Class 17 Case Study: Estimating Causal Effects for Platform Businesses Using Instrumental Variables

Dr. Wei Miao

UCL School of Management

November 27, 2024

Section 1

Causal Questions for Platform Businesses

Class objectives:

- Understand the importance of causal inference for platform businesses.
- Learn how to estimate causal effects using instrumental variables with an application to ride-sharing platforms.

Causal Questions for Platform Businesses

- Platform businesses often need to answer critical causal questions to optimize their operations:
  - **Measuring Network effects**: How does increasing supply affect demand (and vice versa)?
  - **Pricing**: How does surge pricing affect consumer demand and driver supply?
- When relying on secondary non-experimental data, these questions often face **endogeneity** challenges that require careful empirical strategies.

Causal Question Example

How to estimate the causal effect of surge prices on driver work decisions using historical data?

- We can run a linear regression model on Uber's historical data, where the dependent variable is the number of drivers on the road in an hour; the key explanatory variable is the surge multiplier during that hour.

$$\text{NumberDrivers} = \beta_0 + \beta_1 \text{SurgeMultiplier} + \varepsilon$$

- Is there endogeneity in this model?[1]

[1] Answers are available on the HTML version.

Case Study Background

- **Core case question**: How did new COVID-19 cases causally affect drivers' labor supply patterns?

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

- OLS linear regression model
  - $LaborOutcome_{ijt}$: Driver i's labor supply (e.g., whether worked that day) in city j on day t
  - $NewCases_{jt}$: Daily new COVID-19 cases in the city t on day t

Section 2

Data

Driver Daily Trip Data

- Driver-level daily trip data from a ride-sharing platform. About 4000 drivers across 3 UK cities in April 2020.

	driver_id	booking_date	is_work	income	n_order	avg_distance	city
1	1	2020-04-01	0	0	0	0	g
2	1	2020-04-02	0	0	0	0	g
3	1	2020-04-03	0	0	0	0	g
4	1	2020-04-04	0	0	0	0	g
5	1	2020-04-05	0	0	0	0	g

COVID-19 Data

- Daily new cases by city, which serves as key explanatory variable the X

	city	booking_date	new_cases	other_city_new_cases
1	g	2020-04-01	1	0
2	g	2020-04-02	1	0
3	g	2020-04-03	1	0
4	g	2020-04-04	0	0
5	g	2020-04-05	3	0

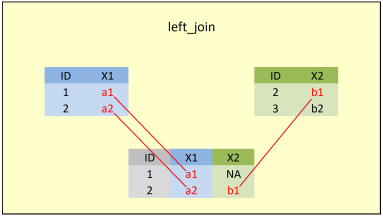
Join Multiple Data Frames

- We can consolidate (merge, join) multiple data frames using the `left_join()` function in the `dplyr` package.
- We need to determine the **main data frame** that will be retained as the final data for analyses. The other data frame will be used as the supplementary data to provide additional information.
  - We often use **the most granular data frame** (usually panel data) as the main data frame
  - The less granular data such as demographic data can be joined onto the main data frame
- In this case, we will use the driver data as the main data frame and join the COVID-19 data onto it.

`left_join()`

- `left_join` keeps everything from the **left data frame** and matches as much as it can from the right data frame based on the chosen IDs.
  - Choose the **longer** data frame as the left data frame. All IDs **in the left data frame** will be retained
  - If a match can be found, value from the right data frame will be filled in
  - If a match cannot be found, a missing value will be returned

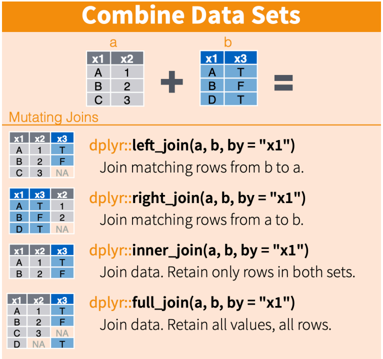
```
df_1 %>%
    left_join(df_2, by = c("ID" = "ID"))
```



- **Exercise**: Join the driver data with the COVID-19 data using the `left_join()` function.

Other joins in `dplyr`

- `inner_join()`: Keeps only the IDs that are present in both data frames
- `right_join()`: Keeps everything from the right data frame and matches as much as it can from the left data frame based on the chosen IDs
- `full_join()`: Keeps everything from both data frames and matches as much as it can based on the chosen IDs



Section 3

Empirical Strategy

OLS Linear Regression

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

- Omitted variable bias: local city policies which affect both COVID cases and driver behavior (lockdowns, mask mandates, etc.)
- Reverse causality: Drivers may reduce labor supply in response to COVID cases, but COVID cases may also increase due to driver behavior
- **Exercise**: Run the linear regression model on Quarto.

Fixed Effects Regression

- Extended model with fixed effects:
  - Driver fixed effects: Control for inherent, time-invariant driver characteristics
  - Time fixed effects: Controls for temporal trends common in all cities
  - City fixed effects: Control for local policies and other time-invariant city characteristics

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

- **Exercise**: Run the fixed effects regression model on Quarto.
- Is there still endogeneity in this model?[2]

[2] Answers are available on the HTML version.

Instrumental Variables Regression

- Instruments that satisfy (1) relevance and (2) exogeneity and (3) exclusion restriction:
  - Candidate 1: Imported new cases from overseas
  - Candidate 2: Cases from neighboring cities

Two-Stage Least Squares (2SLS) Estimation: First Stage

- First stage: Regress endogenous variable on instruments including all control variables.

$$NewCases_{ijt} = \pi_0 + \pi_1 Z_{ijt} + DriverFE + DayFE + CityFE + \varepsilon_{ijt}$$

- Practical considerations:
  - Check for instrument relevance: Instruments should be correlated with the endogenous variable, which can be tested whether the coefficient of Z is significantly different from zero
  - Exogeneity and exclusion restriction are untestable assumptions, and we need to justify them based on the context.
  - The same set of control variables must be included in both stages. In our case, the 3 sets of fixed effects must be included in both stages.
- **Exercise**: Run the first stage regression model on Quarto.

Two-Stage Least Squares (2SLS) Estimation: Second Stage

- Second stage: Regress labor outcome on predicted new cases from the first stage and all control variables

$$LaborOutcome_{ijt} = \beta_0 + \alpha New\hat{C}ases_{ijt} + DriverFE + DayFE + CityFE + \varepsilon_{ijt}$$

- The coefficient α is the causal effect of new COVID cases on driver labor supply.
- **Exercise**: Run the 2SLS regression model on Quarto.

After-Class Reading

- (highly recommended) Encouragement Designs and Instrumental Variables for A/B Testing at Spotify