Estimating Causal Effects for Ride-sharing Platforms with Instrumental Variables: A Case Study*

MSIN0094 Case Study

Dr. Wei Miao

November 27, 2024

1 Industry Background

The sharing economy has been booming in recent years, leading to a rapid increase in jobs in the "gig" economy. According to Hossain (2020), in the US alone, the sharing economy sector has created 6.23 million jobs with 78 million service providers, and 800 million people engage with it. The transportation sector is one of the most salient beneficiaries of the burgeoning sharing economy. For instance, commuting to work by shared bicycle (e.g., Citi Bike) has become an increasingly popular transportation option (Ford et al. 2019). The ride-sharing service (e.g., Uber) allows drivers to enjoy more flexibility in work, which is proven valuable to drivers and has improved capacity utilization (Cramer and Krueger 2016).

The COVID-19 pandemic has brought unprecedented disruptions to many industries, and the transportation industry is among the most disrupted ones. Further, the COVID-19 has raised concerns about the survivability of the sharing economy in general. It is reported that gross bookings on Uber rides were down by 75% in the three months through June 2020, and that Lyft's April ridership was down by 75% from April 2019. Some Uber drivers were extremely cautious about their shift decisions and taking measures to prevent COVID from spreading.

Unlike the traditional taxi market, where taxi drivers rent vehicles from taxi companies and then directly provide transportation services to consumers, modern ride-sharing platforms typically serve as the matching intermediary between drivers and passengers. Due to such two-sided market nature, the profitability of modern ride-sharing platforms (and sharing economy in general) highly depends on the interdependence or externality between the two sides of economic agents (Rysman 2009). Therefore, **a ride-sharing platform would benefit from the network effect if more drivers work for them.** It is thus managerially important for the ride-sharing platform to understand whether COVID-19 has affected drivers' labor supply patterns and if yes, the magnitude of the effect across drivers and over time.

In this case study, we will answer the above causal question using the instrumental variable method.

2 Data Description and Data Wrangling

2.1 Driver Daily Trip Data

In a ride-sharing company's database, the raw trip log records each trip's details, including the driver's ID, the passenger's ID, the booking date and time, the trip's start and end locations, the trip's distance, the trip's fare, and the trip's status (e.g., completed, cancelled, or passenger no shows). The data science team has aggregated the **raw trip-level data** into a **driver-day** level **panel data**.

Panel data structure refers to a dataset that includes multiple observations over time for the same units (in our case, drivers). It combines cross-sectional data (observations at a single point in time) and time series data (observations of a single subject over multiple time periods), thus enabling analysis that captures both individual dynamics and temporal variations.

Our first data set summarizes drivers' daily shift each day in April 2020, right during the period when the pandemic began to spread in the UK. The data set consists of a random sample of around 4000 drivers across 3 UK cities (anonymized as `g`, `s`, and `c`) in 2020.

2.2 COVID-19 Data

To measure the severity of COVID-19, the data science team collected daily number of new cases in each city from the government database.

2.3 Data Wrangling

> **Question 1**
>
> - Join the two datasets using `dplyr`. Please observe the data structure of the two datasets and carefully think about how we should do the data join in this case. Explain your rationale.

2.4 Key Dependent Variables

To facilitate the empirical analysis of drivers' responses to COVID-19, the data science team has computed key outcome variables of interest for each driver, including both extensive margin (i.e., whether to work) and intensive margin (i.e., how much to work) of drivers' labor supply.

1. **Whether or not to work on a day**, a binary outcome variable which equals 1 if a driver has at least one ride request on the day and 0 otherwise. We can use this variable to measure drivers' shift decision, i.e., willingness to work on a day, which proxies for the extensive margin of drivers' labor supply. It is ambiguous ex-ante how the number of new cases affects a driver's shift decision. On the one hand, more new cases may increase the risk of infection, which decease drivers' expected wellbeing, and therefore discourage drivers from working on a specific day; on the other hand, fewer drivers on the street suggest less competition among drivers and therefore higher chances of getting a passenger and potentially higher

hourly earnings, which may motivate drivers to work. It is important for the ride-sharing company to understand how the severity of COVID-19 affects drivers' willingness to work, so that the company can adjust their stimulus plans for drivers accordingly.

2. **Total number of completed orders**, which contain three aspects of information which are of policy and managerial interest. First, the variable can proxy for the length of drivers' daily labor supply. Conditional on working, if a driver decides to work for longer hours, then we expect the driver to have a larger number of requests/orders. Second, both variables contain information on consumer demand. We expect the total number of requests/orders to decrease if there is a lower demand for ride-sharing service from consumers due to the COVID-19 outbreak. Finally, both variables can measure the intensity of competition among drivers. Keeping the level of demand fixed, the total number of requests/orders would be larger when there are fewer drivers working on the day. Due to the complexity of information contained, ex-ante, it is not straightforward how the COVID-19 measures affect the total number of orders for individual drivers.
3. **Earnings**. Earnings measure the driver's income from providing ride-sharing services, which is highly correlated with the number of completed orders and total trip distance. It allows us to directly assess the impact of the COVID-19 on drivers' financial wellbeing.
4. **Average trip distance**. In our empirical context, drivers cannot reject a booking request once being matched with a passenger, therefore, the trip distance is largely determined by passengers. Since passengers may be reluctant to take long distance trips during the pandemic, we expect a negative impact of the number of new cases on the average trip distance.

3 Empirical Analysis

3.1 OLS Linear Regression

To empirically investigate the causal impact of COVID-19 cases on driver behavior, we can first try linear regressions to regress the labor supply measures of driver i, in city j, on day t on the COVID-19 measure and other covariates as follows:

$$LaborOutcome_{ijt} = \beta_0 + \alpha NewCases_{jt} + \varepsilon_{ijt} \tag{1}$$

where $LaborOutcome_{ijt}$ is the dependent variable of interest, $NewCases_{jt}$ is the daily new COVID-19 cases in city j on day t, and ε_{ijt} is the error term.

> **Question 2**
> Please run linear regressions based on the above equation, with the outcome being the aforementioned dependent variables and explanatory variable being new cases. Please report the results in a single table.

3.2 Fixed Effect OLS Regressions

What confounding factors do we need to control in the above OLS regressions to mitigate the omitted variable bias?

We first need to include **driver fixed effects** to control for **driver-specific** characteristics that may affect drivers' labor supply patterns. Such characteristics include, but are not limited to, the driver's socio-demographic characteristics (e.g., gender and age), the driver's degree of risk aversion, whether a driver is driving full-time or part-time, and the driver's innate abilities to search for passengers, etc.

For instance, less risk-averse drivers may prefer to work on days when there are more new cases because they expect less competition from peer drivers and potentially higher profitability on such days. Another example is that, full-time drivers can be more subject to the impact of new cases compared to part-time drivers, because full-time drivers' income largely comes from providing ride-sharing services via the focal company. Driver fixed effects can mitigate such driver-specific time-invariant confounding effects and help us obtain more accurate estimates for our focal explanatory variable `NewCases`.

In addition to driver fixed effects that remove cross-sectional confounding effects across drivers, we also include **time fixed effects** in Equation (1) to mitigate the inter-temporal confounding effects. We consider time fixed effects at the day level.

Moreover, given that the local government in each city may have enacted different policies on fighting COVID-19 and/or stimulating economy (e.g., subsidizing drivers) during our data period, we further control for **city fixed effects**.

$$LaborOutcome_{ijt} = \beta_0 + \alpha NewCases_{jt} + DriverFE + DayFE + CityFE + \varepsilon_{ijt}$$

> **Question 3**
> Run the fixed effect regressions for the dependent variables. Please report the results in a single table.

4 Instrumental Variable Analyses

4.1 Potential Endogeneity

After including the driver, city, date fixed effects fixed effects in the above regression, the remaining challenge to obtaining causal inference is the potential reverse causality of `NewCases`.

Equation (1) could be subject to **simultaneity issues** because drivers' labor supply decisions and number of new cases may be interdependent. On the one hand, drivers may adjust their labor supply accordingly to the number of new cases. On the other hand, prior research has demonstrated the potential effect of mobility on the COVID-19 case growth rate. If a city has a higher volume of private transportation through ride-sharing services, given the highly contagious nature of COVID-19, the city may have a higher number of new cases.

4.2 Instrumental Variables

To tackle the potential endogeneity issue, we use the instrumental variable (IV) method, leveraging exogenous sources of variation in the explanatory variable that are uncorrelated with the error term in Equation (1) using two-stage least squares (2SLS). We can potentially select two instrumental variables.

The first instrumental variable is *imported new cases*, which measures the number of infected travelers from overseas in each city as disclosed by local government. Because the imported cases relate to travelers from overseas, it should be exogenous to local confirmed cases and meet the **exogeneity requirement**.

The second instrumental variable is *other city new cases*, which is the number of new cases confirmed in neighboring cities. Since confirmed cases in other cities should not directly affect the focal city's ride-sharing market, the variable *other city new cases* should also satisfy the **exogeneity requirement**.

The first-stage regression is specified below in Equation Equation 2, where the definitions of variables are the same as in Equation Equation 1:

$$NewCases_{jt} = \pi_0 + \pi_1 OtherCityNewCases_{jt} + DriverFE + DayFE + CityFE + \varepsilon_{ijt} \tag{2}$$

> **Question 4**
> Run the first stage regression and report the results.

In the second stage regression, we regress the outcome variables on the predicted new cases from the 1st stage, controlling for the same set of control variables:

$$LaborOutcome_{ijt} = \beta_0 + \alpha \hat{NewCases}_{jt} + DriverFE + DayFE + CityFE + \varepsilon_{ijt}$$

> **Question 5**
> Run the second stage regression and report the results.

References

Cramer, Judd, and Alan B. Krueger. 2016. "Disruptive Change in the Taxi Business: The Case of Uber." *American Economic Review* 106 (5): 177–82. https://doi.org/10.1257/aer.p20161002.

Ford, Weixing, Jaimie W. Lien, Vladimir V. Mazalov, and Jie Zheng. 2019. "Riding to Wall Street: Determinants of Commute Time Using Citi Bike." *International Journal of Logistics Research and Applications* 22 (5): 473–90.

Rysman, Marc. 2009. "The Economics of Two-Sided Markets." *Journal of Economic Perspectives* 23 (3): 125–43.