Class 15 Endogeneity

Dr Wei Miao

UCL School of Management

November 22, 2023

Section 1

Causal Inference with OLS

Causal Effect from Linear Regression Models

- ***Task***: Tesco would like to understand the causal impact of customer $Income$ on customer $Spending$
- Please run the two regressions on your laptop:
  - Regression 1: $Spending \sim Income$
  - Regression 2: $Spending \sim Income + Kidhome$

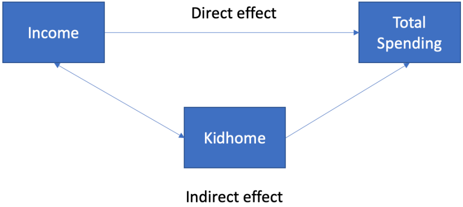
Regression Results

	(1)	(2)
(Intercept)	-556.823***	-299.119***
	(21.654)	(28.069)
Income	0.022***	0.019***
	(0.000)	(0.000)
Kidhome		-230.610***
		(16.945)
Num.Obs.	2000	2000
R2	0.629	0.661
R2 Adj.	0.629	0.660
AIC	29 306.1	29 130.7
BIC	29 317.3	29 147.5
RMSE	367.45	351.51
Std.Errors	IID	IID

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

- ***Question***: if we want to evaluate income's causal effect on spending, which value (0.022, 0.019) should we use?

Direct and Indirect Effects



- Direct Effect (causal effect)
  - Keeping other variables fixed (ceteris paribus)
  - Direct effect only
- Total Effect
  - Including side effects through other variables
  - Direct + indirect effects

Causal Inference from Regression Models

- To obtain causal inference, we need to obtain the **direct effects** of an X variable on the outcome variable Y.
- **Total effects** include both direct effects and indirect effects (i.e., the impacts of other confounding variables).
- Therefore, it is important to include **all confounding variables**, which affect income and total spending at the same time, to control for the side effects from other variables.

Practical Suggestions for Running Regression Models

- For causal inference tasks, we need to use business senses to decide which confounding variables to control.
    - good controls and bad controls
- Sometimes, control variables may be statistically insignificant, they should **NOT** be removed because they still serve the purpose of control variables.
- If some variables are mechanically correlated, then we should not put all of them in the regression, to avoid perfect collinearity problems.

Question: what is the best you can do with `data_full` *to estimate the causal effect of income on spending?*

Causal Inference from Regressions

Now we have included `Kidhome` to tease out the effect of kids, what problems do we still have which hinder us from getting causal effect of income on total spending?

- Due to data availability, we are never able to include all confounding variables in the regression.
- Strictly speaking, we can **never obtain causal effects from simple regression models** based on **non-experiment data**.
- Mathematically speaking, because we can never control all confounding factors, the error term is very likely to be correlated with income, violating $E[\epsilon|X] = 0$.

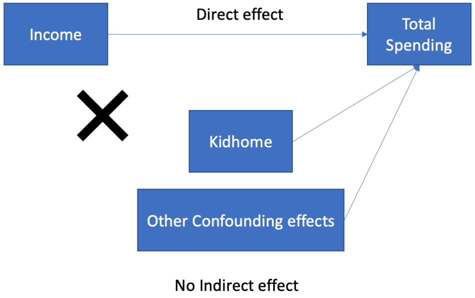
RCTs and Causal Inference

- Why RCTs are the gold standard for causal inference?
  - If we are able to randomize people into different income groups, we can then collect the `total_spending` for each individual in each `income` group.
  - We can run a linear regression to examine the impact of `income` on `total_spending`.

$$Spending = \beta_0 + \beta_1 Income + \epsilon$$

- In the above regression
  - Are there still any confounding effects?
  - Is $Income$ correlated with any of the confounding effects?

RCTs and Causal Inference (cont.)



Section 2

Endogeneity and Its Causes

Endogeneity

Endogeneity

Endogeneity refers to an econometric issue with OLS linear regression, in which an explanatory variable is correlated with the error term, such that the requirement for OLS linear regression $E[\epsilon|X] = 0$ is violated.

Cause I: Omitted Variable Bias

Omitted Variable Bias (OVB)

An omitted variable is a determinant of the outcome variable y_i that is correlated with the focal explanatory variable x_i, but is not included in the regression, either due to data unavailability or ignorance of data scientists.

Two conditions for omitted variable bias

- The variable affects the dependent variable.
- The variable is correlated with the focal explanatory variable.

Examples of OVB

- If we would like to understand the causal effect of years in education on a person's salary.

$$Salary_t = \beta_0 + \beta_1 Education_t + \epsilon_t$$

- Can we get causal effect from this regression? What would be the issue here?

Examples of OVB

- When building Marketing Mix Modeling, the common practice in the industry is to regress the sales in each period on the price in each period.

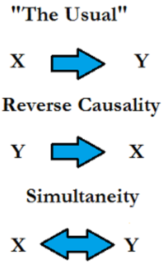
$$Sales_t = \beta_0 + \beta_1 Price_t + \epsilon_t$$

- However, is this regression correct?
  - Very often, if we regress sales on price, we get a positive coefficient for price.

Cause II: Reverse Causality (Simultaneity)

Reverse Causality

Reverse causality refers to the phenomenon that the independent variable X_i affects the dependent variable y_i and the dependent variable y_i also affects the independent variable X_i at the same time.



Examples of Reverse Causality (Simultaneity)

- Besides potential omitted variable biases, there may also exist reverse causality problems with marketing mix modelling.

$$Sales_t = \beta_0 + \beta_1 Price_t + \epsilon_t$$

- Price affects demand, and demand affects sellers' price setting decisions.
  - Higher price leads to lower sales. (X => Y)
  - If sellers expect higher demand, sellers may increase the price to increase profits. (Y => X)

Examples of Reverse Causality (Simultaneity)

- UberEat interview question: If we have historical data on **number of restaurants on UberEat** in each month, and **the total number of orders in each month**, can we run an OLS regression to get the causal effect?

$$NumOrders_t = \beta_0 + \beta_1 NumRestaurants_t + \epsilon_t$$

- If not, how can we measure the causal effects for UberEat?
- This question is not just limited to UberEat; it is in fact related to any platform business with network effect!
  - Amazon; Airbnb; Uber Ridesharing; etc.

Main Takeaway

- Common threats to causal inference from secondary data include
  - Omitted Variable Bias
  - Reverse Causality
- We can overcome the endogeneity problem using **instrumental variable method**.