ESTIMATING THE CAUSAL EFFECTS OF RIDE-HAILING TECHNOLOGIES ON TRAFFIC CONGESTION

A PREPRINT

Ryan M. Sander
Department of Economics
Massachusetts Institute of Technology
Cambridge, MA 02139
rmsander@mit.edu

December 30, 2019

ABSTRACT

Identifying and quantifying the causal effects of ride-hailing technologies on urban traffic congestion delays, costs, and emissions is confounded not only by the presence of omitted variables bias, but also econometric endogeneity. This paper aims to mitigate these econometric issues through the use of spatial and temporal fixed effects, controls, and instrumental variables in panel regression models. By mitigating the presence of endogeneity and omitted variables bias, we are better positioned to provide policy-makers, researchers, industry leaders, and the general public alike with less biased estimates of the causal effects these ride-hailing technologies have on urban traffic congestion. We find that introducing controls, fixed effects, and in particular an instrumental variable for age into our regression models decrease the effects of ride-hailing technologies on traffic congestion and congestion-related emissions. Additionally, we find a statistically-significant negative coefficient capturing the effect of Uber on congestion costs per auto-commuter, indicating that ride-hailing technologies may in fact lead to decreased traffic congestion.

***K*eywords** ride-hailing · transportation · traffic congestion

1 Introduction

1.1 Policy Motivation

As ride-hailing technologies continue to reshape the on-demand driving landscape for increasingly many urban and rural regions across the globe, it is imperative to ascertain, to the greatest extent possible, estimates for the causal effects of these technologies on traffic congestion. This research analyzes the effects of ride-hailing on time delays resulting from traffic congestion, and their associated costs, as well as excess fuel consumption produced by traffic congestion. Developing a more rigorous and causal understanding of this technology's effects will position policymakers, business leaders, and the general public to make more informed decisions about the future of these ride-hailing technologies.

Obtaining accurate estimates of the effects of these ride-hailing technologies is of particularly pressing importance today, as these technologies are ramping up their penetration and operations across the globe, as can be seen in Figure 1. Enhancing our understanding of the effects of ride-hailing technologies positions these technologies to be introduced and operationally managed in a more intelligent way.

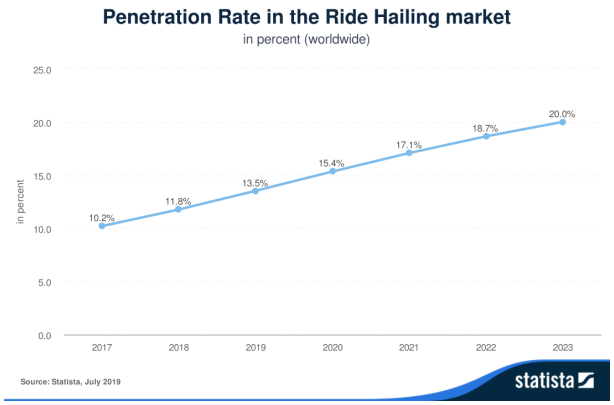

Figure 1: Current and projected worldwide ride-hailing penetration rates.

1.2 Econometric Motivation

Central to the goals of better understanding the effects of ride-hailing on traffic congestion and congestion-related emissions are finding methods that mitigate econometric issues associated with this specific analysis. In this study, omitted variables bias and endogeneity via reverse causality are of particularly strong concern. We explain our reasoning for these concerns below.

Omitted variables bias is largely an econometric concern because traffic congestion and congestion-related emissions are highly coupled to a multitude of macro-scale demographic variables, resulting in confounding effects and complex system dynamics. For example, traffic congestion is strongly tied to productivity: worsening traffic congestion results in longer commute times for auto-commuters, reducing their productivity in their personal and/or occupational lives. In turn, productivity influences traffic congestion: if less people need to commute to travel to work (a manifestation of lower productivity), traffic congestion will increase. It is these two-way relationships between traffic congestion and a myriad of other variables that necessitate the need to mitigate omitted variables bias in our econometric models. In addition to this concern, we believe that self-selection of ride-hailing technologies into certain cities introduces another set of econometric issues: endogeneity.

Endogeneity, the latter concern above, is mentioned because there is likely a degree of reverse causality associated with estimating the effects of ride-hailing on traffic congestion. Under the assumption that Uber, Lyft, and other ride-hailing technology firms seek to maximize their profits, they will choose to enter into different cities in the order in which they can return the largest margins. One factor that may influence a ride-hailing company's expected margin (and therefore their decision of whether or not to enter a city) is the level of traffic already present in that city. All things equal, higher rates of traffic indicate higher demand for mobility, and therefore a larger market to penetrate. Furthermore, larger cities, with consequently larger market demand compared to smaller cities, tend to have worse traffic congestion [9]. This analysis shows that ride-hailing technology companies likely self-select into cities that are already experiencing severe traffic congestion, and separating the causal effects of these ride-hailing technologies on traffic congestion from the reverse causal effect will provide us with more accurate estimates of the true effects of these technologies.

A final motivation stems from the connection ride-hailing technologies have to a similar emerging technology: autonomous vehicles. Though we do not quantitatively investigate the effects of self-driving on traffic congestion and emissions in this work, the conclusions we draw here may serve as a good litmus test for the potential effects self-driving cars may have on traffic congestion and traffic-related emissions. Namely, we believe this litmus test holds here because autonomous vehicles, through their autonomy, have the potential to deliver ride-hailing services once these vehicles reach levels of full autonomy. We arrive at this conclusion because self-driving cars, intuitively, are capable of picking up and dropping off passengers without the need to make an eventual "return trip" back to the passenger's pickup point from the drop-off point, a key feature of ride-hailing services.

2 Related Work

Other empirical works have sought to answer similar and near-identical questions of how ride-hailing technologies affect traffic congestion and congestion-related emissions. We leverage these related works' econometric techniques and insights, and also hope to make meaningful expansions of key insights and conclusions from these works.

2.1 Effects of Ride-hailing on Emissions, 2014

Our research aims to revisit the analysis carried out in Li, Hong, and Zhang [3], using a modified set of regression controls and a newer edition of the dataset utilized in this paper. Our research leverages fixed effect and instrumental variable models that are similar to the models utilized in this paper, and our final choice for our instrument was motivated by the success seen by leveraging the same instrument used in [2]. Using the Texas Transportation Institute's Urban Mobility Report with observations up to 2014 (our analysis utilizes observations up to 2017) in tandem with Uber and Lyft entry and U.S. Census data, Li, Hong, and Zhang find that ride-hailing technologies such as Uber have net negative effects on different measures of traffic congestion.

Revisiting this study promotes further insight into this econometric research question, as ride-hailing technologies and penetration expanded substantially between 2014 and 2017. Our results corroborate the initial results seen, though in some cases, with less statistical significance and certainty than this original paper. We also extend this study beyond just Uber to both Uber and Lyft, encapsulating a greater share of the ride-hailing industry than just Uber alone.

2.2 Further Instruments to Investigate

Due to the presence of econometric endogeneity, as discussed above, instrumental variables play a critical role in our analysis. The analysis carried out in Haldun Anil, Mark43, and Sara Fisher Ellison et al. [5]. analyze the effects of regulatory policies on Uber's entry decisions into urban and metropolitan areas. The insight into how we can both use additional instrumental variables to mitigate the endogeneity discussed above, as well as relate our instrumental variables to the effects of regulatory policies on ride-hailing data, enabling us to intuitively evaluate the quality of the instruments we use.

2.3 Expanding Upon Previous Studies

Other related works through which we frame our research are reports that both claim (a) ride-hailing technologies increase traffic congestion, and (b) do not make any specific mentions to instrumental variables, endogeneity, or omitted variables bias. Though we aim to show through our research that it is possible for ride-hailing technologies to decrease traffic congestion, these articles also provide substantial insight into some of the ways in which these technologies may increase traffic congestion. In turn, these conclusions that describe some of the phenomena that may well lead to increased traffic congestion can be passed on to policymakers for actionable traffic congestion mitigation efforts that fall outside the notion of ride-hailing.

Namely, Clewlow and Mishra [1] report that 49-61% of ride-hailing trips are trips that would have been made by walking, biking, transit, or avoided altogether, citing the potential for ride-hailing technologies to simply create more traffic than what existed before. Schaller's findings [4] also suggest that ride-hailing technologies may be creating more traffic in urban areas: This study finds that ride-hailing companies added 976 million miles of driving to New York City streets from 2013 to 2017. One potential argument that serves as a countervail to the arguments above is that ride-hailing technologies, while perhaps creating more riding trips than what would have resulted from the counterfactual outcome, exhibit a mobility phenomenon of "no back trip". A ride-hailing car will not have to make a subsequent return trip after dropping off a rider, and though another vehicle may need to make this return trip to transport the rider back to where they came from, it is likely (through path planning optimization) that this second vehicle was traveling in that direction to begin with. For instance, Uber has a feature in which drivers can enter a destination location, and their potential rides are filtered to only display rides which take the rider further in that direction[10]. Nonetheless, regardless of the outcomes of these studies, they will position policymakers to consider the effects of ride-hailing more holistically. With our research frame now set, we are ready to discuss our research question and dataset.

3 Empirical Economic Setting and Dataset

3.1 Empirical Environment

Research Goal: Formally, our research goal is to develop panel regression models for estimating the causal effects of ride-hailing entry on congestion cost per auto-commuter, a commuting time index, excess hours lost per auto-commuter, and excess emissions per auto-commuter. Our empirical economic setting, which is amenable to analysis with a variety of different variables, provided us with a statistically-powerful environment with which we could develop and test models in.

Empirical Economic Setting: We conduct our analysis using panel data over American cities from 1982-2017, and analyze both micro and macro-economic effects gathered during this time period. As mentioned above, the rich sources of data that comprise this economic setting enable us to control for many omitted effects.

Top-Level Dataset Overview: For this project, I combined panel mobility and transportation data with cross-sectional ride-hailing data, as well as panel unemployment rate and estimated panel population distribution data for our instrumental variables panel regression approach.

3.2 Base Datasets

We leveraged four open data sources for this project:

1. **Texas Transportation Institute's 2019 Urban Mobility Report (TTI UMR)** [2]: The TTI UMR panel dataset contains time series observations from 101 different urban centers across the United States, and has observations for traffic and demographic variables such as average price of gas, number of commuters, number of miles driven on freeway and arterial roads, or excess fuel wasted in traffic.
2. **Uber/Lyft Urban penetration dataset** [3]: This cross-sectional dataset captures the entry year for both Uber and Lyft into 92 of the 101 urban centers studied in the TTI UMR.
3. **Unemployment Data** [7]: This panel dataset manually collected from the Bureau of Labor Statistics captures the yearly unemployment rate in the 92 urban centers present in both the Uber/Lyft penetration dataset and the UMR TTI.
4. **Age Distribution Data** [8]: This cross-sectional dataset was manually collected via the U.S. Census, and provides the 5-year estimate of the percentage of individuals aged 65 or older for each of the 92 urban centers present in the three datasets above. To avoid perfect collinearity, we will estimate city-specific time series values for this instrument by incorporating national rates for the percentage of population aged 65 or older over the past 10 years. This estimation technique will be discussed below.

3.2.1 Table of Summary Statistics for Our Base Datasets

Variable	Mean	Standard Deviation	Min	Max
City Population (1000s)	1546.4	2394.9	75	19095
Number of Auto-commuters (1000s)	643.7	839.2	27	6003
Annual Aggregated distance on freeway (miles)	12626.2	18701.3	110	139275
Annual Aggregated distance on arterial roads (miles)	12983.5	17869.9	435	126010
General Value of Money ($/hr)	12.79	3.42	7.20	18.12
Commercial Value of Money ($/hr)	33.15	8.26	23.31	52.14
Travel Time Index	1.16	0.08	1	1.51
Annual Excess Fuel Consumed per Auto-commuter (gallons/year)	13.28	7.42	1	45
Annual Hours of Traffic Delay per Auto-commuter (hrs/year)	33.26	16.49	1	119
Unemployment Rate (%)	6.85	2.48	2.41	19.66
Percentage of Population 65 or Older (%)	10.90	2.29	4.71	26.4

Table 1: Summary statistics for our panel Texas Transportation Institute Urban Mobility Report (TTI UMR) dataset, unemployment data, and population distribution data.

3.2.2 Traffic Congestion Insights From the 2019 Urban Mobility Report

The most pertinent summary statistics above for the 2019 Urban Mobility Report dataset are plotted below, using an average across each cross-section at each time index (in this case, year). From these plots and the summary statistics

above, it is clear that: (1) traffic congestion, on average, has been worsening in a near-monotonic fashion since well before ride-hailing technologies were even introduced into these cities, and (2) our output regressands of interest (Figures 3, 4, and 5) are all strongly correlated with the number of auto-commuters in an urban area, one of our demographic control variables (Figure 2).

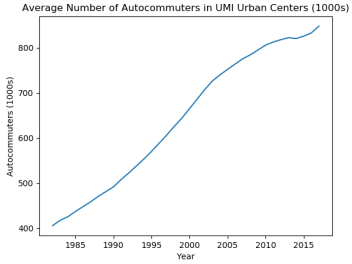

Figure 2: Number of auto-commuters in a city (1000s; averaged across urban centers) as a function of time. A quick analysis of the different graphs for traffic congestion also indicates a positive correlation with this variable.

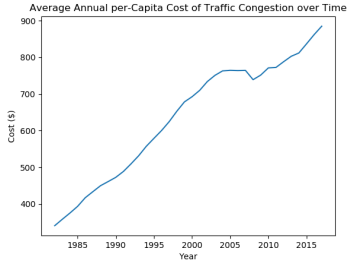

Figure 3: Annual congestion costs per auto-commuter (dollars/(auto-commuter $\times$ year); averaged across urban centers) as we vary time. This has been increasing since well before the introduction of ride-hailing technologies.

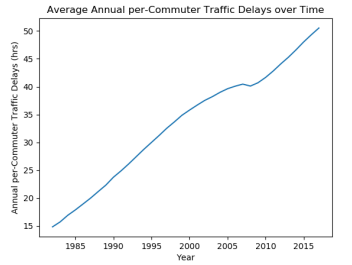

Figure 4: Traffic delay (hours/(auto-commuter $\times$ year); averaged across urban centers) as a function of time for auto-commuters.

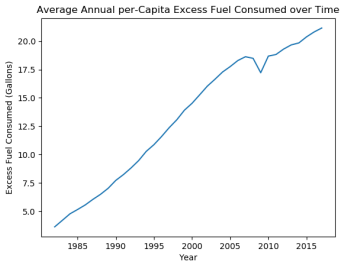

Figure 5: Excess emissions (gallons/(auto-commuter $\times$ year); averaged across urban centers) as a function of time.

3.2.3 Construction of Uber/Lyft Dataset, and Insights

This Uber/Lyft dataset is a panel dataset indexed by year and city. Each element of this dataset is an indicator random variable denoting whether the ride-hailing company in question is operating that city during that year:

$$u_{ct} = \begin{cases} 1 & \text{Uber operating in city c during year t} \\ 0 & \text{otherwise} \end{cases}, \quad l_{ct} = \begin{cases} 1 & \text{Lyft operating in city c during year t} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Below is a histogram depicting the entry years for Uber and Lyft for the urban centers contained within the Urban Mobility Report using these indicator variables:

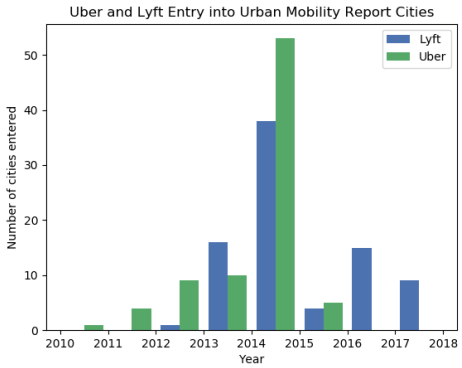

Figure 6: Uber and Lyft's paths to entering new urban areas across the urban areas used in the 2019 Urban Mobility Report.

3.2.4 Quantitative and Qualitative Details of Unemployment Dataset

As mentioned before, the unemployment data was collected for use as an instrumental variable in our empirical models, as this variable was in [3]. The data was taken from the U.S. Bureau of Labor Statistics [3], and each local unemployment panel data slice captures the monthly unemployment rate of the locality/urban area in question from the years 2007-2017. The average unemployment rate below is consistent with our intuition about the unemployment rate falling after its peak during the Great Recession.

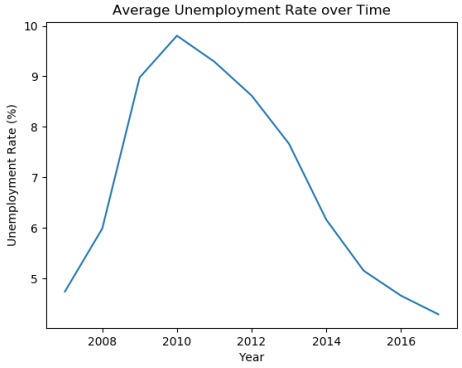

Figure 7: A time series graph of the unemployment rate averaged over all 93 cities included in the Urban Mobility Report and Uber/Lyft dataset.

3.2.5 Estimating Data for Our Age Instrument

As discussed above, we used an estimation technique to generate data for our second instrument, which captured the percentage of an urban area aged 65 or older. The motivation for this estimation was to avoid collinearity for our instrumental variables regression, which we observed empirically when we tried leveraging age data from solely 2017. Though we only had access to one year of data (2017) for each city for our age instrument, we were able to obtain national aggregated measurements of the percentage of population aged 65 or older over the past 10 years. We make use of these national aggregated measurements through the estimation procedure described below.

Under the assumption that the percentage of population aged 65 or older has been growing exponentially in all urban areas included in the TTI UMR for the past ten years, we first estimate exponential coefficients for this growth using national aggregates for this instrument.

Estimation Procedure: Using the national measurements for percentage of population aged 65 and older, assume these data points are generated according to the (noiseless) time series model given by:

$$y^{(t)} = ae^{bt} \tag{2}$$

We use the above model and our national measurement data for age in tandem with exponential regression to determine values for the coefficients a and b. Using the a and b coefficients obtained via this regression, we assume that each city has the same b coefficient as found above. To avoid collinearity, we add i.i.d. Gaussian noise of mean 0 and variance of 0.01 (1%) to each estimate, which (a) helps to mitigate the collinearity we observed without the addition of this noise, and (b) captures some uncertainty with each estimate, since the assumptions we make above likely do not hold in actuality. Therefore, each (city, time) estimate (with t = 0 corresponding to the year 2008) is given by:

$$z_c^{(t)} = a z_c^{(10)} \exp\{bt\} + g_c^{(t)} \tag{3}$$

Where $g_c^{(t)}$ corresponds to our i.i.d. Gaussian noise:

$$g_c^{(t)} \sim \mathbb{N}(0, 0.01) \tag{4}$$

Though we make many assumptions with these estimates, we believe these assumptions to be reasonable and consistent with our intuition, and account for some of this uncertainty through the addition of Gaussian noise. We chose the years 2008-2017 because the study in [3] chose to use their age and unemployment instruments over a 10-year period for their panel dataset. Furthermore, we believe years before this time period are simply less significant from an instrumental variables regression perspective simply because the endogenous regressors (our Uber and Lyft indicator random variables) are zero for all cities for all time periods before the year 2011.

With all of our datasets spelled out, we need to merge and combine these datasets in order to test our panel regression models.

3.3 Data Pre-Processing

To prepare these datasets for empirical analysis, some pre-processing steps were required:

1. **Merging the TTI UMR and Uber/Lyft datasets**: For this step, the entry date for Uber and Lyft (if applicable) was added as another observation for each different city. This was accomplished using a Python script that created 0-1 dummy variables for each (city, year) pair for new Uber and Lyft observation columns. A 0 indicates that the ride-hailing platform in question (in this case, Uber or Lyft) has not entered an urban area yet, while a 1 indicates that the ride-hailing platform has entered the urban area.
2. **Cleaning the combined panel data**: For this step, we manually removed city-year observation blocks that we were unable to find Uber and Lyft entry data for. After completing this manual filtering process, we still had 92 cities worth of data, and 35 observations for each city-specific time series.
3. **Adding unemployment rates and population distribution data to the combined dataset**: For this pre-processing step, we wrote a script that matched each unemployment rate key to its specified index in the panel dataset using year and city.

With our dataset curated and ready to test, we are ready to begin discussing our different empirical models tested in this paper.

4 Empirical Models

As previously discussed in the introduction, in this paper we investigate the use of empirical models that leverage control regressors, fixed effects variables, and instruments. These models, and the motivation for using them, are described in each of the sub-sections below. Before proceeding, I will define the different variables leveraged in these models.

Our table of our output dependent variables of interest is below:

Regressand Name	Definition
EFPA	annual excess fuel consumed per auto-commuter (gallons)
HDPA	annual hours delay per auto-commuter
TTI	travel time index
CCPA	annual congestion cost per auto-commuter (dollars)

Table 3: Table of our regressands of interest from this model. This data was collected through the TTI UMR dataset.

Variable	Definition
y_{ct}	output regressand of interest (EFPA, HDPA, TTI, or CCPA)
β	regression coefficients for ride-hailing regressors
u_{ct}	indicator for Uber operating in city c at time t
uPl_{ct}	sum of Uber and Lyft indicators for city c at time t
l_{ct}	indicator for Lyft operating in city c at time t
ϵ_{ct}, η_{ct}	error terms for city c at time t
z_{1ct}	unemployment rate in city c at time t
z_{2ct}	percentage of population aged 65 or older in city c at time t
Φ	regression coefficients for control variables
c_{ct}	regression control variables

Table 2: Table of variable definitions for our empirical models.

We are specifically interested in how $EFPA_{ct}$, $HDPA_{ct}$, TTI_{ct}, and $CCPA_{ct}$ depend on u_{ct} and l_{ct}. We will leverage our other controls, fixed effects, and instrumental variables in tandem with our regressors of interest to estimate the causal effect of these ride-hailing indicators. For each type of model, we will regress on each of our regressands of interest individually (as opposed to together), since they are likely highly correlated with each other (e.g. excess fuel and traffic delay times are highly correlated with one another).

For clarity, let y_{ct} denote any of our regressands in question ($EFPA_{ct}$, $HDPA_{ct}$, TTI_{ct}, and $CCPA_{ct}$). The models below, specified for a regressand of y_{ct}, are symbolically duplicated for each of our regressands of interest (i.e. the same model, with likely different coefficients produced, is run on each regressand).

4.1 Panel Regression

Panel regression models make no assumptions about the idiosyncratic features of each cross-sectional or time series slices of the dataset, and thus there are no variables (aside from the constant factor) that are shared between multiple observations, aside from the estimated coefficients. The empirical panel regression models we use here are given by:

$$y_{ct} = \beta_0 + \beta_1 u_{ct} + \beta_2 l_{ct} + \epsilon_{ct} \quad (\textbf{Panel Regression}) \tag{5}$$

In addition, for our panel regression model and for all our subsequent models, we estimate robust standard errors.

4.2 Fixed Effects

Fixed effects models are ubiquitous to many empirical studies involving the use of panel data. They enable for the automated estimation and incorporation of temporal and spatial idiosyncrasies that standard multiple regression models simply cannot always capture. Mathematically, these regression models are specified by:

$$y_{ct} = \beta_0 + \alpha_c + \omega_t + \beta_1 u_{ct} + \beta_2 l_{ct} + \epsilon_{ct} \quad (\textbf{Fixed Effects}) \tag{6}$$

4.3 Fixed Effects with Controls

Though fixed effects model enable for the estimation and use of many idiosyncratic features that are common in panel datasets, using additional control regressors can enable for us to more accurately ascertain what the true causal effects of ride-hailing are on traffic congestion and travel time index. Our regression models using fixed effects with controls are given below.

$$y_{ct} = \beta_0 + \alpha_c + \omega_t + \beta_1 u_{ct} + \beta_2 l_{ct} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{c}_{ct} + \epsilon_{ct} \quad (\textbf{Fixed Effects with Controls}) \tag{7}$$

Where $\mathbf{\Phi} \in \mathbb{R}^d$ is our vectors of d control coefficients, and $\mathbf{c}_{ct} \in \mathbb{R}^d$ is our vector of control variable observations for city c and time t:

$$\mathbf{\Phi} \equiv [\phi_1 \quad \phi_2 \quad \dots \quad \phi_d]^{\mathrm{T}} \in \mathbb{R}^d \tag{8}$$

$$\mathbf{c}_{ct} \equiv [c_{1ct} \quad c_{2ct} \quad \dots \quad c_{dct}]^{\mathrm{T}} \in \mathbb{R}^d \tag{9}$$

The controls leveraged in this model, as mentioned previously, were included in this regression model in order to reduce the presence of omitted variable bias, thus ultimately providing us with ride-hailing traffic effect estimators that are (more) unbiased. The controls we used were (note: these controls were all indexed with respect to city and year):

Control Variable (city c, time t)
Population (1000s)
Number of auto-commuters (1000s)
Daily miles of freeway traveled
Daily miles of “arterial road" traveled
Value of time (general and commercial)
Cost of gasoline and diesel

Table 4: Table of our control variables used for controlling omitted variables bias.

4.4 Instrumental Variables, Fixed Effects, and Controls

Finally, one of the main motivations for this paper was the inherent presence of endogeneity from ride-hailing platforms such as Uber and Lyft selecting to enter into certain urban areas at certain times. This leads (at least to some degree) to reverse causality, which makes disentangling the causal effects of ride-hailing on traffic congestion and congestion-related emissions substantially more difficult to estimate. One way in which we an develop a more accurate estimate of this causal effect is through the use of instrumental variables. These regressors mitigate the effect of endogeneity by varying with the regressors of interest (u_{ct} and l_{ct}), but not the regressands (y_{ct}). In this case, the conditions for an instrument z_{ct} to be valid are the following:

$$\text{Cov}(u_{ct}, z_{ct}) \neq 0, \quad \text{Cov}(l_{ct}, z_{ct}) \neq 0, z_{ct} \text{ only co-varies with } y_{ct} \text{ through } u_{ct} \text{ and } l_{ct}. \tag{10}$$

These conditions necessitate that our instrumental variables be related to the regressors for which we are seeking to determine the causal effects of (relevance condition), and only related to the output regressands of interest through the endogenous variable (exclusion condition). Two instruments were proposed by Li, Hong, and Zhang [3]: unemployment rate and percentage of population over the age of 65. Each of these instruments is discussed individually below:

- The motivation for why the unemployment rate was chosen was because if the unemployment rate in a city is high, this may yield economic circumstances that make people more likely to drive for ride-hailing companies at a lower salary, which, for profit-maximizing firms such as Uber and Lyft, is in their interest. Thus, these ride-hailing companies, all things equal, will likely self-select into cities with higher unemployment rates. Additionally, these companies will likely self-select into cities with more transportation, which, again all things equal, likely means that these ride-hailing companies self-select into cities with high traffic congestion. However, it is also worth noting that a higher unemployment rate implies there will be less traffic on the road, reducing traffic congestion and therefore being related to an output regressand through means unrelated to ride-hailing. Therefore, the validity of this instrument is questionable.
- The second instrument explored in [3] was the percentage of population aged 65 or older. The intuition behind this instrument was that because people aged 65 or older have a history of not using ride-hailing services [3] in the same frequency as other age groups, ride-hailing services such as Uber and Lyft will be less inclined to enter new urban areas with higher populations of older people. The efficacy behind this instrument is more intuitive than in the case above, though further testing should be done to confirm this.

For this paper, we implemented instrumental variables using a two-stage least squares approach, with the goal of mitigating endogeneity via reverse causality. Because we only had access to a single instrument, we introduced a new variable that is the sum of our Uber and Lyft indicators. We introduced this variable because with our instrumental variables model, we cannot have more instruments than endogenous regressors; hence we have now constrained ourselves to having one endogenous regressor. With this established, we can now use two-stage least squares (note: the controls here are the same as the controls above):

$$\hat{uPl}_{ct} = \pi_0 + \alpha_c + \omega_t + \pi_1 z_{ct} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{c}_{ct} + \eta_{ct} \quad (\textbf{Stage 1}) \tag{11}$$

$$y_{ct} = \beta_0 + \alpha_c + \omega_t + \beta_1 \hat{uPl}_{ct} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{c}_{ct} + \epsilon_{ct} \quad (\textbf{Stage 2}) \tag{12}$$

From econometric theory, if the instrument used is valid, then our estimates for the effects of our regressors of interest on our output regressands of interest will no longer be subjected to endogeneity and reverse causality.

5 Results

Using our combined and pre-processed dataset, we tested our panel regression, fixed effects, fixed effects with controls, and two instruments (separately) with fixed effects and controls models on our dataset to determine the estimated coefficients for Uber/Lyft entry on our output variables of interest. The results from running our dataset on these models are given below:

5.1 Quantitative Results

Testing our 5 models on our data produced the table of results below. There are two observations from this table which we wish to highlight: (1) There is a negative statistically-significant regression coefficient for congestion cost per auto-commuter (obtained via Fixed Effects + Controls), and (2) Generally, as we control for more omitted variables and add instruments, the effects of ride-sharing on traffic congestion tend to decrease, suggesting that there is substantial omitted variables bias and endogeneity present in a simple panel regression model.

Model	**Regressor**	**Regressand**							
-	-	**EFPA**		**HDPA**		**TTI**		**CCPA**	
-	-	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value
Panel OLS	uber (u_{ct})	7.91^{***} (0.67)	0.00	18.0^{***} (1.61)	0.00	0.07^{***} (0.01)	0.00	288^{***} (37.0)	0.00
Panel OLS	lyft (l_{ct})	1.92^{***} (0.74)	0.01	3.61^{**} (1.81)	0.05	0.01 (0.01)	0.60	44.8 (41.0)	0.28
Fixed Effects (FE)	uber (u_{ct})	0.84^{***} (0.25)	0.00	2.20^{***} (0.47)	0.00	0.004^{*} (0.03)	0.09	5.07 (10.8)	0.64
Fixed Effects (FE)	lyft (l_{ct})	0.68^{***} (0.24)	0.01	1.40^{***} (0.44)	0.00	0.004 (0.02)	0.11	-11.2 (10.2)	0.27
FE + Controls	uber (u_{ct})	-0.22 (0.23)	0.33	1.12^{**} (0.45)	0.01	0.00 (0.002)	0.99	-37.7^{***} (9.65)	0.00
FE + Controls	lyft (l_{ct})	0.30 (0.21)	0.16	0.87^{**} (0.42)	0.04	-0.00 (0.002)	0.80	-13.8 (9.09)	0.13
FE, Controls, IV = $\mathbf{z_{1ct}}$	uber + lyft (uPl_{ct})	-0.85 (2.34)	0.72	1.89 (3.39)	0.58	0.04 (0.03)	0.24	40.5 (71.2)	0.57
FE, Controls, IV = $\mathbf{z_{2ct}}$	uber + lyft (uPl_{ct})	-16.91 (36.3)	0.64	-1.18 (11.1)	0.92	-0.04 (0.095)	0.67	-590.5 (1250.7)	0.64

Table 5: HAC standard errors are given in parentheses after each coefficient, and the significance of each coefficient is denoted according to: $^{***}: p \leq 0.01$, $^{**}: p \leq 0.05$, $^{*}: p \leq 0.1$.

5.2 Analysis of Results

The results above are particularly interesting because we find that as we add more variables and controls to our models, we become less confident that Uber, Lyft, and likely other ride-hailing services increase traffic congestion and emissions. As our models increase in complexity, the observed p-value tends to increase as well. One explanation for why this may be happening is because regardless of the effect that ride-hailing has on traffic congestion and traffic problems, there exist other variables in this panel dataset that also lead to increased traffic congestion and emissions, and not including these variables in our regression models simply adds the effect that would have been seen through other regressors to the estimator for ride-hailing (definition of omitted variable bias). Therefore, from this we can conclude that substantial endogeneity exists, and because we see high p-values from our full model, we cannot conclude that ride-hailing technologies increase traffic congestion and emissions.

This conclusion also comes with the caveat that we cannot make many broad conclusions about ride-hailing technologies decreasing traffic congestion and emissions. One exception to this, which may have important policy implications, is the regression coefficient for CCPA using our fixed effects with controls model. This coefficient is negative and statistically-significant, indicating that it is likely that ride-hailing services actually decrease congestion costs per auto-commuter. One interpretation of this is that while ride-hailing does cost money, perhaps part of the revenue that ride-hailing drivers bring in through ride-hailing trips is reflected as countervailing positive revenue relative to the congestion costs imposed on the riders who use ride-hailing services.

5.2.1 Explanation of Observed Results

As mentioned previously, one reason why the estimates for the effects of these ride-hailing technologies decreased over time is because of the mitigation of omitted variables bias and endogeneity:

- **Analyzing Omitted Variables Bias**: Mitigating the presence of omitted variables bias essentially means removing the effect from the estimator that other regressors have on the regressand. In this case, since traffic congestion has simply grown worse over time due to growing urban populations, and ride-hailing technologies have simply happened to come into existence during this worsening of traffic congestion, without including any other variables it would appear as though ride-hailing technologies are the case of this worsening of traffic congestion, when in reality they are simply correlated with it.
- **Analyzing Endogeneity**: As discussed above in the description of the instrumental variables model, ride-hailing technologies tend to self-select themselves into cities that are already subjected to bad traffic congestion. Not taking this into account could lead us to mistakenly conclude that ride-hailing technologies were the cause of this traffic congestion, as opposed to the reverse case (i.e. reverse causality).
- **Instruments**: From our analyses above, and by analyzing the consistency of the signs of the coefficients on each regressand of interest for our instrumental variables model when our instrument was unemployment rate, it is unlikely (but not impossible) that unemployment rate was a good instruments. On the contrary, our

intuition and the quantitative results for using instrumental variables regression with percentage of population aged 65 or older suggests that it has the potential to be both a valid and powerful instrument.

6 Conclusions

The results from these models have provided us insight into the effects that ride-hailing platforms such as Uber and Lyft have on traffic congestion and travel time index. The results here don't have the statistical significance to confirm that ride-hailing technologies decrease traffic congestion and cost, but also cannot validate that they do (which is nonetheless a large step for this specific research topic). The results here provide policymakers, ride-hailing business leaders, and the general public with results that will enable for the intelligent penetration and integration of these technologies into new and existing regions across the globe, as well as related transportation technologies such as autonomous ride-hailing systems.

6.1 Future Work in Improving These Results

6.1.1 Additional Instruments

More work can be done to create additional instruments and leverage other control variables in order to obtain more accurate estimates of the effects of ride-hailing on traffic congestion and emissions. Namely, we could further investigate the use of regulatory controls as instruments for ride-hailing entry into cities, since it is quite unlikely these are correlated with traffic congestion and/or emissions.

6.1.2 Additional Control Variables

One control I would like to see introduced is the inflation rate, particularly when analyzing the effects of ride-hailing technologies on the cost of congestion. This bias created by inflation is in part mitigated by the inclusion of the value of time, but it would be interesting to see if including the inflation rate explicitly could also produce more accurate and definitive estimates of these effects.

6.2 Summary of Findings

This work finds that endogeneity is largely a reason for ride-hailing technologies to seemingly be the principal causes of traffic congestion and traffic-related emissions. This endogeneity manifests because Uber, Lyft, and other ride-hailing technologies self-select into the cities they bring their services to based off of salient demographic and economic features. We chose to use this insight to design a combined instrumental variables, fixed effects, and controlled panel regression model to mitigate the presence of this endogeneity. Though our findings from testing this model on our dataset, which consisted of a merging and filtering of the 2019 Urban Mobility Report dataset, Uber and Lyft entry data, unemployment data from the Bureau of Labor Statistics, and age data from the U.S. Census are inconclusive, we are still able to conclude that ride-hailing technologies do not provably increase traffic congestion and emissions. Furthermore, we identified a statistically-significant negative coefficient for congestion cost per auto-commuter for Uber, indicating that the presence of Uber in fact decreases congestion costs per auto-commuter.

While additional work on this problem remains to be done, this work helps policymakers, business-people, and the public to separate the (still-to-be-determined) causal effects of these ride-hailing technologies from sources of endogeneity and omitted variables bias.

References

[1] Regina R Clewlow and Gouri S Mishra. "Disruptive transportation: The adoption, utilization, and impacts of ride-hailing in the United States". In: (2017).

[2] Phil Lasley. "2019 URBAN MOBILITY REPORT". In: (2019).

[3] Ziru Li, Yili Hong, and Zhongju Zhang. "Do on-demand ride-sharing services affect traffic congestion? evidence from uber entry". In: *Evidence from Uber Entry (August 30, 2016)* (2016).

[4] Bruce Schaller. "The new automobility: Lyft, Uber and the future of American cities". In: (2018).

[5] Anil, Mark43, Ellison et al. “Regulatory Distortion: Evidence from Uber’s Entry Decisions in the US”.

[6] U.S. Bureau of Labor Statistics (2019). BLS Data Finder (version 1.1).

[7] U.S. Census Bureau. Older Population and Aging.

[8] Chang et al. “Scaling Relationship between Traffic Congestion versus Population Size of 164 Global Cities".

[9] Fig. 1. Penetration Rate in the Ride Hailing Market; statista.com, https://www.statista.com/outlook/368/100/ride-hailing/worldwide.

[10] “Set a Driver Destination.” Uber Help, Uber Technologies, https://help.uber.com/driving-and-delivering/article/set-a-driver-destination?nodeId=f3df375b-5bd4-4460-a5e9-afd84ba439b9.