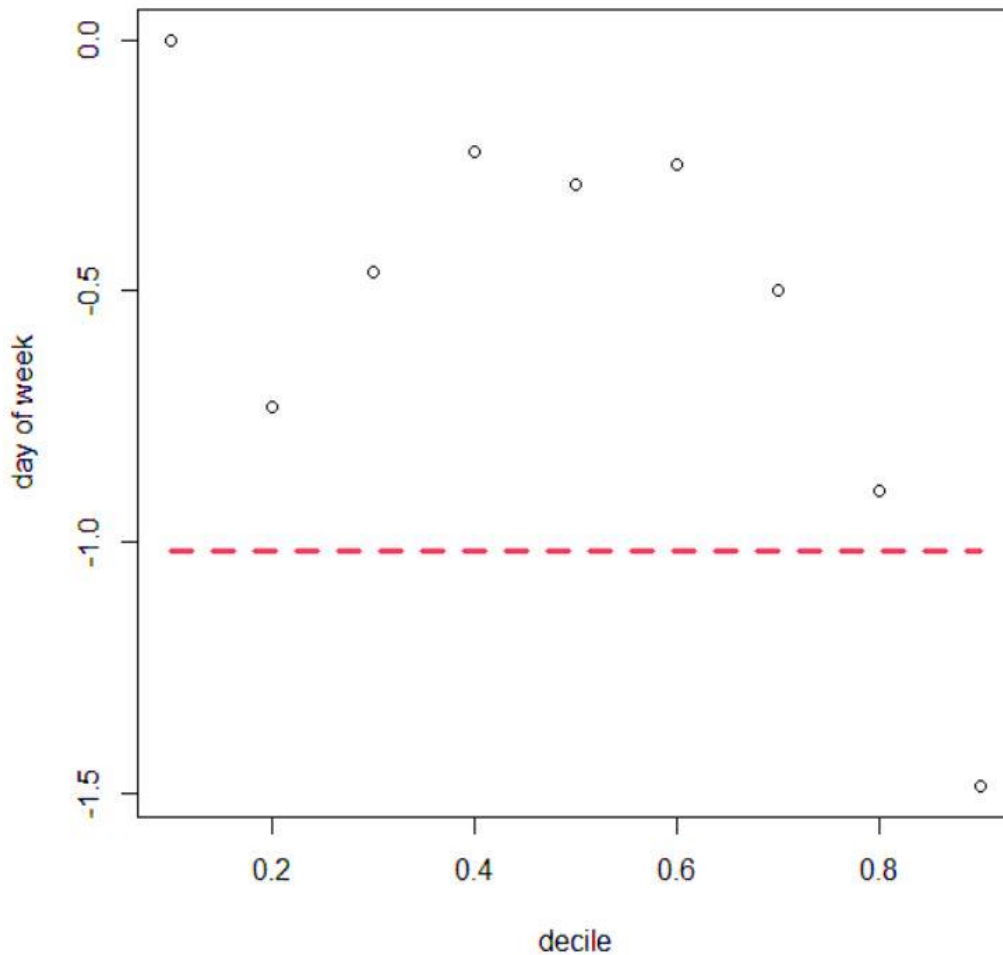
1. Intercept estimate: (Refer Fig 1) The intercept estimate is relatively low for lower deciles of the fare. The estimates tend to increase as we look into higher deciles. This is expected, as this implies that when all covariates are zero, the lower conditional decile of the cab fare are lower

than the higher conditional decile. The OLS line overestimates the intercept for the majority of the fare.



(Fig 2)

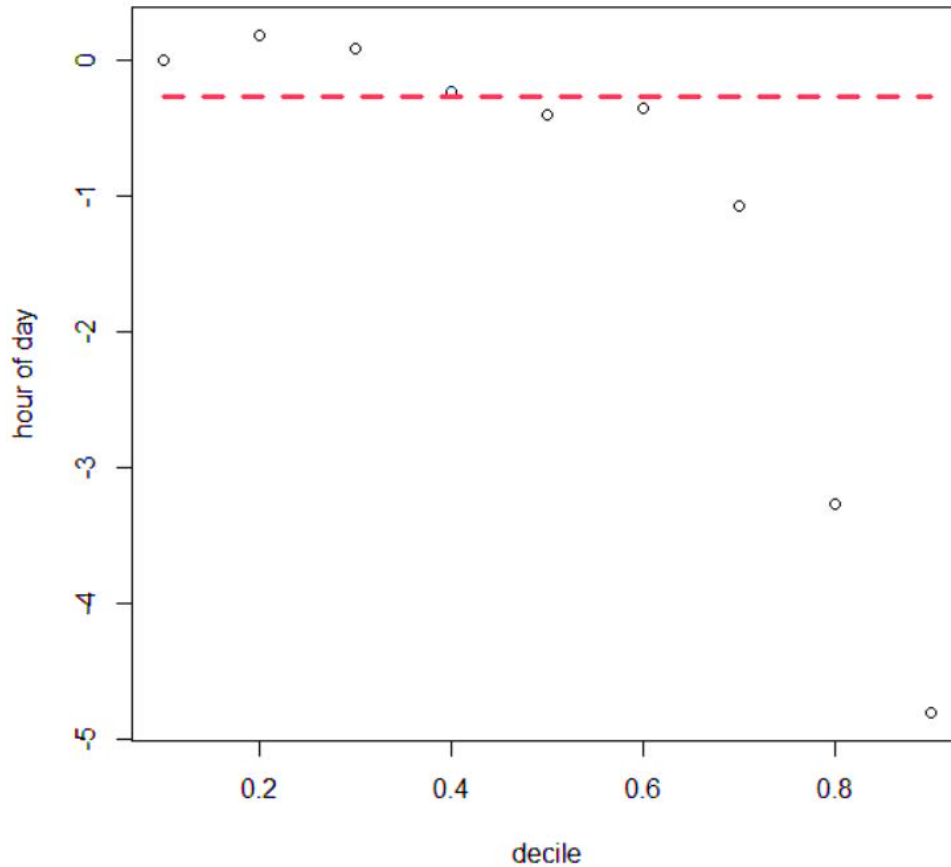
2. Distance: (Refer Fig 2) For most part of the data, the estimate remains more or less the same, i.e. in the range of 2.7 to 3. This value could be interpreted as the change in the respective conditional decile of the cab fare (measured in US dollars), when the trip distance increases by 1 mile, keeping all other covariates fixed. However, the estimate for the 1st decile is unusually low, close to zero. This could be a specific extreme case in our data. This part of the data seems to have slightly affected the OLS estimate, which is slightly lower than all the other estimates of the conditional decile lines. This could be owing to the fact that quantile measures are more robust to such extreme data values.



(Fig 3)

3. Day of week: (Refer Fig 3) The OLS estimate has come out to be around -1.0173. This means that as we go from a weekend to a weekday (keeping other factors fixed) the cab fare decreases on an average by 1.0173 dollars. This effect, however is a bit different for different conditional deciles of the fare, according to the quantile regression results. We see that, for the lower deciles, this estimate increases, i.e. the decrease in the lower conditional deciles of the cab fare as we go from a weekend to a weekday, is slightly lesser, and lessens more in the roughly the first half of the conditional deciles. After this, the estimate decreases, as we go to higher conditional deciles.

The comparatively higher fare for weekends can be attributed to higher demand of cabs and lesser availability of drivers, causing a surge in price.



(Fig 4)

4. Hour of day: (Refer Fig 4) The estimate for this coefficient decreases as we go to higher conditional deciles. Thus, for higher values of the cab fare, as we go from night to day (keeping other factors fixed), the decrease in the cab fare is more, than as compared to the lower values of the fare. For example, the estimate for this parameter for the 9th decile, or (0.9)th quantile line has come out to be -4.812. This implies that going from night to day decreases the 9th conditional decile of the cab fare by 4.812 dollars, keeping other covariates fixed. The OLS estimate tends to underestimate this decrease in cab fare for the latter conditional deciles.

Here too, the comparatively higher fare during the night is due to lesser availability of drivers, thus causing an increase in the demand for them. The companies tend to surge up the fare, so that more drivers can be encouraged to accept the rides.

GOODNESS OF FIT

In case of quantile regression, we do not have a multiple R square, as for OLS regression. However, we have an analogy for it. The basis of this is to measure the proportion of the variation in the response that can be explained by the model that is being fitted.

For quantile regression, we do have multiple lines that we are fitting. However, we could look at the median line, i.e. the estimate of the conditional median of the response given the predictors, as the most suitably comparable option with the mean fit (OLS estimate).

The measure we use here is called Koenker's R squared or pseudo r squared denoted by R^1 .

$$R^1 = 1 - \frac{\sum \rho_{\tau}(y_i - \hat{y}_i)}{\sum \rho_{\tau}(y_i - \tilde{y})}$$

- $\rho_{\tau}(u) = u(\tau - I(u < 0))$ is the **check function** for quantile regression.
- y_i = actual response value.
- \hat{y}_i = predicted value from the quantile regression.
- \tilde{y} = fitted value from the **intercept-only** model (i.e., the unconditional quantile estimate).
- $\sum \rho_{\tau}(y_i - \hat{y}_i)$ is the sum of **quantile regression residuals** (quantile loss).
- $\sum \rho_{\tau}(y_i - \tilde{y})$ is the **total variation** in the response variable.

From calculations using R, we have conditional median fitted as:

$$= 2.8951 + 2.8569x_{1i} + 0.00245x_{2i} + 1.0407x_{3i} + 1.0757x_{4i} - 0.2871x_{5i} - 0.40508x_{6i}$$

$$\sum \rho_{\tau}(y_i - \hat{y}_i) = 10345.21$$

$$\sum \rho_{\tau}(y_i - \tilde{y}) = 26599.77$$

Thus, $R^1 = 0.611$

This can be interpreted as 61.1% of the variation in y can be explained by the fitted median regression of the response on predictors, which by magnitude, is an improvement over the mean fit.

LIMITATIONS OF QUANTILE REGRESSION

1. Computational Difficulty:

Quantile Regression, unlike least squares regression, has a much more rigorous procedure for estimation of parameters and calculation processes. For large scale statistical analysis, this poses to be a hindrance in being an efficient regression process.

2. Interpretability of Results:

For least squares regression, we can interpret the parameter estimates as the average effect due to the predictors, or the rate of change of the response with respect to a predictor. But the concept of conditional quantiles in quantile regression makes it more cumbersome as we may have different interpretations for different quantiles.

3. Sensitive to Outliers at Specific Quantiles:

Although quantile regression is robust to outliers when modeling central quantiles (e.g., median), it can be sensitive to outliers at the extremes (e.g., 0.05 or 0.95 quantiles). These values can have a disproportionate influence on the estimated coefficients at the boundaries of the distribution, leading to difficulty in reasoning and interpretability of results.

4. Effect due to Multicollinearity:

Much like the estimates for ordinary least squares, the estimates here too are prone to be affected by dependencies among predictor variables.

5. Quantile Crossing:

It is expected that a higher quantile would have a higher fitted value than a lower quantile. However, this might not always be reflected by the estimated lines. This phenomenon is termed as quantile crossing. This might be due to improper model specification.

6. Not a Causal Inference Method:

Quantile regression estimates at different points how the variables are associated, but it is not conclusive enough to infer about a certain variable being a cause of the change in response. For example, if marks scored by students are regressed on the number of tuition classes in a week, and let's say we get a positive coefficient for the fitted 75th quantile, it might indicate that people who attend more classes, get more marks. But this doesn't mean that attending more classes is the cause of higher marks.

CONCLUSION

From the analysis on our data, we actually get to see a more in-depth picture of the varying importance of predictors in affecting the response. For example, the time of the day at which the cab ride is being taken is a more effective predictor for higher values of the fare, than as compared to the lower values. Similarly, predictors such as extras or toll tax seem to be more or less equally effective across all values of the cab fare.

Our analysis also highlights those areas where ordinary least squares regression struggles to give a good fit. This is usually pertaining to underestimation or overestimation of certain predictor coefficients by the least squares method.

Surprisingly, the duration of the trip, seems to be a comparatively less important predictor for the cab fare. This could be possibly due to its dependence with the trip distance (i.e. a longer route or distance would take a longer time to travel than a shorter route). As mentioned in the limitations, multicollinearity among the predictors poses to be a possible reason for unusual parameter estimation.

What is important to note here is, that this is a dataset of New York City, so the pricing systems and fares in general, need not be the same as that for other cities. Sometimes, even the cab company could prove to be a big deciding factor in cab fares.

REFERENCES

- i. [https://search.r-project.org/CRAN/refmans/WRTDStidal/html/goodfit.html#:~:text=The%20goodness%20of%20fit%20measure,non%2Dconditional\)%20quantile%20model.](https://search.r-project.org/CRAN/refmans/WRTDStidal/html/goodfit.html#:~:text=The%20goodness%20of%20fit%20measure,non%2Dconditional)%20quantile%20model.)
- ii. <https://www.spsanderson.com/steveondata/posts/2023-11-29/index.html>