Riding to Wall Street: Determinants of Commute Time using Citibike

Weixing Ford
Texas A&M - San Antonio

Jaimie W. Lien
The Chinese University of Hong Kong

Vladimir V. Mazalov
Russian Academy of Sciences

Jie Zheng
Tsinghua University

Current version: January 17, 2019
Original version: August 25, 2018

Abstract:[1]

Commuting to work by shared bicycle is an increasingly popular transportation option. We examine the trip data from Citibike, the shared bicycle program of New York City, restricting analysis to morning rides into the Wall Street area, a destination for which workers are likely to be sensitive to their arrival time. Among the frequent commute origins, commuting later in the morning lengthens the time on the road, while demographic variables such as gender, age and subscription also significantly predict commute time. Congestion appears to increase for bicycle commuters directly before financial markets open. Weather-related factors have a small impact on biking commute time compared to other factors. When considering the model specifications as potential user models of commute time, the bicycle commuters outperformed the models for shorter commutes and for commutes occurring later in the morning, which is consistent with the ability and incentive to increase efforts to reduce their riding time under these conditions.

Keywords: bicycle commuting, shared bicycle, New York City, trip duration prediction

[1] Ford: Department of Management and Marketing, College of Business, Texas A&M University - San Antonio, weixing.ford@tamusa.edu; Lien: Department of Decision Sciences and Managerial Economics, The Chinese University of Hong Kong, jaimie.academic@gmail.com; Mazalov: Institute of Applied Mathematical Research, Karelian Research Center, Russian Academy of Sciences, vlmazalov@yandex.ru ; Zheng: Department of Economics, School of Economics and Management, Tsinghua University, jie.academic@gmail.com

We are grateful to the Editors and three anonymous referees for their suggestions which greatly improved the paper. For helpful comments and encouragement, we thank Prof. Yacan Wang, as well as Guenter Emberger, Meihan Jin, Takeru Shibayama, and participants of the 6th International Conference on Transportation and Space-time Economics, WCTR Workshop, Beijing Jiaotong University. The authors gratefully acknowledge research funding from Hong Kong Research Grants Council (Project No. 14500516), Chinese University of Hong Kong Direct Grants, National Natural Science Foundation of China (Projects No. 61661136002 and No. 71873074), Tsinghua University Initiative Scientific Research Grant (Project No. 20151080397), the Russian Fund for Basic Research (Project No.16-51-55006). All errors are our own.

Launched in 2013, New York City's successful CitiBike program is currently the largest bike sharing program in the United States, serving as a model for similar programs in other urban areas. Shared bicycle programs have also found remarkable success and competition in cities across China, including Ofo and Mobike, which have subsequently extended their business to outside of China. Shared bicycle programs charge customers a lump sum fee for a time period to ride as many times as the customer chooses. This fee is often quite affordable compared to other transportation options, making cycling to work a plausible and attractive option for many commuters, without having to worry about the storage and maintenance of a personal bicycle.[2]

Despite the obvious appeal of cycling to work, bikes and/or shared bikes comprise a relatively small fraction of commute methods nationwide. New York City is one urban area in the United States where such a program is likely to succeed commercially and yield social benefits. First, the density of residents and workers is high, and the landscape is relatively flat compared to more hilly cities such as San Francisco (Cevero and Duncan, 2013). Workers in New York City are also well-known for having some of the most demanding daily commutes, with an estimated 1/8 of Manhattan workers spending more than 90 minutes each way on their daily commute (Moss, Qing and Kaufman, 2012).[3] NYC commuters also rely heavily on public transportation (3/4th of Manhattan commutes; Moss, Qing and Kaufman, 2012) and less on driving individual cars to work than in other areas of the country. New York City therefore, has served as an important test for the success of similar bike sharing systems in other urban areas.

Although bicycle sharing programs are undoubtedly well-utilized and successful, how practical are they for commuting to work? What are the factors determining commute time on a shared bicycle, and how does the performance of commutes over time and distance compare to basic model predictions of commute time? We study these questions by targeting our analysis to those trips which are most likely to be time sensitive commutes to work: Citibike rides whose destinations are the financial district of Manhattan during the morning commute hours.

In major metropolitan areas, subway and other public transport systems can get overcrowded and congested, creating a demand for alternative transport methods and incentive systems. Wang, Ettema, Zhou and Sun (2018) examine the effectiveness of proposed financial incentives for avoiding peak travel time on the Beijing subway system, finding that commuters are receptive to such incentives. The wearing down of older infrastructure can also pose a challenge for transportation systems, as studied by Wheat and Smith (2008). Another approach to urban congestion is to provide alternative transportation such as shared bicycle systems. However, the implementation and maintenance of such systems can also present logistical challenge, such as in Lian, Si, Jiao and Li (2018) which propose a management scheme for the repair of damaged shared bikes.

A number of studies seek to understand individuals' motives for using bicycles as a form of transportation. Research which examines the effect of weather on bicycling activity includes Brandenburg, Matzarakis and Arnberger (2004, 2007) which focus on the sensitivity of commuting and leisure cyclists to weather factors. Martinez (2017) assesses the impact of weather factors on Citibike ridership, finding that higher temperatures, lower precipitation and wind speed enhance the number of daily trips taken. Compared to our study, Martinez (2017) studies trip volume whereas our

[2] Currently, Citibike charges $169 US per year, or $14.95 per month for unlimited 45 minute rides. A short term pass option for visitors is offered at $12 for 24 hours of access, or $24 for 72 hours of access, unlimited 30 minute rides. Alternatively, a single use pass sells for $3 for a 30 minute ride.

[3] Simonsohn (2006) found that New Yorkers were most tolerant of long commute times, even after moving out of the New York City area to other parts of the country.

current study focuses on trip duration. Several studies examine the psycho-social factors which contribute to the decision to engage in bicycle travel, including Gatersleben and Appleton (2007) and de Geus (2007). Bernhoft and Carstensen (2008) focus on the preferences and behavior of cyclists based on demographic factors.

Faghih-Imani, Anowar, Miller and Eluru (2017) compare the travel times by taxi and by Citibike, finding that the bike sharing mode is competitive or faster than taxi travel. Their study conducts a more general assessment of the Citibike program, not exclusive to commuters per se, but focusing on the relative advantages of taxi versus bike at different times of day. Our study focuses specifically on morning commuters to the financial district, and in doing so hopes to capture the patterns with respect to a time sensitive sub-population of commuters which may not be detectable otherwise among a more general heterogeneous population. Leth, Shibayama and Brezina (2017) study the bike sharing system in Vienna, finding that it plays a supplementary role to the public transport system. Heinen, van Wee and Maat (2009) provide an overall survey the literature of commuting by bicycle, focusing on the determinants of the decision to cycle to work.

Our approach of studying individuals using the Citibike system for rides into the financial district of New York City are threefold. First, individuals biking into the financial district in the morning of a workday are more likely to be commuting to work than those biking into an area of a city which is known for sightseeing, dining or other attractions. This serves our research purpose well since we are interested in isolating commuting behavior and performance to the extent possible, despite not having direct information about whether the ride was a commute or not. Second, workers in the financial district are more likely to be sensitive and face consequences regarding the time that they arrive at work, due to real time financial market activity which likely influences their job performance. Thus, financial district workers have strong incentive for timely commutes, and are simultaneously less flexible in their hours than many other professions. This implies that financial district workers may have quite a high incentive to optimize their commuting strategy, whether by bicycle or other transportation means. Third, workers in the financial district are often expected to dress reasonably nicely, in contrast to other industries which may have a more relaxed dress code. This is likely to serve as a deterrent for many financial district workers in choosing to bike to work, since it may pose an extra inconvenience to them. Thus, the appeal of commuting to work by bicycle must hold a substantial appeal to workers along some dimension in order to outweigh this inconvenience. For these reasons, financial district bicycle rides on Citibike are a potentially interesting subset of rides to study, which can be indicative on some levels, of the feasibility and effectiveness of shared bicycle programs more broadly.

The remainder of the paper proceeds as follows: Section 2 describes the data and sample, Section 3 describes the empirical approaches and main regression results, Section 4 considers the performance and distribution of individual trips compared to the model predictions based on arrival time and distance traveled, Section 5 summarizes and discusses future directions.

2. Data

The trip data from Citibike are publicly available, consisting of trips taken from the initial launch until the present.[4] The datasets consist of starting and ending station information, time information, and basic demographic characteristics of the user. The data is anonymized and does not explicitly identify individual subscribers across trips. We utilize a full year of data from the time period July

[4] https://www.citibikenyc.com/system-data

2016 to June 2017.

The data volume is large and in order to meaningfully narrow down the sample on morning commuters, we focus the analysis using some reasonable assumptions. We focus on those rides which end at one of the 23 stations in the Wall Street area, as identified by the Citibike station map. These stations are shown in Table 1.

Table 1: Destination Stations in Financial District Sample (official Citibike names)

Warren St & Church St
Murray St & Greenwich St
Murray St & West St
Barclay St & Church St
Centre St & Chambers St
Gold St & Frankfort St
Fulton St & Broadway
Fulton St & William St
John St & William St
Liberty St & Broadway
Cliff St & Fulton St
Peck Slip & Front St
Maiden Ln & Pearl St
William St & Pine St
Front St & Maiden Ln
Pearl St & Hanover Square
Old Slip & Front St
South St & Gouverneur Ln
Broadway & Battery Pl
Broad St & Bridge St
Water - Whitehall Plaza
Bus Slip & State St
South St & Whitehall St

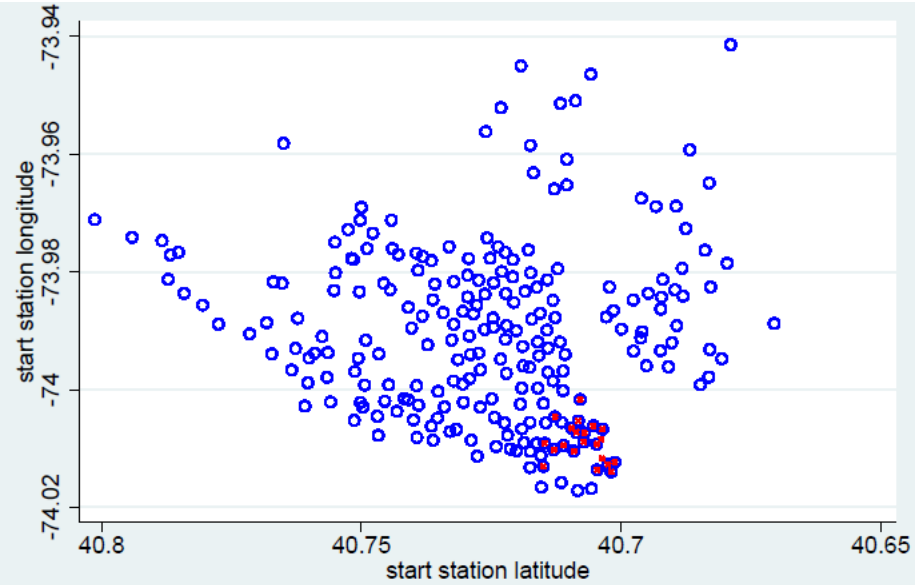
Compared to workers in other industries, those employed in the financial district are plausibly more pressured and motivated to arrive at work at a particular time, due to the volatile occurrences in the financial markets which affect their job performance outcomes. For similar reasons, we consider only the morning commute, those rides commencing between the hours of 6am and 10am, whereas in the afternoons workers are more likely to leave at their discretion and remaining duties which are more heterogeneous across workers. In other words, as much as possible given the limited information in the data, we hope to capture those cases where riders have strong incentive to arrive on time. Weekends and holidays in which the markets are closed are also omitted from the sample.

Although the destination stations have been narrowed down substantially, the total number of routes is still very large due to the wide variety of locations that users originate their rides from. Since we conduct route fixed-effect analysis, we further narrow the routes in the data to those which appear more than 60 times in the data over the 12 month sample period. In other words, the route must be sufficiently common to be included in the analysis. The exact numerical cutoff on the number of times a route appears in the data does not affect the main qualitative results in terms of

significance and magnitude of coefficients. After this criterion is imposed, the map of dock stations in our sample is shown in the longitude and latitude plot in Figure 1. The concentration of red dots corresponds to the financial district area of Manhattan, and the open blue dots are the starting stations in the same or neighboring areas and boroughs. Using this data sample, summary statistics of some main variables in the analysis are provided in the Appendix Table A1.

Figure 1: Location of Dock Stations

Red: end station in sample; blue: start station in sample

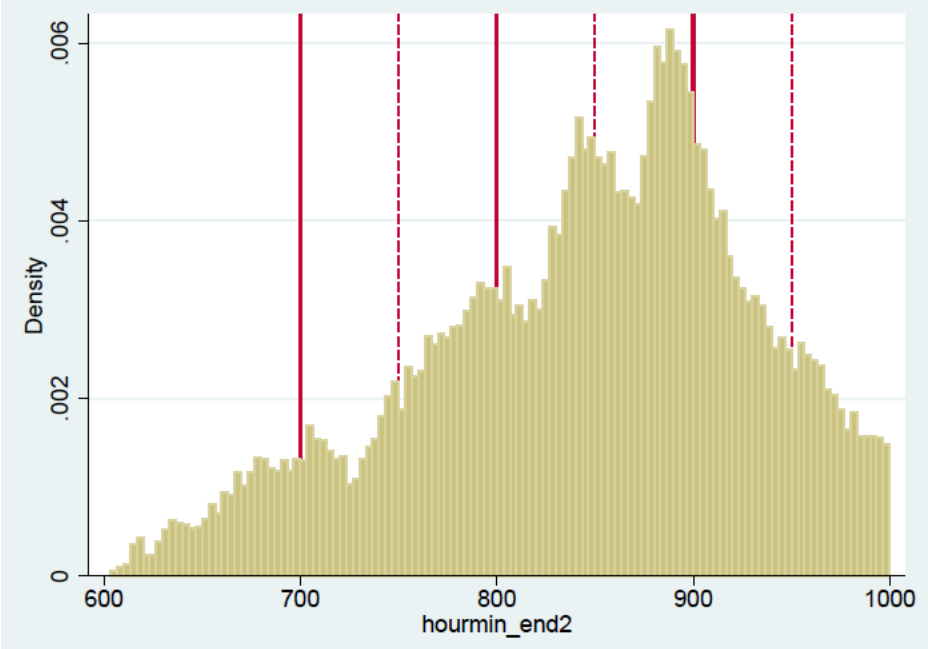


Finally, we download the daily weather information for the Manhattan area from the website Weather Underground, which archives historical daily information on rainfall, humidity, temperature and other weather-related variables.

Figure 2 shows the density of arrival times by hour and 2-minute intervals. As we can see, the peak arrival hour is in the 8'oclock interval. Furthermore, within each hour-long interval, the arrivals tend to be concentrated towards the end of the hour, with the exception of the 9'oclock interval, which has heavier concentration close to 9:00am, and tapers off as 10:00am approaches. In the density plot, we can observe the additional tendency for commuters to arrive before 7:15am, before 8:00am, around 8:30am and before 9:00am. We note that the financial markets officially open at 9:30am, which provides a strong incentive for workers to arrive sometime before then.

A comparison between the density of arrival times for the financial district as in Figure 2, and the analogous density for all destination stations as provided in the Appendix (Figure A1) reveals three main differences. First, the distribution of arrival times across all stations in the city is smoother than the density for the financial district alone. Secondly, the financial district sample has a heightened proportion of bicycle users arriving just before 7:15, 8:00 and 8:30 than the total sample. Finally, the drop-off in arrivals past 9:30am is more severe than in the total sample, density which is distributed throughout earlier in the morning. These differences between the arrival time density of the financial district and the total sample are indicative of the higher time urgency of riders headed for the Wall Street area in the morning hours.

Figure 2: Density of Arrival Times



3. Main Results

Our empirical analysis focuses on trip duration as the dependent variable. We consider three main types of specifications: A least squares regression with route fixed-effects, designed to capture the average influence of covariates within each of the frequent routes; Least squares regressions substituting route distance instead of fixed effects, adding groups of covariates incrementally, intended as a psychologically realistic modeling approach under situations where route frequencies in the data are not necessarily high; Finally we conduct a survival analysis on commute time as a robustness check, which allows for the trip ending probability to vary with time elapsed. In fact, all the modeling approaches yield quite similar results.

We first consider an econometric specification which fully controls for route-specific effects. In other words, any route-specific factors are absorbed into the route fixed-effects, and the coefficients are the average within-route variation across all routes in the sample. Using trip duration (measured in seconds) as the dependent variable, we estimate the least squares regression:

$$tripduration_{j(,i,t)} = \mu_r + \beta \cdot timefeatures_{j(,t)} + \gamma \cdot demogr_{j(,i)} + \delta \cdot weather_j + \varepsilon_j$$

where *timefeatures* is a trip-specific vector of characteristics, including particular features of the start and ending times of the ride. We consider chronological departure time and its squared term, as well as indicator variable for features of the arrival time. "end time before 30, 00" corresponds to and arrival time which is up to three minutes prior to the round number value of minutes 30, or 00 exactly. "End time after 30, 00" corresponds to the three minutes afterwards. The analogous definitions hold for the "15, 45" variables. To account for different hour effects, we include dummy variables for each of the arrival hours in our sample, using the 7 o'clock interval as the comparison point. *Demogr* are the observable demographic features of the account holder, including age, gender and subscription status. We only have these variables attached to each trip observation, but we cannot observe whether the same individual took two different trips in the data based on only these coarse demographic variables. *Weather* are the weather-related variables on the day the trip was taken, thus can be

represented as trip-specific. μ_r are the route fixed- effects, where recall that our sample is restricted to those routes that appear in the total 12 month dataset more than 60 times. The subscripts in parentheses merely remind us that although we are dealing with essentially cross-sectional data, we do have time and person-related demographic information for each observation.

The leftmost column of Table 2 (baseline) shows the result of this specification. Among the time variables, the dummy variables for hour of arrival are statistically significant. Controlling for the specific route, which means that distance is also controlled for, all else equal, an arrival in the 9'oclock interval takes significantly longer than an arrival in the 8'oclock interval, which is significantly longer than the 7'oclock intervals, followed by the 6'oclock interval which has the shortest estimated commute time. The demographic variables indicate a slightly convex relationship between birth year and trip duration, with the younger riders getting to their destination faster. Females lag behind males in travel time by slightly over a minute, all else equal, while non-subscribers complete their journeys around 6 minutes slower.

Out of the weather variables considered, only daily low temperature has a statistically significant coefficient, and the magnitude of the effect is small, with higher daily low temperatures corresponding to shorter commute times. Since our weather data is on a daily average basis, it is also possible that it does not capture the exact weather conditions at the time of commute. For this reason, we use the low temperature variable, which may be more likely to correspond to an early morning temperature. Another likely possibility is that weather variables affect the decision about whether commute by bicycle, but conditional on choosing to commute by bicycle, the effect on the commute time itself is not substantial. Indeed, in our other empirical specifications considered in this paper, whenever the effect of weather is significant, it tends to be a small effect compared to that of other explanatory variables, and often not robust across model specifications.

In the middle column of Table 2 (market open), we eliminate most of the time related variables, except that we allow for a market opening rush hour interval around 9:30am. Arriving in the few minutes directly before 9:30am corresponds to significantly longer travel time, which is indicative of traffic congestion leading up to the market opening. Demographic characteristics remain of similar significance and magnitude.

In the right column of Table 2 (basic), we eliminate those variables from the baseline model which might be psychologically less plausible. For example, riders may have some idea that starting times and their own age affects how long their bicycle commute would take (ex. an older person may recognize that commute speed declines with age), but they may not know the nonlinear features of this relationship. Similarly, commuters may not know the average temperature, but they may know the morning temperature which corresponds to the daily low. Humidity may also be difficult for individuals to gauge, particularly in terms of how it would affect their commuting time. The results for this specification are much the same as in the baseline specification, including the general pattern that younger riders arrive at their destinations faster. When average temperature and humidity are omitted, the effect of an increase in daily low temperature is to slightly increase the commute time, while wind speed reduces it.

We note here that the relatively small impact of weather variables on commute time could be in part driven by a selection effect of willing bicycle commuters under particular weather conditions. For example, under very cold and adverse conditions, there could be two driving effects. For one, the adverse weather conditions could intimidate lower ability riders away from commuting by bicycle, lowering the average commute time by route among the remaining willing riders. Adding to this,

high skilled riders may be motivated by the cold conditions to ride even faster than they would otherwise. The other effect of course, is that the adverse conditions could actually slow down those participating riders in the data set. Although we are unable to distinguish the skill level of riders in the current data, the aggregated results seem to suggest that the weather adversity effect may slightly dominate among the participating riders.

Table 2: Route Fixed Effect Models: Dependent Variable: Trip Duration

	baseline	**market open**	**basic**
start time	0.0000***	0.0000***	0.0000***
	0.0000	*0.000*	*0.0008*
start time^2	0.0000***	0.0000***	
	0.0000	*0.000*	
end time before 30, 00	1.4515		1.4083
	0.6806		*0.6897*
end time after 30, 00	-6.2911		-6.2334
	0.0803		*0.0832*
end time before 15, 45	-0.0760		-0.1009
	0.9832		*0.9777*
end time after 15, 45	-0.9814		-0.8972
	0.7812		*0.7996*
end time before 9:30am		25.7256**	
		0.0070	
end time after 9:30am		16.3213	
		0.0886	
arrive at 9 something	66.5549***		65.9238***
	0.0000		*0.0000*
arrive at 8 something	36.7328***		36.4335***
	0.0000		*0.0000*
arrive at 6 something	-25.4239***		-25.9271***
	0.0000		*0.0000*
birth year	-173.3669***	-166.8534***	-2.6832***
	0.0000	*0.0000*	*0.000*
birth year ^2	0.0432***	0.0416***	
	0.0000	*0.0000*	
gender (female = 1)	75.8021***	78.1895***	76.5742***
	0.0000	*0.0000*	*0.0000*
Subscriber	-360.3943***	-369.9227***	-360.9093***
	0.0000	*0.0000*	*0.0000*
temperature low	-1.2623**	-1.3273**	0.2581***
	0.0080	*0.0054*	*0.0009*
average temperature	0.8508	0.7903	
	0.0624	*0.0841*	
humidity high	0.0182	0.0377	
	0.8300	*0.6582*	
wind high	-0.2351	-0.2601	-0.4813*

	0.3438	*0.2960*	*0.0459*
average rainfall	-0.0266	0.0025	0.0050
	0.9080	*0.9915*	*0.9824*
avg temp*rainfall	0.0004	-0.0002	-0.0003
	0.9207	*0.9514*	*0.9390*
R squared	0.686	0.684	0.685
Obs	111,775	111,775	111,775

p-values in italics; *p<0.05, ** p< 0.01, ***p< 0.001

While so far, the analysis completely allows for unobserved heterogeneity in route characteristics, we also consider the possibility that commuters are not so familiar with the features of specific routes, which is highly plausible in reality. Suppose that a Wall Street worker is commuting this morning from a new starting station. How might he or she assess the needed cycling time?

In order to understand how accounting for route heterogeneity makes a difference in the model, we remove the route fixed effects from the analysis and replace them with simply the geographic distance between the starting and ending dock station. Since this distance-based approach is a more heuristic approach than the models estimated previously, we use the basic model previously considered as a baseline, thus, only those variables in that model are considered here.

$$tripduration_{j(,i,t)} = \alpha \cdot dist_j + \beta \cdot timefeatures_{j(,t)} + \gamma \cdot demogr_{j(,i)} + \delta \cdot weather_j + \varepsilon_j$$

Table 3 shows the results from this specification, adding each of the groupings of variables as a possible specification, as a decision-maker with a heuristic approach might do. For example, some commuters might mostly focus on their personal characteristics in assessing their commute time. Others may additionally account for time factors, but not weather, and so on.

As expected, the distance variable is highly significant throughout, and indicates about a 7 minute trip time increase per mile on average. Age remains significant in the linear term, as is gender. In fact, without accounting for route-effects, the gender difference in travel time appears larger than is actually the case. Turning to the weather column, increasing daily low temperature significantly increases commute time, which is consistent with the basic model in Table 2. Finally, for those models with time variables included, the 8'oclock and 9'oclock range remain significant explanatory variables for lengthier commute times. For those distance-based specifications with weather variables included, we observe a modest temperature effect in that as the daily low temperature increases, the trip duration shortens.

Table 3: Distance-based Models: Dependent Variable: Trip Duration

	distance	**personal**	**weather**	**timing**	**full**
ride distance	418.2942***	417.8449***	417.7254***	417.6662***	417.5241***
	0.0000	*0.0000*	*0.0000*	*0.0000*	*0.0000*
birth year		-2.0338***	-2.0449***	-2.1161***	-2.1325***
		0.0000	*0.0000*	*0.0000*	*0.0000*
gender		95.0167***	94.8100***	90.1255***	89.8923***
		0.0000	*0.0000*	*0.0000*	*0.0000*

subscriber		-353.7219***	-353.7326***	-347.3475***	-346.0748***
		0.0000	*0.0000*	*0.0000*	*0.0000*
temperature low			0.2600***		-0.2738***
			0.0005		*0.0008*
wind high			-0.5629*		-0.5565*
			0.0300		*0.0315*
average rainfall			0.1125		-0.0803
			0.6372		*0.7417*
avg temp*rainfall			-0.0030		0.0020
			0.5084		*0.6810*
start time				0.0000***	0.0000***
				0.0000	*0.0002*
end time before 30, 00				1.6921	1.7649
				0.6535	*0.6396*
end time after 30, 00				-4.8252	-4.8436
				0.2082	*0.2064*
end time before 15, 45				3.7513	3.5964
				0.3310	*0.3513*
end time after 15, 45				-1.8998	-2.0113
				0.6145	*0.5939*
arrive at 9 something				55.3407***	56.1080***
				0.0000	*0.0000*
arrive at 8 something				32.8802***	33.1959***
				0.0000	*0.0000*
arrive at 6 something				-25.5361***	-25.7391***
				0.0000	*0.0000*
constant	165.4427***	4510.814***	4527.089***	5908.535***	5600.617***
	0.0000	*0.0000*	*0.0000*	*0.0000*	*0.0000*
R squared	0.562	0.631	0.631	0.633	0.633
Obs	114,225	111,775	111,775	111,775	111,775

p-values in italics; *p<0.05, ** p< 0.01, ***p< 0.001

Finally, we consider a duration analysis which allows for the possibility that the underlying duration or hazard probability depends on the sustained length of the trip duration. Duration analysis is a common modeling approach in transportation studies (Hensher and Mannering, 1994). Nevertheless, our previous distance-based and fixed effects approaches account for distance, and therefore may be less uniformly subject to trip lengths which are highly influenced by the elapsed trip time. As a robustness check, we implement a duration model with Weibull and Exponential baseline hazard distributions, respectively, while omitting the distance related variables.[5] The results, provided in Table 4, show that the same variables as in the previous models are the robust predictors of commute time, and in the same directions, where here a hazard ratio less than 1 indicates a slower end to the commute duration.

[5] Squared start time is omitted from the regression due to collinearity.

We note that the Weibull distribution model obtains some findings which were not significant in previous models, namely those arriving within three minutes after a round number time (15,30,45,00) had shorter commutes. In addition, the wind and humidity weather variables are significant. However, these effects are not robust to alternative underlying hazard distributions, therefore we may view these as less robust results.

Overall, the duration model provides some assurance that the directions, significance and relative magnitudes of the effects in the linear model are robust.

Table 4: Duration Model: Dependent Variable: Trip Duration

Hazard ratio shown, p-value in italics

	Weibull	**Exponential**
start time	1.000***	1.000
	0.001	*0.352*
end time before 30, 00	1.005	0.995
	0.615	*0.627*
end time after 30, 00	1.028**	1.008
	0.007	*0.422*
end time before 15, 45	1.000	0.992
	0.999	*0.420*
end time after 15, 45	1.028**	1.011
	0.008	*0.299*
arrive at 9 something	0.882***	0.936***
	0.000	*0.000*
arrive at 8 something	0.923***	0.948***
	0.000	*0.000*
arrive at 6 something	1.152***	1.109***
	0.000	*0.000*
birth year	0.570***	0.732***
	0.000	*0.000*
birth year ^2	1.000***	1.000***
	0.000	*0.000*
gender (female = 1)	0.798***	0.841***
	0.000	*0.000*
subscriber	1.779***	1.498***
	0.000	*0.000*
temperature low	1.001	1.001
	0.652	*0.273*
average temperature	0.998	0.997*
	0.063	*0.013*
humidity high	0.999	1.000
	0.085	*0.724*
wind high	1.004***	1.001
	0.000	*0.051*
average rainfall	1.001	1.000
	0.150	*0.618*

avg temp*rainfall	1.000	1.000
	0.053	*0.335*
Log Rho	0.393***	
	0.000	
Obs	111,775	111,775

p-values in italics; *p<0.05, ** p< 0.01, ***p< 0.001

4. Models as a Commute Prediction Tool

We now consider the possibility of using the previous regression analyses as prediction tools for commute times. The notion of regression as a candidate decision tool is discussed in Gigerenzer, Hertwig and Pachur (2011), which argues that simple heuristics often outperform more complex analyses including regression.

Our main question of interest is about the distributions of late and early arrival times using the regression models as the prediction. This helps us to obtain an understanding of whether our models capture a balanced assessment of commute times, or whether in practice, the bicycle commutes are substantially overtime or under-time relative to the prediction. Viewing it in another way, if commuters were to use these models to generate a prediction of their commute time, for example through an app or other method (Emberger and Shibayama, 2018 provide a summary of relevant technologies), how does the distribution of actual commute times compare?

To address this question, for each trip observation j, we define $overtime_j = tripduration_j - fittedvals_j$, where fitted values simply apply the estimated coefficients to the observable characteristics of trip observation j. In other words, we gauge overtime and under-time based on the residuals of the estimation. We focus on the basic model with route fixed-effects, and the full distance-based model, each in the rightmost columns of Tables 2 and 3, respectively. We consider two dimensions along which the overtimes may vary, namely by hour of arrival, and by distance traveled.

Figures 3 and 4 show the distributions of overtimes by each hour block of arrival times to the destination station. Figure 3 shows the distributions for the route fixed-effects model, while Figure 4 shows the distributions for the distance-based model. A normal distribution is fitted to each histogram, using the empirical mean and standard deviation of the data in the histogram (displayed in each of the Figure descriptions), and can thus be interpreted as a best fit under the assumption of a normal distribution. Cumulative distribution plots are also provided in the bottom panel of each Figure to illustrate the comparison of each subgroup by arrival hour. Out of practical considerations for graphing purposes, we restrict attention to those overtimes and undertimes within 20 minutes of the predicted commute time, which is denoted by the red dashed line.

We first notice that each of the distributions regardless of hour display excess kurtosis relative to a normal distribution.[6] This pattern is persistent across all of our arrival time and distance traveled subcategories, and might be at least partially attributed to heterogeneity in rider characteristics which are not observable in this data set, including riders' cycling skills, experience level and commitment. In addition, there is a time trend in the distribution, with later hours in the morning generally skewing

[6] Recall that one measure of a model's proper specification is that the errors follow a normal distribution. While our models generally do not satisfy this assumption, we nevertheless examine the distribution of errors to understand the sources of this misspecification.

increasingly to arriving earlier than the model predicts. In other words, while a worker arriving around 6am or 7am is almost about as likely to arrive earlier than anticipated compared to later than anticipated, there are substantial proportions of commuters in the 9am block that beat the prediction, as seen by the peaks in the distributions to the left of the red dotted line. These distributions are successively statistically different from one another via Kolmogorov-Smirnov tests ($p < 0.001$).

Figure 3: Overtimes: Fitted Values by hour of arrival, Route Fixed Effects Model

Top panel: empirical density by hour of arrival, fitted with normal distributions based on actual (mean, stddev): 6: (-0.065, 2.012); 7: (-0.056, 2.101); 8: (-0.069, 2.383); 9: (-0.172, 2.895)

Bottom panel: empirical cumulative densities by hour of arrival (statistically different from one another via Kolmogorov-Smirnov tests, p<0.001)

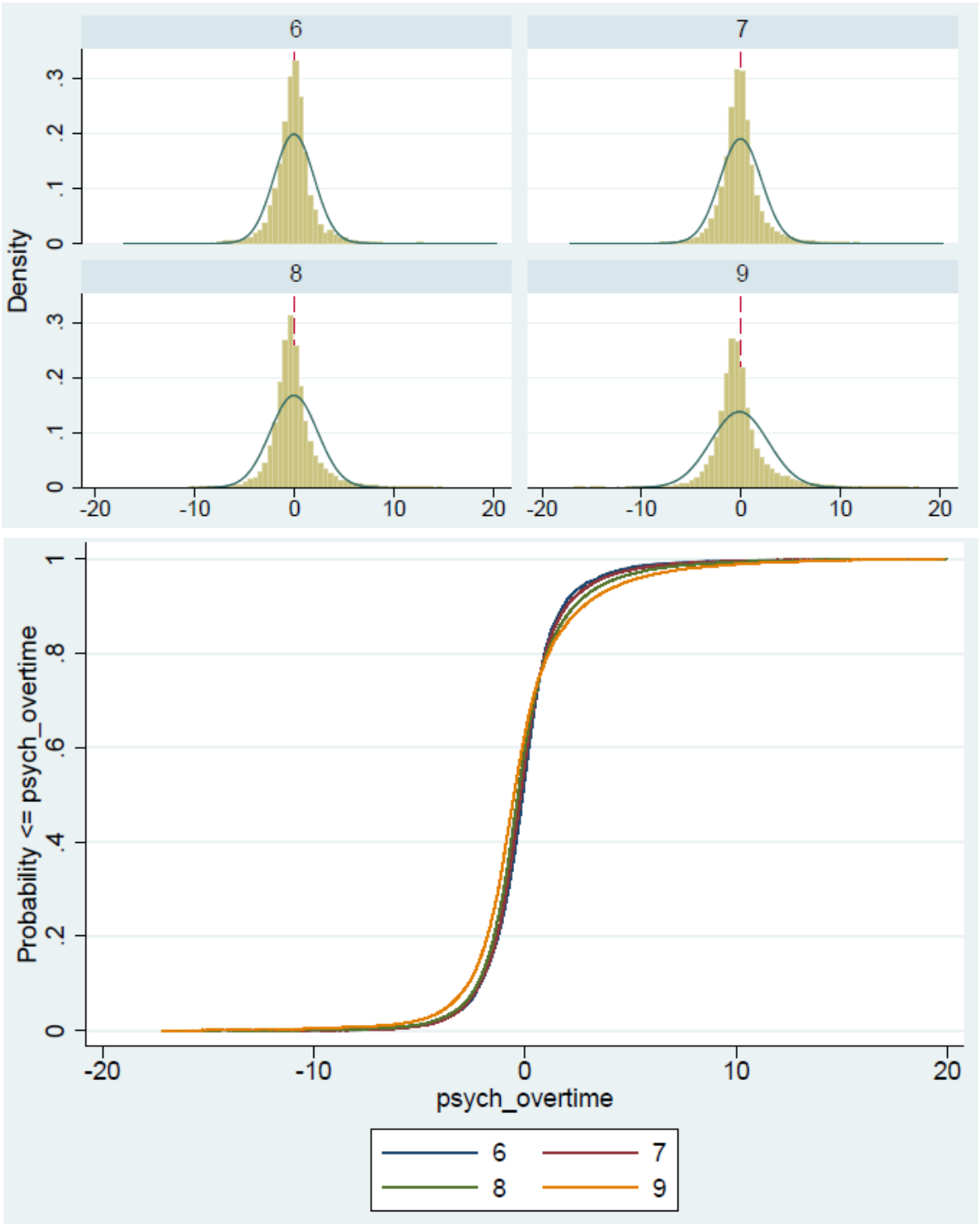
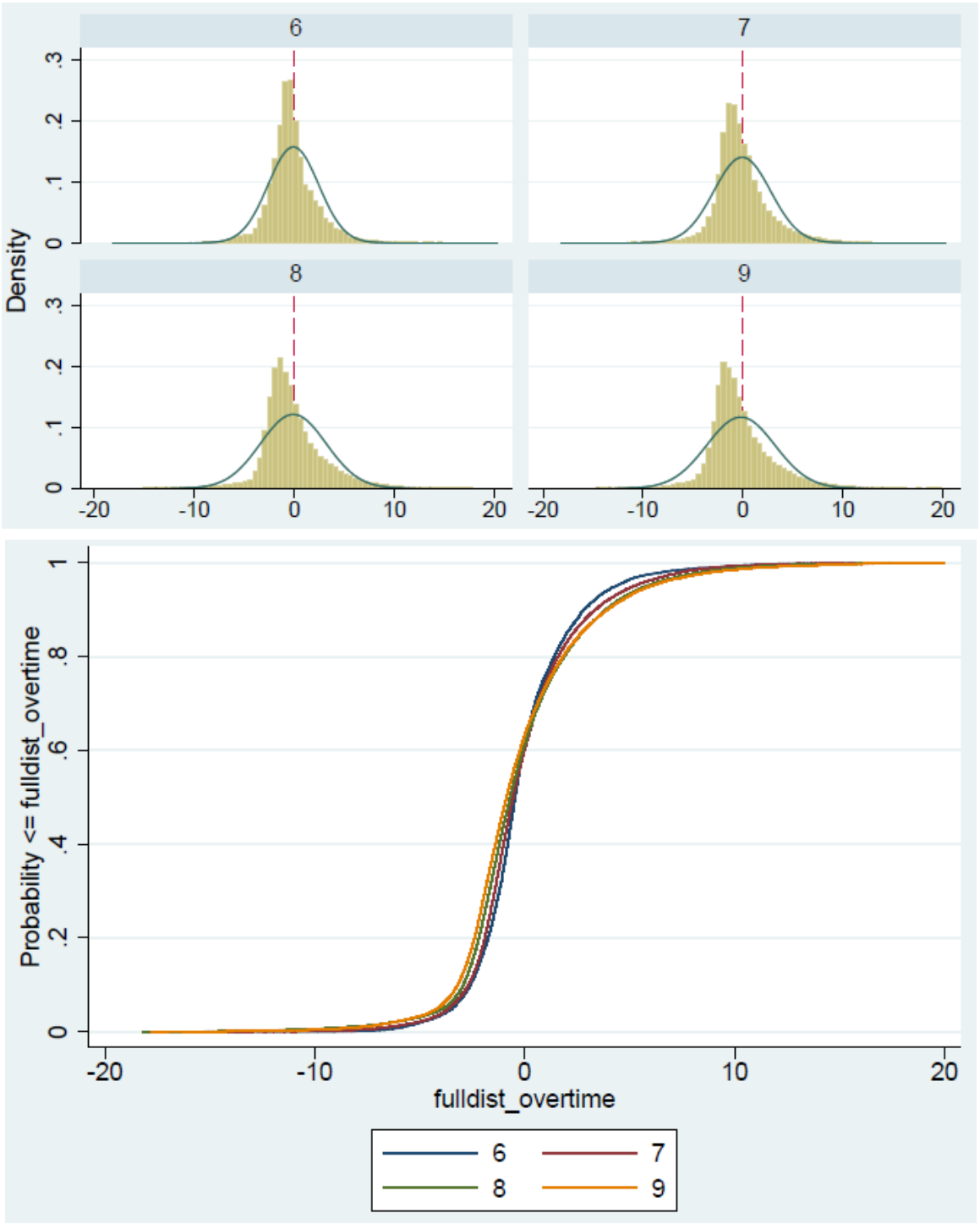


Figure 4: Overtimes: Fitted Values by hour of arrival, Distance-based model

Top panel: empirical density by hour of arrival, fitted with normal distributions based on actual (mean, stddev): 6: (-0.100, 2.529); 7: (-0.065, 2.833); 8: (-0.111, 3.283); 9: (-0.206, 3.407)
Bottom panel: empirical cumulative densities by hour of arrival (statistically different from one another via Kolmogorov-Smirnov tests, p<0.001)



The pattern is consistent with the idea that for those bike commuters who need to get to work more urgently, as those arriving later in the morning likely do, adjustments can be made to their riding strategy to enable an earlier-than-expected commute time. When comparing Figure 3 with Figure 4, this time effect of the arrival distribution is even more pronounced for the distance-based

model than the route fixed-effects model. This indicates that accounting for route fixed-effects in the model allows relatively less room for commuter maneuvering when compared to the distance-based model. In other words, some of the strategies bicycle commuters used to get to work in the desired time frame could be route-specific.

We next examine the overtime distributions by distance commuted. We divide the trips in the data into quartiles based on the distance traveled. The bottom 25% of the trips traveled 0.526 miles or less, the next 25% of trips traveled between 0.526 miles and 1.113 miles, while the next 25% of trips traveled between 1.113 miles and 2.015 miles, and finally the top 25% traveled over 2.015 miles. We use the same 20-minute undertime and undertime interval as previously for the distribution plots. The distributions are shown in Figures 5 and 6.

From the bottom panels of Figures 5 and 6 which show the empirical cumulative distributions of overtimes, we can see clearly that the distributions by distance quartiles are noticeably even more different from one another than the distributions by hour of arrival depicted in Figures 3 and 4. As in Figures 3 and 4, the distributions by distance quartile are successively statistically different from one another via Kolmogorov-Smirnov tests ($p < 0.001$).

In addition, similarly to the distributions of overtimes by hour, there is generally excess kurtosis in the distribution of residuals. However, in the top panel of Figure 5 for the fixed-effects model, we can notice a distinct pattern, which is that as the distance of travel is longer, the distribution of arrival times becomes noticeably closer to the fit of a normal distribution. In the distance-based model (Figure 6), the effect is further pronounced, with an even clearer additional pattern that all distance quartiles display substantial early arrival times at the peak of their distributions.

The increasing normality of the overtimes as a function of distance indicates that the models seem to be more accurately specified for long distance commutes than shorter distance commutes, and one apparent reason is the ability of commuters to outperform the model for the shorter rides in the data. This may be intuitive in the sense that for longer rides, the majority of time on the ride is likely spent actually covering geographical distance, whereas for the shorter rides, traffic conditions and road density in the commuting vicinity may present more opportunities for maneuvering and hastening arrival times.

Indeed, no specific variable in the empirical models can capture the effort and maneuvering flexibility of the bicycle commuters. While variables such as time, distance, demographics and weather characteristics are observable to the researcher, individual effort levels of the cyclists are not observable. We can only infer the ‘faster than expected’ rides through their placement in the distribution of arrival times, potentially attributing such deviations from normally distributed residuals to unobservable factors including riders’ efforts.

Figure 5: Overtimes: Fitted Values by distance quartiles, Route Fixed Effects Model

Top panel: empirical density by distance quartile, fitted with normal distributions based on actual (mean, stddev): 1st: (-0.183, 2.020); 2nd: (-0.068, 1.915); 3rd: (-0.047, 2.446); 4th: (-0.073, 3.177)
Bottom panel: empirical cumulative densities by distance quartile (statistically different from one another via Kolmogorov-Smirnov tests, p<0.001)

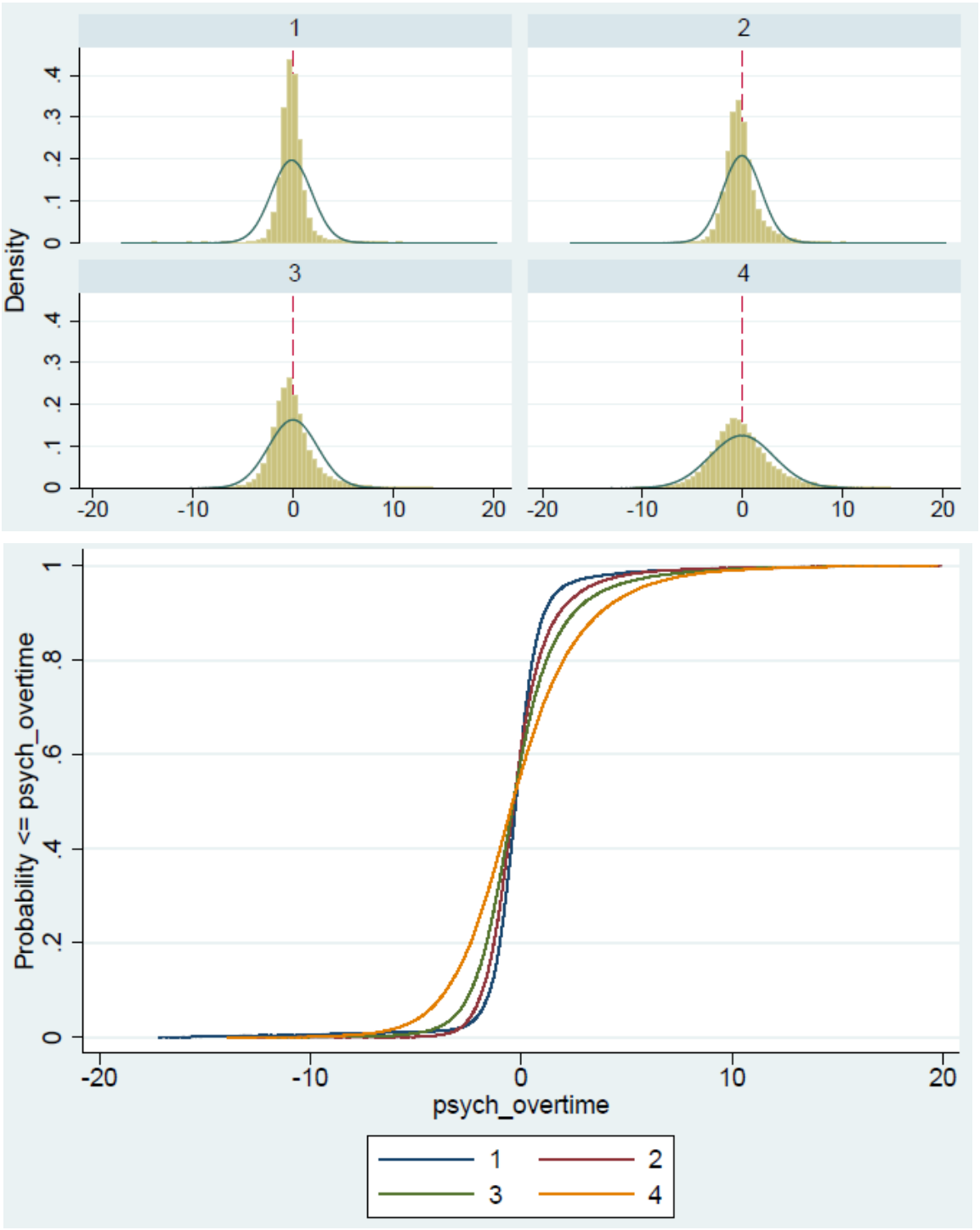
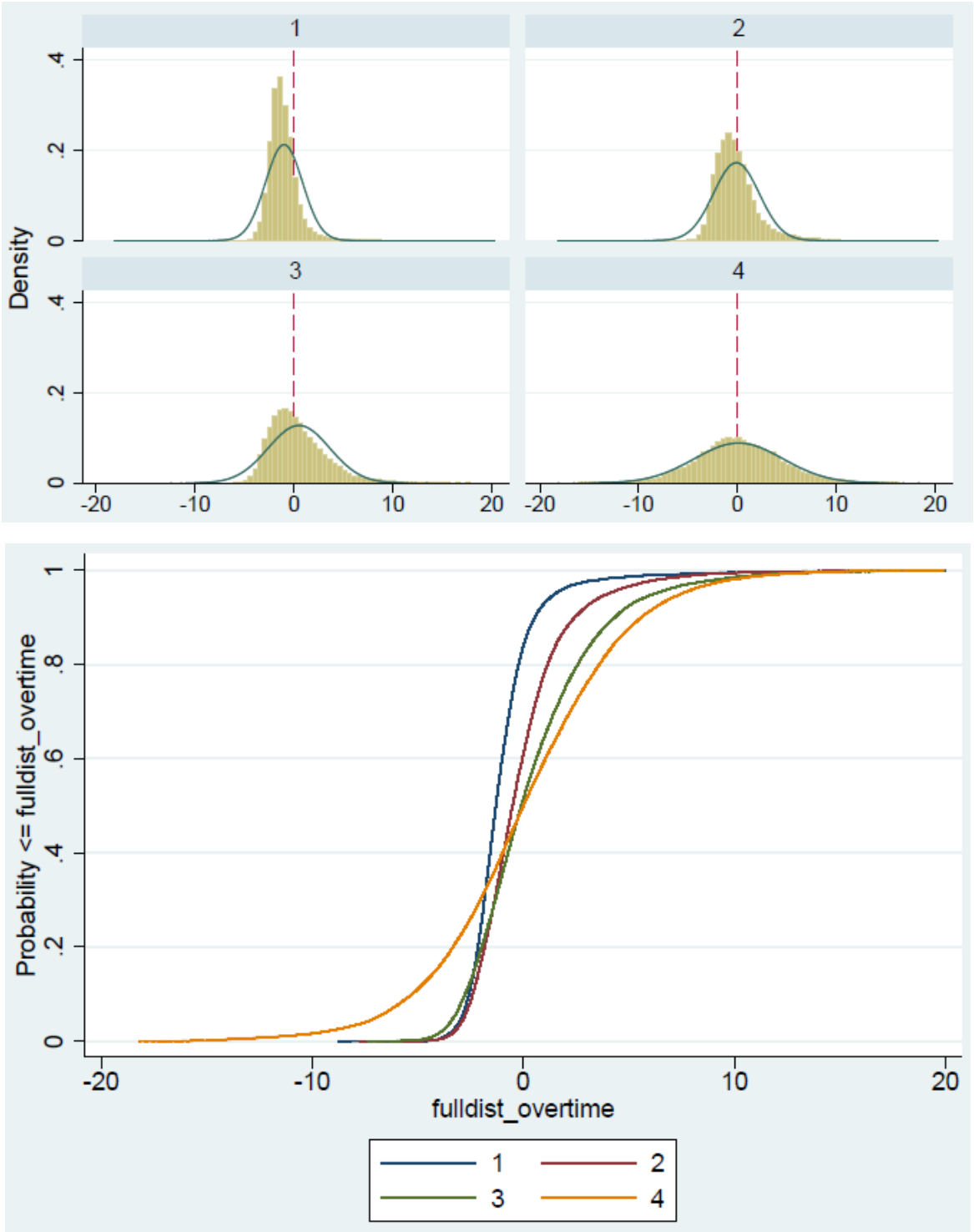


Figure 6: Overtimes: Fitted Values by distance quartiles, Distance Based Model

Top panel: empirical density by distance quartile, fitted with normal distributions based on actual (mean, stddev): 1st: (-1.009, 1.872); 2nd: (-0.113, 2.309); 3rd: (0.501, 3.121); 4th: (0.114, 4.496)
Bottom panel: empirical cumulative densities by distance quartile (statistically different from one another via Kolmogorov-Smirnov tests, p<0.001)



5. Conclusions

In this study we have examined the commuting patterns and ride times for a subpopulation of New York City bicycle commuters, whose arrival time at work is likely to be more uniformly time-sensitive than other subsets of the working population. This allows us to assess the effectiveness and travel time influences of the NYC Citibike program from a commuting standpoint, compared to the broader set of bicycle trip types previously considered in the literature.

We consider different types of econometric models and assess their implications for "late" and "early" commute arrivals with respect to the model predictions. In our first set of models, route-specific characteristics which may be unobservable to individual commuters are fully accounted for in a route fixed-effects framework. Controlling for these unobservable features, the results show that in the most comprehensive model, the main significant factors in commute time are demographics and hour of arrival. Weather factors appear to play a relatively smaller role in predicting the commute time, but may instead be influential on the decision to cycle itself, as suggested in other studies (Martinez, 2017). The potential selection of ridership population for commuting purposes is an important topic for future research.

In the route fixed-effect approach, our minor variations on the model specification, such as the psychologically plausible one (basic), or one which focuses on opening bell timing (market open), yield only minimal differences in terms of the estimated late arrivals and early arrivals in the trip sample. We emphasize that our assessment of a trip observation being "overtime" or "early" in the analysis operates under the assumption that a commuter forms an expectation on the commute time in a similar manner that we run our regression analysis. One possible mechanism for a commuter to obtain such a prediction which is generated using a large volume of data is through a program or an app. Commuters may also in practice form expectations heuristically based on their own prior experiences and the experiences of their acquaintances. The route fixed-effects approach averages the effects of the independent variables across all routes, where individual commuters are unlikely to be so knowledgeable about all of the frequently used routes in the system. An alternative prediction approach would be to run individual estimations for each of the frequent routes in the data – however this approach has the drawback of being less comprehensive and generalizable in the findings, and further may suffer from a relatively small sample for those relatively infrequently traveled routes. One interpretation of the prediction obtained by the fixed effects model is that it is an average of general predictions generated by commuters who are individually familiar with their own commute route, perhaps similar to some crowdsourced traffic apps.

Although the Citibike data is limited in its ability to provide commuter-specific behavioral insights due to lack of a user ID variable, we approximate slightly more plausible statistical decision models using a distance-based estimation rather than a route fixed-effects one. The idea is that a worker who commutes by bike on an infrequent basis, which is plausible given many of the commute volumes in the data, may not have a complete picture of the Citibike system and all the factors that influence commute times within it. Some commuters may focus on their own abilities and characteristics in assessing the commute time, while others may be more cognizant of weather and/or timing factors. We find that regardless of the exact specification, those main significant variables from the fixed-effects approach are still robustly significant, while of course the distance variable is highly statistically and economically significant. Robustness checks using duration models also yielded the same variables being significantly and robustly influential on commute time.

Some intuitive results are found when evaluating the overtimes or undertimes compared to the

main models, which suggest that bicycle commuters successfully allocate their effort into arriving at work at a particular time. Subdividing the residuals by hour of arrival, those individuals arriving later in the morning are significantly more likely to outperform the model, which is not by coincidence consistent with workers' intensity of incentives for having a timely commute. In addition, the distance of the commute also predicts how well the regression model performs. While the overtimes for the long-distance rides are relatively closer to a normal distribution, the shorter distance rides tend to be far from normally distributed and skew towards early arrival. This is again suggestive of riders' abilities to maneuver and find faster routes for short distance commutes, enabling them to arrive to work on time.

We view several directions for future research. Firstly, our approach may not realistically reflect how actual commuters may make their assessment of bicycle commute durations. A more behavioral approach may be to incorporate further limitations in commuters' assessments, including for example, adaptive expectations or expectations formed upon immediately previous or best/worst case scenario examples. Secondly, our current analysis focuses only on the data provided in the Citibike system and does not incorporate traffic flows on alternative modes of transportation (for example, in the nature of Faghih-Imani et al, 2017)). While bicycles are well-known for being able to weave in and out of traffic in a way that cars cannot easily achieve, it is quite plausible that automobile traffic conditions impact the effective bike speeds and affect the assessment and realization of bike commute times. For this approach, merging of the vehicular traffic data with the current bicycle system data would be needed. Finally, the current data are limited in being able to distinguish between effects on the extensive (ie. rider selection effects) and intensive margins. Meaning that with only trip-level data available and no unique rider identifier available, our results are due to a combination of changes in commuter participation in Citibike for a particular morning, and actual impacts on the commute time for commuters who decided to use Citibike. One important direction for future work is to disentangle these two types of effects, possibly through additional data collection, survey methods, or advanced simulation methods. We leave these directions for future research.

References:

Bergström, A. and Magnussen, R. 2003. Potential of transferring car trips to bicycle during winter. *Transportation Research Part A*, 37: 649–666

Bernhoft, I. M. and Carstensen, G. 2008. Preferences and behaviour of pedestrians and cyclists by age and gender. *Transportation Research Part F*, 11(2): 83–95.

Brandenburg, C., Matzarakis, A. and Arnberger, A. 2004. "The effects of weather on frequencies of use by commuting and recreation bicyclists". In *Advances in Tourism Climatology*, Edited by: Matzarakis, A., De Freitas, C. R. and Scott, D. Vol. 12, 189–197. Freiburg: Berichte des Meteorologischen Instituts der Universität Freiburg.

Brandenburg, C., Matzarakis, A. and Arnberger, A. "Weather and cycling – a first approach to the effects of weather conditions on cycling" *Meteorological Applications*, Vol. 14 (2007), Issue 1, p. 61 – 67.

Cervero, R. and Duncan, M. 2003. Walking, bicycling, and urban landscapes: evidence from the San Francisco Bay Area. *American Journal of Public Health*, 93(9): 1478–1483.

de Geus, B. 2007. "Cycling to work: psychosocial and environmental factors associated with cycling and the effect of cycling on fitness and health indexes in an untrained working population". Department of Human Physiology and Sports Medicine, Vrije Universiteit Brussel. Doctoral dissertation

Dickinson, J. E., Kingham, S., Copsey, S. and Hougie, D. J. P. 2003. Employer travel plans, cycling and gender: will travel plan measures improve the outlook for cycling to work in the UK?. *Transportation Research Part D*, 8(1): 53–67.

Emberger, Guenter and Takeru Shibayama, "Emerging types of mobility services and vehicle technologies, employed ICTS and implication for transport planning and policy", working paper 2018.

Faghih-Imani, Ahmadreza, Sabreena Anowar, Eric J. Miller, and Naveen Eluru, "Hail a Cab or Ride a Bike? A Travel Time Comparison of Taxi and Bicycle-Sharing Systems in New York City", *Transportation Research Part A, Policy and Practice*, July 2017.

Gatersleben, B. and Appleton, K. M. 2007. Contemplating cycling to work: attitudes and perceptions in different stages of change. *Transportation Research Part A*, 41(4): 302–312.

Gigerenzer, Gerd, Ralph Hertwig and Thorsten Pachur, Heuristics: The Foundations of Adaptive Behavior, Oxford University Press, 2011.

Hensher, David A., and Fred L. Mannering, "Hazard-based duration models and their application to transport analysis", *Transport Reviews*, Vol. 14 (1994), No. 1, p. 63 – 82.

Leth, Ulrich, Takeru Shibayama and Tadej Brezina, “Competition or Supplement? Tracing the Relationship of Public Transport and Bike-Sharing in Vienna”, *GI Forum*, Issue 2, p. 137 – 151.

Liang, Xuedong, Guangsen Si, Leilei Jiao and Zhi Li, “Recycling scheduling of urban damaged shared bicycles based on improved genetic algorithm” *Journal of Logistics Research and Applications*, Feb 26, 2018.

Martens, K. 2004. The bicycle as a feedering mode: experiences from three European countries. *Transportation Research Part D*, 9: 281–294.

Martinez, Mark (2017) "The Impact Weather Has on NYC Citi Bike Share Company Activity," *Journal of Environmental and Resource Economics at Colby*: Vol. 4 : Iss. 1 , Article 12.

Moss, Mitchel L., Caron Y Qing, and Sarah Kaufman, “Commuting to Manhattan: A study of residence location trends for Manhattan workers from 2012 to 2019”, Working Paper (March 2012), New York University Wagner School of Public Service.

Noland, R. B. and Kunreuther, H. 1995. Short-run and long-run policies for increasing bicycle transportation for daily commuter trips. *Transport Policy*, 2(1): 67–79.

Simonsohn, Uri (2006) "New-Yorkers Commute More Everywhere: Contrast Effects in the Field," *Review of Economics and Statistics*, V88 (1) 1-9.

Wang, Yacan, Dick Ettema, Huiyu Zhou and Xiangrui Sun, “Understanding peak avoidance commuting by subway: an empirical study in Beijing”, *Journal of Logistics Research and Applications*, May 7, 2018.

Wheat, Phill; Smith, Andrew S.J., “Assessing the marginal infrastructure maintenance wear and tear costs for Britain's railway network”, *Journal of Transport Economics and Policy*, Vol. 42 (2008), p. 189-224.

Appendix: Tables and Charts

Table A1: Summary Statistics for Citibike financial district destinations (N = 114,225)

Variable	*Mean*	*Std Dev*	*Min, Max*
Trip duration (sec)	771.65	638.79	62, 89662
Min temperature (F)	55.15	15.48	14, 82
Rain average	7.51	21.87	0, 210.01
Birth year (self-report)	1976.66	11.33	1937, 1999
Gender (female)	0.221	0.415	0,1
Frequency of route	246.26	253.76	61, 1282
Ride distance (miles)	1.449	1.145	0, 7.04

Figure A1: Arrival Times for all Ending Stations