An Empirical Analysis of Taxi Pricing Schemes, Fraudulent Detours, and Passenger Volition

111266012 莫喬丹 Jordan Murillo | 111266004 白季堯 Josef Blazek | 111266007 張靜宜

International Master's Program of Applied Economics and Social Development

National Chengchi University

266850001: Applied Microeconometrics with R

Prof. 廖仁哲 Jen-Che Liao

12 January, 2024

Replication of: *Liu, Ting and Vergara-Cobos, Estefania and Zhou, Yiyi, Pricing Schemes and Seller Fraud: Evidence from New York City Taxi Rides (March 16, 2017)*

Introduction

The paper by Liu, Vergara-Cobos, and Zhou examines an important issue in markets with information asymmetries – the problem of seller fraud and how pricing schemes affect the incentives for sellers to defraud consumers. As the authors discuss, seller fraud is prevalent across many expert service markets, including healthcare, auto repairs, and taxi rides (Wolinsky 1993; Dulleck & Kerschbamer 2006). However, seller fra[1]ud is difficult to study empirically due to challenges in identifying fraudulent behavior and quantifying its costs (Levitt & Syverson 2008; Grytten et al. 2011).

The taxi ride market provides an ideal setting to analyze seller fraud given the clear delineation between overtreatment (unnecessary detours), undertreatment (failing to reach destination), and overcharging (exaggerating distance/time). The paper makes a novel contribution by constructing "detour indexes" to measure overtreatment and examining how these indexes correlate with different pricing schemes and occupancies in New York City. Their difference-in-differences methodology accounts for potential confounds by comparing trips priced under a two-part tariff versus a flat fare before and after a rate change.

The paper provides compelling evidence that drivers do respond to financial incentives by taking more detours when subject to a two-part tariff, especially when occupancy rates are lower and variable rates higher. The authors estimate such detours impose annual costs of $0.5-0.8 million on passengers. These findings have important regulatory implications regarding taxi pricing policies and reducing information asymmetries. The paper makes key strides in identifying and quantifying seller fraud in a credence-good market. However, the analysis also has some limitations that provide opportunities for extension. First, the detour indexes are inferred rather than directly observed, as the authors note (Liu et al. 2017). Future research could aim to more directly measure overtreatment through audits or tracking. Second, the study is restricted to two airports in New York City, limiting generalizability. Expanding the analysis to other cities would provide valuable robustness checks and additional policy insights tailored to different contexts.

[1] The replication code for this project can be accessed from https://github.com/jormur/taxi-fare-fraud

Additionally, while the paper focuses specifically on overtreatment in the form of detours, analyzing undertreatment and overcharging could also merit attention. For instance, DiNardo and Lee (2004) and Davis (2008) both exploit policy changes to examine impacts on labor and environmental outcomes. A similar approach could help identify effects on underprovision of services or overcharging if regulations shifted. Finally, the welfare impact on consumers remains unquantified - do they detect and avoid detours or fraudulent taxis? Incorporating consumer responses could better elucidate the full welfare effects.

Nonetheless, within the scope of analyzing overtreatment and pricing incentives, the paper makes an important empirical contribution. Seller fraud is notoriously difficult to study due to limitations on information. The paper's innovative methods and credible identification strategies significantly advance understanding of fraud incentives and costs in the taxi market. The results offer valuable insights for regulation in taxi and related expert service markets. The analysis helps address gaps in quantifying and deterring fraudulent behavior when information asymmetries exist between sellers and consumers. While opportunities remain for extensions, the paper marks an important step toward empirically identifying and reducing seller fraud across credence good markets.

Other than the limitations mentioned above, the inability to include tips as a control variable in the regression model was also mentioned as an important factor. Not controlling for the amount of tip is problematic, mainly because the correlation with factors that influence the driver's route choice might be disregarded. The exclusion was therefore seen as a new area for further analysis. Realizing this and the importance of tipping in the US service industry with its significant influence on (especially low-paid) workers' salaries, we decided to place it at the center of our contribution and observe the effects selected variables have on tipping (Azar, 2011).

Based on the assumptions that are introduced in the extension part of this paper, we hypothesize that local residents tend to tip taxi drivers more. Previous research has already identified motivations that stimulate customers' desire to provide taxi drivers with tips. Lynn (2015) illustrates the connection between reward and altruistic motives with increased amounts of tips. Azar emphasizes the altruistic motives, stating that tipping is more prevalent in

lower-wage occupations when a close relationship is established between the customer and the worker. Furthermore, Azar (2011) concludes that the level of psychological utility for consumers is a crucial factor in determining the decision to provide a tip. In contrast, Chandar et al. (2019) found that both the likelihood of tipping and the size of non-zero tips decrease when drivers engage in behaviors such as speeding, hard braking and sudden acceleration.

As the original paper placed importance on seeing the effect of price increase on the drivers' behavior through price schemes they use, we decided to also test whether it is true that residents tip drivers more as they believe they will not be overcharged. Residents are familiar with services offered in the city as well as what kind of price scheme is established in individual airports. Therefore, they tip more if they take a taxi charging flat rates as they believe those will not perform any detour. To this end, we decided to employ a similar approach methodology-wise in the use of a two-way fixed-effects estimation so as to control for trip and time variation, relying on the formulated variables constructed in the replication process.

Data

Just as in the paper, we sought out data from the Taxicab Passenger Enhancement Program (TPEP) data. The entire database of all the records in the time period that the authors used, 2010 to 2013, contained many hundreds of millions of trip records, so they therefore filtered out a sample according to a number of criteria that served the purpose of their empirical aims and were purportedly supported by their theoretical bases. The authors narrowed down the kinds of trips into those that only originated from either of two airports in NYC, LaGuardia Airport (LGA) or John F. Kennedy International Airport (JFK). This directly serves in the testing of one of the authors' claims of there being a difference in relative detour (and thus, fraud probability) between the two fare schemes: two-part tariffs and flat fare charging. According to NYC's pricing regulation, trips from LGA to Manhattan are priced using a two-part tariff that consists of a fixed fare and a variable fare that depends on trip distance and duration. In contrast, trips from JFK to Manhattan are subject to a single flat fare, regardless of trip distance or duration.

Thereafter, to the end of narrowing down the kinds of passengers to serve as the treatment and control group in their testing of their hypotheses, they only used trips that ended in

the approximate area of a select handful of popular hotels and those in a particular residential area in reasonable proximity to either airport and that would have residents of particular means who would be predisposed to taking such a taxi trip.

Taking the steps to replicate this paper and its findings, we employed what are approximately identical sampling criteria. Additionally, the sample was constructed based on a few other parameters and with a few other considerations in mind. The sample excludes all trips happening on major holidays due to the likely severe difference in traffic and trip demand over "normal" days. Like the authors, we also excluded the trips that have extreme values of distance, duration, or fare that are more than three standard deviations away from the mean. Our sample results in approximately 400,000 trip records versus their 177,981 records.

It is here that we note the first divergence from the original methodology. We have decided to draw on the years 2011 to 2014, as opposed to the original 2010 to 2013. This decision is made in consideration of the accuracy of empiricism. The NYC Taxi Association decided to rehaul their data coding format sometime during/at the end of 2010. One of the most crucial variables used in this study is that of the trip location recording and it is this very same that was changed from being recorded as exact coordinates to a more organized, yet consolidated zone ID scheme. The discrepancy in between year schemes makes the joining of their data challenging and most likely inaccurate. Therefore, we opt to simply shift the observation period a year later. This is not done without consideration, as we believe this period lends itself more appropriately to the research aim, whereby the policy intervention of interest would occur very much near the midway point in the span of time (the policy was enacted on 4 August, 2012).

We present the summary statistics of our sample as follows, for comparison with the original paper's sample. Table II coincides with Table II of the paper, comparing the driver characteristics between departure locations/airports. Table III presents the sample statistics of our constructed sample, for comparison to the original.

Table II: Driver Characteristics

Airport	num_trips	working_hours	occupied_hours	occupancy
JFK	124224	143087	94521.23	0.6605857
LGA	155516	159039	84842.34	0.5334687

Table III: Summary Statistics by Departure Location and Trip Type

Airport	trip_type	num_trips	avg_trip_distance	avg_duration	avg_fare
JFK	Hotel	133444	17.857917	47.22825 mins	49.15881
JFK	Residential	34769	17.308539	39.90870 mins	49.19070
LGA	Hotel	165064	9.051135	33.14373 mins	28.55425
LGA	Residential	49859	7.891844	24.75080 mins	24.04569

Methodology & Extension

The key methodology utilized by Liu et al. (2017) is a difference-in-differences approach leveraging a natural experiment - a change in taxi fare pricing schemes in New York City. Specifically, the authors construct "detour indexes" that measure the extra distance, time, and fare incurred by passengers traveling from airports to hotels versus those traveling to nearby residential areas. The residential trips proxy for informed local passengers, while hotel travelers are perceived as less informed about optimal routes. Differencing isolates the portion of detours likely attributable to driver fraud motives rather than other factors like traffic. Liu et al. then examine how these detour indexes change pricing shifts occur in 2012. The difference-in-differences framework compares detour indexes across pricing schemes and over time while controlling for various trip characteristics. The compiled sample data comprises of the key feature being that the fare rates changed during the sample period: both the variable rate of the two-part tariff and the rate of the flat fare increased in September, 2012. This rate change serves as an exogenous variation in drivers' financial incentives to detour and allows us to disentangle drivers' fraud motives from alternative explanations. By comparing trips before and after the fare rate change, we are able to identify how the fraudulent detour varies in the fare rate and the expected post-dropoff occupancy. This methodology allows the authors to empirically identify the causal effect of changing pricing incentives on the extent of seller fraud in the taxi

market. Replicating this approach is crucial for extending the analysis to quantify unexplored outcomes like impacts on passenger avoidance behavior. The difference-in-differences methodology leveraging the 2012 NYC pricing change provides a credible strategy for causally assessing how pricing schemes influence seller fraud.

The authors identify their main empirical model specification as follows:

$$detour_{ijt}^{k} = \beta_1^k occ_i + \beta_2^k g_t + \beta_3^k g_t * occ_i + \lambda_j^k + f^k(date_t) + \epsilon_{ijt}^k, \quad (1)$$

where $detour_{ijt}^{k}$ is a detour index of a trip i that originates from airport k and ends in hotel j on date t, occ_i is the expected occupancy at trip i's dropoff hour, g_t is a dummy variable that equals one if the date t is after the rate increase and zero otherwise, $g_t \times occ_i$ is the interaction of g_t and occ_i, λ_j^k are airport-hotel-specific factors affecting the detour index, $f^k(date_t)$ is a flexible polynomial in the trip's date that controls for omitted time-varying factors, and ϵ_{ijt}^k is an error term.

It is here that we make note and detail the empirical model employed. The authors refer to their evaluation of equation (1) as being *"akin to"* a difference-in-difference estimation. Through rigorous review and implementation testing, we distill from their descriptions that their methodology is not a classical difference-in-differences design despite their wording. By constructing the detour index as the difference between hotel and residential trips, they have essentially created a treated versus control comparison built into their outcome variable.The authors have already constructed a pseudo-counterfactual for each hotel trip using the residential trips. To that end, we have sought to employ a two-way fixed effects estimation. The detour index contrasts each hotel trip to "informed local" residential trips departing from the same airport at nearly the same time. This effectively nets out common factors like traffic or road closures that would affect trips to both destinations equally. What remains in the index isolates the portion of extra distance/time/fare attributable to driver fraud rather than other reasons. In a sense, the authors have implicitly built a differences-in-differences style counterfactual into their detour index outcome measure. Therefore, the subsequent FE regression we use then controls for any remaining confounds not already differenced out by the index construction.

The authors sought to investigate a number of claims/hypotheses that, to the end of replicating their findings, we see to corroborating to the ends of what can be distilled from their main set of results (known as Table V in the original paper):

Claim 1: The driver will not take the passenger on unnecessary detours if the trip is subject to a flat fare.

Claim 2: The detour rate increases with the variable rate. A higher variable rate increases the driver's marginal benefit from detouring but does not affect the marginal cost of doing so. As a result, the driver will take the passenger for a longer detour when the variable rate is higher.

The details of our replication of these variables are detailed in the following.

Expected Occupancy

The authors describe how, based upon their theoretical analysis, drivers have stronger incentives to take fraudulent detours if they expect a lower occupancy after dropping off the current passenger. Therefore, it is necessary to measure the occupancies that are expected by drivers. The expected occupancy variable aims to capture the demand conditions facing drivers when deciding whether to take fraudulent detours. To measure actual occupancy, the authors first calculate each driver's hourly occupancy rate based on the proportion of time their taxi is occupied in a given hour. They then take the average across all drivers active during that hour to get the citywide hourly occupancy rate.

To estimate the occupancy rate drivers would expect in a given hour, the authors regress the actual occupancy on control variables including year indicators (3), hour of day by day of week (167), and week of year (51). The predicted values from this regression represent the anticipated occupancy component, while the residuals reflect unanticipated deviations from expectations.

The authors thus use the predicted values from the regression analysis to measure the hourly occupancy rates taxi drivers would expect based on each trip's dropoff time. This expected occupancy variable is then used as a proxy for demand conditions relevant to drivers' detour incentives and decisions. Constructing the variable in this way allows the analysis to focus on the component of occupancy that is anticipated by drivers when making routing

calculations. We go about creating our reconstruction using the same variables/indicators to the best of our interpretation and utilizing R's built-in regression functions (via *lm* and *predict* to derive residuals).

Detour Index

The detour index aims to isolate the portion of extra distance, time, and fare incurred on a trip that is likely attributable to the taxi driver taking an unnecessarily long route due to fraud motives. To construct this index, the authors focus on trips from airports to hotels versus nearby residential areas.

Hotel passengers are assumed to be less familiar with optimal routes than local passengers traveling to residential addresses. For each airport-hotel trip, the authors identify a set of comparison trips from the same airport ending in the nearby residential area within a short time window (e.g. 15 minutes). They then calculate the detour index as the difference between the hotel trip's actual distance, duration, and fare and the average values for the residential comparison trips. This differences out any common factors like traffic or road closures that would affect trips to both destinations similarly. The resulting detour indexes provide a metric capturing the excess distance, time, and fare incurred specifically on the hotel trips relative to informed local passengers. Since residential and hotel passengers hailing taxis at the airport face similar initial conditions, this approach isolates the portion of the longer hotel trips plausibly explained by driver fraud. By constructing these indexes, the authors develop a novel variable measuring overtreatment attributable to driver fraud. Examining how the indexes correlate with pricing schemes and occupancies provides insights into determinants of fraudulent behavior.

In line with our research goals, we identify a non-local passenger as the one traveling from either of two hotels in our dataset to specified hotel locations. This is an important assumption because we establish passenger's residency as a treatment variable. The trip duration, distance and fare are outcome variables. To see the effect of the treatment variable on the outcomes, we construct a detour index which reflects extra duration, distance and fare relative to being a local. It thus serves as a proxy of average treatment effect on the treated.

In order to establish apple-to-apple comparison and eliminate bias, our identification strategy relies on the fact that passengers are randomly assigned to taxi drivers as they are

prohibited by law to reject a customer based on race, disability, or destination within NYC once they enter the taxis. Conditioning on that, for each trip realized by non-local (observed outcome), we calculate average outcomes of trips that originated from the same airport within a 15-minute time period but ended up in a specified residential area (proxy for unobserved outcome). We then subtract those two values and consider the results as the ATE.

Table IV presents the mean detour per respective airport for reference and comparison.

Table IV: Mean Detours by Airport

Airport	avg_detour_distance	avg_detour_duration	avg_detour_fare
JFK	4.868641	8.111029	11.274777
LGA	-3.877009	-5.885130	-8.931812

Extension

The dataset used for the extension estimation covers the same time period and variables. With regards to our hypothesis, it is desirable to make a short comment on cash and credit card transactions. Tips information was only collected for credit card payments, resulting in a significant number of missing values as the majority of trips were paid by cash. This unfortunately leaves us with a smaller sample size than the one employed in the previous estimation. Another limitation worth mentioning is the fact that the dataset does not provide passenger-related characteristics. These sources of information are especially important as they appear to play a significant role in tip prediction. They would therefore represent key ingredients in our estimation. For this reason, we expect to have unobserved variable bias.

The idea behind our extension and its related assumptions are reflecting the findings in literature - psychological motivations are an important factor when deciding whether to tip a driver or not. We outline the assumptions as follows. Firstly, we state that residents regard taxi drivers who use a flat fare price scheme in their fare calculation to be less fraudulent. Secondly, we believe that local passengers are more familiar with the types of taxi services offered in the city and where they are available. Lastly, in line with the literature, we assume that locals are more likely to establish amicable relationships with drivers since they may expect to see them again and/or want to encourage good drivers to stay in the market (Neto et al. 2019).

Based on these assumptions, we hypothesize that passengers are aware JFK taxi drivers charge a flat fare rate for trips and therefore assume that those drivers will not take a detour. This leads to a higher tip amount compared to a driver who uses a two-part tariff.

Having this in mind, we construct the following model which we built upon the needs of our hypothesis and with reflection of models used in the literature, namely in Neto et al. (2019) and Conlisk (2022):

$$tip_{it} = \beta_1 local_{it} + \beta_2 duration_{it} + \beta_3 distance_{it} + \beta_4 fare_{it} + \beta_5 passnum_{it} + \gamma_t + \varepsilon_{it} \quad (2)$$

Where tip_{it} stands for a tip that originated during trip i on date t, $local_{it}$ is a dummy variable equal to one if a passenger during trip i on date t is resident in NYC and zero otherwise. $duration_{it}$ is the the time in minutes a trip i took on date t. $distance_{it}$ equals the trip i's distance in miles on date t, $fare_{it}$ is the fare amount charged by driver for trip i on date t, $passnum$ is the number of passenger taking the same taxi trip i on date t, γ_t is the fixed effect for a date t and finally ε_{it} represents the trip's i error term on date t.

We use a fixed effect model to control time. Time-related factors, such as seasonality, day-of-week effects, or time-of-day patterns, can introduce variability in the data. Controlling for time allows the model to capture and account for these patterns, providing more accurate estimates of the effects of other variables of interest. These fixed effects help control for unobserved hour-specific factors that might otherwise affect the tips but are not explicitly included in the model.

We check the robustness across model specifications with and without fixed effects utilizing OLS linear regression, a panel data fixed effect pooling model, and the original two-way fixed effects for comparison.

Results

Our replication of authors' methodology culminates in our reconstruction of Table V found in the original paper, where ours can be found below for reference and comparison.

Table V: Estimations Results of Equation (1)

Model:	LGA Distance (1)	LGA Duration (2)	LGA Fare (3)	JFK Distance (4)	JFK Duration (5)
Variables					
occ	8.016***	-84.82**	1.690	7.535***	-95.07***
	(0.7979)	(16.21)	(4.938)	(0.2359)	(7.863)
g	0.0558	32.72***	12.78***	-0.0660	7.527**
	(0.3619)	(4.199)	(2.002)	(0.4220)	(1.390)
occ × g	0.6280	-60.68***	-24.29***	0.3017	-4.928
	(1.272)	(6.795)	(3.680)	(0.7013)	(2.465)
Fixed-effects					
Trip Unit	Yes	Yes	Yes	Yes	Yes
Date	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	155,516	155,516	155,516	124,224	124,224
R^2	0.14639	0.11780	0.19665	0.36909	0.07603
Within R^2	0.00145	0.00700	0.00067	0.00590	0.00529

Clustered (Trip Unit) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

In our replication, as seen in Table V, our estimates are wildly different from the original ones. In particular, the coefficients on the occupancy variable and the interaction term are much larger in magnitude and opposite in sign for most models. For example, for LGA distance, we find that occupancy increases the distance by 8.016 km, while the original authors find that it decreases the distance by 0.515 km. Similarly, for LGA duration, we find that occupancy decreases the duration by 84.82 minutes, while they find that it increases the duration by 3.185 minutes.

Additionally, Liu et al. found no significant effects for any determinants of JFK detour indexes, whereas two of our JFK coefficients are significant. This suggests our models may not fully capture latent relationships or clean heterogeneity as effectively. Small sample issues, functional form assumptions, error correlations, and other modeling choices may also contribute to quantitative differences.

There are several possible explanations for this discrepancy. Firstly, even with a meticulous approach to sample creation and variable reconstruction, subtle differences in data

handling and processing can lead to significant variations in outcomes. The statistical models employed are sensitive to these nuances. For instance, the choice of unit of analysis (trip or driver-day), the treatment of outliers and missing values, the definition of trip origin and destination, and the calculation of distance and duration can all affect the results. Secondly, the original authors might have used specific exclusion or inclusion criteria or data cleaning processes that are not entirely clear or detailed in their publication. This lack of transparency can often lead to replication issues. For example, Liu et al. state that they exclude trips with negative or zero fare, but do not specify how they handle trips with negative or zero distance or duration.

Results of Extension

The model employed for the estimation of our hypothesis yielded results as follows.

Table VI: Estimations Results of Equation (2) with Robustiness Checks

	LGA OLS	JKF OLS	LGA PLM	JFK PLM	LGA FE	JFK FE
Local dummy	−0.201***	0.774***	−0.248***	0.773***	−0.248***	0.773***
	(0.019)	(0.031)	(0.019)	(0.031)	(0.017)	(0.031)
Fare Amount	0.134***	0.159***	0.084***	0.036***	0.084***	0.036**
	(0.002)	(0.003)	(0.003)	(0.008)	(0.003)	(0.012)
Trip Distance	−0.138***	−0.226***	−0.001	−0.196***	−0.001	−0.196***
	(0.006)	(0.009)	(0.007)	(0.009)	(0.009)	(0.010)
Trip Duration	0.000	−0.013***	0.000	−0.015***	0.000	−0.015***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Passenger Count	−0.100***	−0.155***	−0.083***	−0.135***	−0.083***	−0.135***
	(0.006)	(0.009)	(0.006)	(0.009)	(0.006)	(0.009)
Hour of Week			−0.002***	−0.002***	−0.002***	−0.002***
			(0.000)	(0.000)	(0.000)	(0.000)
Num.Obs.	213 932	167 551	213 932	167 551	213 932	167 551
R2	0.044	0.028	0.020	0.016	0.085	0.057
R2 Adj.	0.044	0.028	0.013	0.007	0.078	0.049
AIC	1 135 698.2	1 015 340.4	1 126 455.8	1 010 224.1	1 126 455.8	1 010 224.1
BIC	1 135 770.1	1 015 410.6	1 126 527.7	1 010 294.3	1 126 527.7	1 010 294.3
Log.Lik.	−567 842.110	−507 663.218				
RMSE	3.44	5.01	3.37	4.93	3.37	4.93

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

In the checks for robustness in the table above, among the OLS, PLM, and two-way fixed-effects models, we consistently observed strongly statistically significant influences of the residency dummy variable on tipping.

In the model considering double-tariff rate (flexible rate), our investigation revealed that, conditioning on other regressors, the results for residential passengers is similar across the OLS and fixed effects models. For the former, residential passengers tip approximately $0.201 less than tourists, for the latter they tip $0.248 less (*see Table VI*) - these changes are on average and controlling for other variables. This outcome deviates from our initial hypothesis. Explanation for this observed variance may be explained by the fact that non-local passengers possess greater financial resources or that their expenses are potentially covered by their employers if they visit the city for a business trip. Our findings align with those observed by Neto et al. (2019), who propose that tourists tend to tip more than local individuals.

On the other hand, the model including flat rate exhibits the opposite phenomenon - local passengers spend more on tipping than their counterparts. This result fulfilled our initial hypothesis that local passengers are more likely to be regular customers who frequent the same establishments regularly. This familiarity with service providers could lead to a stronger sense of loyalty and a desire to reward good service through higher tips. The effect is yet again not largely different between the OLS and FE models, both account for approximately, on average and ceteris paribus, $0.773 more tip when we consider for locals.

When comparing the results we estimated for both types of price schemes, we can see that our results confirm our hypothesis. Locals indeed tip more when they opt for a taxi ride which is charged according to flat fare.

Conclusion

The taxi drivers' behavior is predetermined by vast numbers of factors. Personal-characteristics matter as well as the trip conditions and the vision of better financial possibilities. The initial study convincingly shows that drivers react to financial incentives by taking extra detours under a two-part tariff, especially with lower occupancy rates and higher variable rates. These detours were estimated to cost passengers $0.5-0.8 million annually.

In our extension, we explored potential disparities in tipping patterns between local residents and tourists, with a particular focus on New York City. Utilizing the geo-coordinates of drop-off locations, we distinguished passengers going to a specific set of hotels to be tourists and to a specific residential area within close proximity to hotels to be locals .

Our analysis reveals a noteworthy distinction: tourists exhibit a higher propensity to tip compared to locals when we consider the double-tariff price scheme. To be specific, the findings indicate that tourists tend to tip on average around $0.25 more than their local counterparts, given on double-tariffs rates and control of other variables. On the other hand, the outcome is different for controlling flat rate, locals tip approximately $0.773 more on average given the set of controls. While the impact is marginally more positive when conditioning on local passengers and considering flat rates versus double tariffs, we were unable to substantiate our hypothesis and corresponding assumptions. The observations indicating higher tipping expenditures by tourists compared to resident passengers align with existing literature (Neto et al. 2019).

For future research, it would be valuable to gather additional insights from both passengers and drivers. Factors such as passengers' gender, income, cultural norms, and education level should be considered. Additionally, investigating the quality of drivers' service appears to be desirable as well. Interesting alternative to our extension is exploring the relationship between stock market performance and tipping behavior, as existing research indicates that employees from public companies are inclined to tip taxi drivers more generously when their company performs well in the stock market.

References

Azar, O. H. (2011). Business strategy and the social norm of tipping. Journal of Economic Psychology, 32(3), 515–525.

Chandar, B., Gneezy, U., List, J. A., & Muir, I. (2019). The drivers of social preferences: Evidence from a nationwide tipping field experiment (No. W26380). National Bureau of Economic Research.

Davis, L.W. (2008). The effect of driving restrictions on air quality in Mexico City. Journal of Political Economy, 116(1), 38-81.

DiNardo, J., & Lee, D. S. (2004). Economic impacts of new unionization on private sector employers: 1984–2001. The Quarterly Journal of Economics, 119(4), 1383-1441.

Dulleck, U., & Kerschbamer, R. (2006). On doctors, mechanics, and computer specialists: The economics of credence goods. Journal of Economic literature, 44(1), 5-42.

Grytten, J., Sørensen, R., & Sorensen, R. J. (2011). Price and quality when professionals meet patients: an empirical study of physicians in primary care. International Journal of Health Care Finance and Economics, 11(1), 31-47.

José I. Castillo-Manzano and Antonio Sánchez-Br. (2010)An Evaluation of the Establishment of a Taxi Flat Rate from City to Airport: The Case of Seville. Urban Studies, Vol. 48, No. 9 1909-1924.

Levitt, S. D., & Syverson, C. (2008). Market distortions when agents are better informed: The value of information in real estate transactions. The Review of Economics and Statistics, 90(4), 599-611.

Liu, T., Vergara-Cobos, E., & Zhou, Y. (2017). Pricing schemes and seller fraud: Evidence from New York City taxi rides. Working Paper.

Lynn, M. (2015). Explanations of service gratuities and tipping: Evidence from individual differences in tipping motivations and tendencies. Journal of Behavioral and Experimental Economics, 55, 65–71.

Neto, A. B. F., Nowak, A., & Ross, A. (2019). Do Tourists Tip More Than Local Consumers? Evidence from Taxi Rides in New York City. International Regional Science Review, 42(3-4), 281-306.

Wolinsky, A. (1993). Competition in a market for informed experts' services. The RAND Journal of Economics, 380-398.

Appendix

Summary Statistics Tables

```
# This script calculates summary statistics for the data based upon each respective paper

library(dplyr)
library(tidyr)
library(knitr)
library(lubridate)
library(data.table)
library(kableExtra)


# Create the summary table
driver_summary_table <- taxi %>%
  group_by(PULocationID) %>%
  summarize(
    num_trips = n(),
    # Round up to the nearest whole number
    working_hours = sum(ceiling(ifelse(dropoff_hour < pickup_hour, 24
                                       + dropoff_hour - pickup_hour, trip_duration / 60))),
    # Convert trip_duration to hours
    occupied_hours = sum(ifelse(passenger_count > 0, trip_duration / 60, 0)),
    occupancy = occupied_hours / working_hours
  )

#Rename PUlocationID to Airport
driver_summary_table <- driver_summary_table %>%
  rename(Airport = PULocationID)

#Change Airport values to JK and LGA
driver_summary_table$Airport[driver_summary_table$Airport == 132] <- "JFK"
driver_summary_table$Airport[driver_summary_table$Airport == 138] <- "LGA"

driver_summary_table %>%
  kbl(caption = "Table II: Driver Characteristics") %>%
  kable_classic(full_width = F, html_font = "Georgia")

# TABLE III #
#Create table summary statistics including the number of trips,
#the average trip distance, duration and fare
#This is done for each respective departure location further breaking down into drop off location
#Each departure location is to be broken down into "All" and then residential and hotel
#Residential is defined as a trip that ends in DOLocationID = 229, otherwise it is a hotel trip

table_summary <- taxi %>%
  group_by(PULocationID, trip_type) %>%
  summarize(
    num_trips = n(),
    avg_trip_distance = mean(trip_distance),
    avg_duration = mean(difftime(tpep_dropoff_datetime,
                                 tpep_pickup_datetime, units = "mins")),
    avg_fare = mean(fare_amount)
  )
```

```r
#Rename PUlocationID to Airport
table_summary <- table_summary %>%
  rename(Airport = PULocationID)

#Change Airport values to JK and LGA
table_summary$Airport[table_summary$Airport == 132] <- "JFK"
table_summary$Airport[table_summary$Airport == 138] <- "LGA"

table_summary %>%
  kbl(caption = "Table III: Summary Statistics by Departure Location and Trip Type") %>%
  kable_classic(full_width = F, html_font = "Georgia")
```

Main Formulation and Estimation

```r
# This script calculates summary statistics for the data based upon each respective paper

library(dplyr)
library(tidyr)
library(knitr)
library(lubridate)
library(data.table)

#First, we read the parquet file into R
library(arrow)
# taxi <- read_parquet("sample_data.parquet")
taxi <- read_parquet("sample_data_big.parquet")

# Create a new variable to identify residential or hotel trips
taxi <- taxi %>%
  mutate(trip_type = ifelse(DOLocationID == 229, "Residential", "Hotel"))

# Feature engineering: Extract hour and calculate trip duration
taxi <- taxi %>%
  mutate(
    pickup_hour = hour(tpep_pickup_datetime),
    dropoff_hour = hour(tpep_dropoff_datetime),
    trip_duration = as.numeric(difftime(tpep_dropoff_datetime,
                                        tpep_pickup_datetime, units = "mins")),
    # Round up to the nearest whole number
    # working_hours = ceiling(as.numeric(difftime(dropoff_hour, pickup_hour, units = "hours")))
  )

#Remove trips where the pickup = 132 (JFK) has a distance less than 10 and more than 40
#And where the pickup = 138 (LGA) has a distance less than 5 and more than 20
taxi <- taxi %>%
  filter(trip_duration > 0) %>%
  filter(!(PULocationID == 132 & (trip_distance < 10 | trip_distance > 40))) %>%
  filter(!(PULocationID == 138 & (trip_distance < 5 | trip_distance > 20)))

#Create a new variable to identify the respective trip type based on departure and drop off location
#By combining the ID's of the departure and drop off location, we can identify the trip type
#For example, if the departure location is 132 (JFK) and the drop off location
```

```r
#is 229 (Residential), then the trip type is 132229
taxi <- taxi %>%
  mutate(tripID = paste(PULocationID, DOLocationID, sep = ""))

#Create a policy treatment variable to identify when the policy intervention took place:
#1 = after, 0 = before
#The policy took effect on September 4, 2012, at 12:01 am
taxi <- taxi %>%
  mutate(treatment = ifelse(tpep_pickup_datetime >= as.POSIXct("2012-09-04 00:01:00"), 1, 0))




############# VARIABLE CONSTRUCTION ############
## EXPECTED OCCUPANCY ##
#This is a trial attempt at recreating the expected occupancy from the paper to be used as a covaria
#This is done with regressing the actual occupancy on a number of time based indicator controls
#This includes year indicators (0-3), hour of day by day of week (0-167), week of year (0-51)

#Create year indicator (0-3)
taxi <- taxi %>%
  mutate(
    year = year(tpep_pickup_datetime) - 2011
  )

#Create hour of day by day of week indicator (0-167)
taxi <- taxi %>%
  mutate(
    hour_of_day = hour(tpep_pickup_datetime),
    day_of_week = wday(tpep_pickup_datetime),
    hour_of_day_by_day_of_week = hour_of_day + (day_of_week - 1) * 24
  )

#Create week of year indicator (0-51)
taxi <- taxi %>%
  mutate(
    week_of_year = week(tpep_pickup_datetime) - 1
  )

#Reformulation of actual occupancy
taxi <- taxi %>%
  mutate(
    pickup_hour = hour(tpep_pickup_datetime),
    dropoff_hour = hour(tpep_dropoff_datetime),
    trip_duration = as.numeric(difftime(tpep_dropoff_datetime,
                                        tpep_pickup_datetime, units = "mins"))
  )

taxi <- taxi %>%
  mutate(
    working_hours = ceiling(ifelse(dropoff_hour < pickup_hour, 24
                                   + dropoff_hour - pickup_hour, trip_duration / 60)),
    occupancy = trip_duration / (working_hours*60)
```

```
  )

#Regression of controls on actual occupancy
occ_model <- lm(occupancy ~ year + hour_of_day_by_day_of_week + week_of_year,
                data = taxi)

#Extract the predicted occupancy
occ_predict <- predict(occ_model, newdata = taxi)

#Add the predicted occupancy to the taxi data
taxi$pred_occ <- occ_predict



### DETOUR INDEX ###
# !!!! RUN THE DETOUR INDEX SCRIPT FIRST !!!!
#Join the detour indices from result_df to the main taxi data, matching based on
#"VendorID", "tpep_pickup_datetime", "PULocationID" and "DOLocationID".
#This would add the detour_distance, detour_duration, and detour_fare columns to the main taxi data
taxi <- taxi %>%
  left_join(result_df, by = c("VendorID", "tpep_pickup_datetime",
                              "PULocationID", "DOLocationID"))



### ESTIMATION PREPARATION ###
#Make tripID numeric
taxi <- taxi %>%
  mutate(tripID = as.numeric(tripID),
         pred_occ = as.numeric(pred_occ))

#Drop duplicates for tripID and dropoff time
taxi <- taxi %>%
  distinct(tripID, tpep_dropoff_datetime, .keep_all = TRUE)

#Create a date variable as a categorical variable denoting the day out of the entire data period (1,2,
# taxi <- taxi %>%
#   mutate(date = as.numeric(as.Date(tpep_pickup_datetime) - as.Date("2011-01-01")))

#Create a date variable from the datetime of dropoff
taxi <- taxi %>%
  mutate(date = as.Date(tpep_dropoff_datetime))


#Export the taxi data to a stata dta file
# library(haven)
# write_dta(taxi, "taxi.dta")

#Drop records where the detour distance, detour duration, or detour fare are in the
#lowest 1% or highest 1% of their respective distributions
#This is done to remove outliers, excluding NaN values
taxi <- taxi %>%
  filter(detour_distance > quantile(detour_distance, 0.01, na.rm = TRUE)
```

```
        & detour_distance < quantile(detour_distance, 0.99, na.rm = TRUE)) %>%
  filter(detour_duration > quantile(detour_duration, 0.01, na.rm = TRUE)
        & detour_duration < quantile(detour_duration, 0.99, na.rm = TRUE)) %>%
  filter(detour_fare > quantile(detour_fare, 0.01, na.rm = TRUE)
        & detour_fare < quantile(detour_fare, 0.99, na.rm = TRUE))



#### "DiD" IMPLEMENTATION ####
#The following seeks to replicate Table V from the paper
#We will estimate each column of the table separately to fall in line with the empirical specification
library(fixest)

# Column 1: LGA Distance
lga_d <- taxi %>%
  filter(PULocationID == 138) %>%
  feols(detour_distance ~  pred_occ + treatment + treatment*pred_occ | tripID + date,
        cluster = "tripID")

# summary(lga_d)

# Column 2: LGA Duration
lga_t <- taxi %>%
  filter(PULocationID == 138) %>%
  feols(detour_duration ~  pred_occ + treatment + treatment*pred_occ | tripID + date,
        cluster = "tripID")

# summary(lga_t)

# Column 3: LGA Fare
lga_f <- taxi %>%
  filter(PULocationID == 138) %>%
  feols(detour_fare ~  pred_occ + treatment + treatment*pred_occ | tripID + date,
        cluster = "tripID")

# summary(lga_f)

# Column 4: JFK Distance
jfk_d <- taxi %>%
  filter(PULocationID == 132) %>%
  feols(detour_distance ~  pred_occ + treatment + treatment*pred_occ | tripID + date,
        cluster = "tripID")

# summary(jfk_d)

# Column 5: JFK Duration
jfk_t <- taxi %>%
  filter(PULocationID == 132) %>%
  feols(detour_duration ~ pred_occ + treatment + treatment*pred_occ | tripID + date,
        cluster = "tripID")

# summary(jfk_t)
```

```
models <- list(lga_d, lga_t, lga_f, jfk_d, jfk_t)

#Export
myDict <- c("pred_occ" = "occ",
            "treatment" = "g",
            "treatment:pred_occ" = "g x occ",
            "detour_distance" = "Distance",
            "detour_duration" = "Duration",
            "detour_fare" = "Fare",
            "tripID" = "Trip Unit",
            "date" = "Date",
            "PULocationID" = "Pickup Location",
            "DOLocationID" = "Dropoff Location",
            "tip_amount" = "Tip Amount",
            "local" = "Local dummy",
            "fare_amount" = "Fare Amount",
            "trip_distance" = "Trip Distance",
            "trip_duration" = "Trip Duration",
            "passenger_count" = "Passenger Count",
            "hour_of_day_by_day_of_week" = "Hour of Week")

etable(lga_d, lga_t, lga_f, jfk_d, jfk_t,
       dict = myDict,
       title = "Estimations Results of Equation (1)",
       headers = c("LGA Distance", "LGA Duration", "LGA Fare",
                   "JFK Distance", "JFK Duration"),
       # file = "tableV.tex",
       # replace = TRUE,
       view = TRUE,
       depvar = FALSE)




########### EXTENSION #############
#Create a dummy variable for whether the trip is done by a local
taxi <- taxi %>%
  mutate(local = ifelse(DOLocationID == 229, 1, 0))

#OLS
# Column 1: LGA
ols_lga <- taxi %>%
  filter(PULocationID == 138) %>%
  lm(tip_amount ~ local + trip_duration + trip_distance + fare_amount +
       passenger_count, data = .)

# Column 2: JFK
ols_jfk <- taxi %>%
  filter(PULocationID == 132) %>%
  lm(tip_amount ~ local + trip_duration + trip_distance + fare_amount +
       passenger_count, data = .)

#PLM
```

```
library(plm)

# Column 3: LGA
plm_lga <- taxi %>%
  filter(PULocationID == 138) %>%
  plm(tip_amount ~ local + trip_duration + trip_distance + fare_amount +
        passenger_count + hour_of_day_by_day_of_week, data = ., index = c("date"),
      model = "within")

# Column 4: JFK
plm_jfk <- taxi %>%
  filter(PULocationID == 132) %>%
  plm(tip_amount ~ local + trip_duration + trip_distance + fare_amount +
        passenger_count + hour_of_day_by_day_of_week, data = ., index = c("date"),
      model = "within")

#TWFE
library(estimatr)

# Column 5: LGA
twfe_lga <- taxi %>%
  filter(PULocationID == 138) %>%
  lm_robust(tip_amount ~ local + trip_duration + trip_distance +
              fare_amount + passenger_count + hour_of_day_by_day_of_week,
            fixed_effects = ~ date,
            data = .,
            se_type = "stata")

# Column 6: JFK
twfe_jfk <- taxi %>%
  filter(PULocationID == 132) %>%
  lm_robust(tip_amount ~ local + trip_duration + trip_distance +
              fare_amount + passenger_count + hour_of_day_by_day_of_week,
            fixed_effects = ~ date,
            data = .,
            se_type = "stata")

model_ext <- list("LGA OLS" = ols_lga, "JKF OLS" = ols_jfk, "LGA PLM" = plm_lga,
                  "JFK PLM" = plm_jfk, "LGA TWFE" = twfe_lga, "JFK TWFE" = twfe_jfk)

#Table Export
library(modelsummary)
library(kableExtra)
modelsummary(model_ext,
             stars = TRUE,
             coef_map = myDict,
             title = "Estimations Results of Equation (2) with Robustiness Checks",
             output = "ext_results.tex"
)
```

Detour Indices Construction

```
# Consolidated solution for Detour Index
library(dplyr)
library(lubridate)

# Assuming your dataset is named 'taxi'
taxi$tpep_pickup_datetime <- as.POSIXct(taxi$tpep_pickup_datetime)
taxi$tpep_dropoff_datetime <- as.POSIXct(taxi$tpep_dropoff_datetime)

# Filter for non-local passengers
nonlocal_trips <- taxi %>%
    filter(PULocationID %in% c("132", "138") & DOLocationID %in%
        c("163", "161", "162", "164"))

# Set time window (e.g., 15 mins)
time_window <- 15 * 60

# Define a function to calculate detour indexes for a given
# airport-hotel trip
calculate_detour_indexes <- function(trip, comparison_trips,
    time_window) {
    comparison_trips <- comparison_trips %>%
        filter(tpep_pickup_datetime >= (trip$tpep_pickup_datetime -
            minutes(time_window)) & tpep_pickup_datetime <= (trip$tpep_pickup_datetime +
            minutes(time_window)) & DOLocationID != trip$DOLocationID)

    avg_distance <- mean(comparison_trips$trip_distance)
    avg_duration <- mean(comparison_trips$trip_duration)
    avg_fare <- mean(comparison_trips$fare_amount)

    detour_indexes <- trip %>%
        summarise(detour_distance = trip_distance - avg_distance,
            detour_duration = trip_duration - avg_duration, detour_fare = fare_amount -
                avg_fare)

    return(detour_indexes)
}

# Apply the function to each airport-hotel trip
detour_indexes <- nonlocal_trips %>%
    group_by(across(c("VendorID", "tpep_pickup_datetime", "PULocationID",
        "DOLocationID"))) %>%
    do(calculate_detour_indexes(., nonlocal_trips, time_window))

# Combine the results into a data frame
result_df <- bind_rows(detour_indexes)



#### Index Stats ####

# View the mean of each detour index based on pickup
```

```r
# location
detour_summary <- result_df %>%
    filter(!is.na(detour_distance) & !is.na(detour_duration) &
        !is.na(detour_fare)) %>%
    group_by(PULocationID) %>%
    summarise(avg_detour_distance = mean(detour_distance), avg_detour_duration = mean(detour_duration)
        avg_detour_fare = mean(detour_fare))

# Rename PUlocationID to Airport
detour_summary <- detour_summary %>%
    rename(Airport = PULocationID)

# Change Airport values to JK and LGA
detour_summary$Airport[detour_summary$Airport == 132] <- "JFK"
detour_summary$Airport[detour_summary$Airport == 138] <- "LGA"

detour_summary %>%
    kbl(caption = "Table IV: Mean Detours by Airport") %>%
    kable_classic(full_width = F, html_font = "Georgia")
```

Task Allocation

I am assuming the project is equally contributed by team members. If this is not the case, it is optional for you to let me know the division of labor by assigning a percentage contribution of each of the team members on the last page of the paper, provided there is a consensus of the contributions.

111266012 莫喬丹 Jordan Murillo	111266004 白季堯 Josef Blazek	111266007 張靜宜
40%	40%	20%