Class 15 Endogeneity and Its Causes

Dr. Wei Miao

UCL School of Management

November 19, 2025

Section 1

Causal Inference with OLS

Class Objective

- Understand the reasoning why linear regression can almost never provide causal effects from non-experimental data.
  - Direct and Indirect Effects
  - Causal Inference from Regression Models
  - Understand the difference between RCT and non-experimental data.
- Understand the concept of endogeneity and its causes.
  - Omitted Variable Bias
  - Reverse Causality
  - Measurement Error (Optional)

Causal Effect from Non-Experimental Secondary Data

- ***Task***: M&S wants to understand the causal impact of customer $Income$ on customer $Spending$, i.e., the Marginal Propensity to Consume (MPS).[1]
- Please run the two regressions on your Quarto document and export the regression table:
  - Regression 1: $Spending \sim Income$
  - Regression 2: $Spending \sim Income + Kidhome$

[1] In economics, MPC refers to the proportion of an additional unit of income that is spent on consumption.

Regression Results

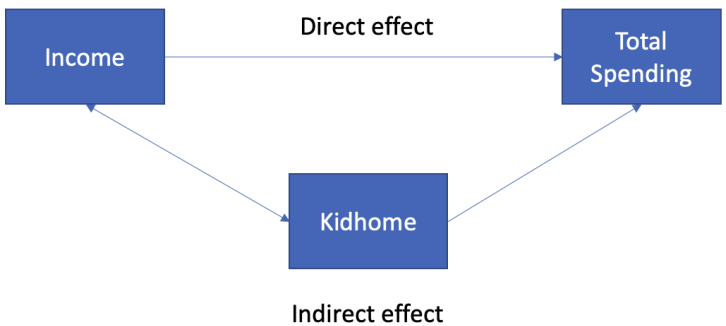
	(1)	(2)
(Intercept)	-556.823***	-299.119***
	(21.654)	(28.069)
Income	0.022***	0.019***
	(0.000)	(0.000)
Kidhome		-230.610***
		(16.945)
Num.Obs.	2000	2000
R2	0.629	0.661

$+$ 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

Using our common sense, let's think about how income can causally affect total spending:



- Causal effect
  - Direct effect keeping other variables fixed
- Total Effect
  - Direct effect + indirect effects through other variables

Causal Inference from Regression Models

- To obtain causal effects from non-experimental data, we refer to obtaining the **direct effects** of a focal X variable on the outcome variable Y.
- However, if we do not include `Kidhome` in the regression, the regression coefficient 0.022 measures the **total effects of income,** including
  - **direct effects** of income on total spending, 0.019
  - **indirect effects** of income on other intermediate variables, which in turn affect income. These intermediate variables are called **confounding variables** or **confounders**.
- Therefore, it is important to **include all other confounding variables**, which affect income and total spending at the same time, to **control for the indirect effects** via other variables, in order to tease out the clean direct effect of income on total spending.

Practical Tips for Running Regression Models for Causal Inference

1. For causal inference tasks, we need to use business senses to decide which confounding variables to control. We face the **good control and bad control problems**.[2]

> *"Some variables are bad controls and should not be included in a regression model, even when their inclusion might be expected to change the short regression coefficients. Bad controls are variables that are themselves outcome variables in the notional experiment at hand. That is, bad controls might just as well be dependent variables too. Good controls are variables that we can think of having been fixed at the time the regressor of interest was determined."*

[2] Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Practical Tips for Running Regression Models (Cont.)

2. Sometimes, control variables may be statistically insignificant, they should **NOT** be removed from the regression because they still serve the purpose of control variables.
3. A high correlation between independent variables is generally not an issue in practice. However, if some variables are mechanically correlated, then we should not put them altogether in the regression to avoid perfect collinearity problems.
4. For correct statistical inference, we should construct the correct standard errors
   - **Robust standard errors** for cross-sectional data. This is to account for the heteroskedasticity of errors. `vcov = "hetero"` in `feols()`.
   - **Clustered robust standard errors** for panel data or longitudinal with group structures. This is to account for the correlation of errors within the same group. `cluster = ~ ID` in `feols()`.

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 that we have included `Kidhome` to tease out the effect of kids, what problems do we still have that prevent 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. Therefore, strictly speaking, we can **never obtain causal effects** from **non-experimental data** by merely controlling confounding variables in a linear regression.
- Mathematically speaking, because we can never control all confounding factors, the error term is always correlated with income to some extent, violating the **exogeneity assumption** of a linear regression model $E[\epsilon|X] = 0$.

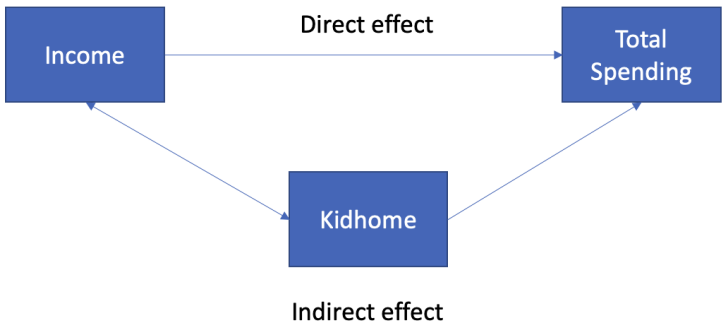
Revisit RCT: the Gold Standard of Causal Inference

- Why RCTs are the gold standard for causal inference? Why we can obtain causal inference from primary data collected from RCTs?
  - If we randomise 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`.

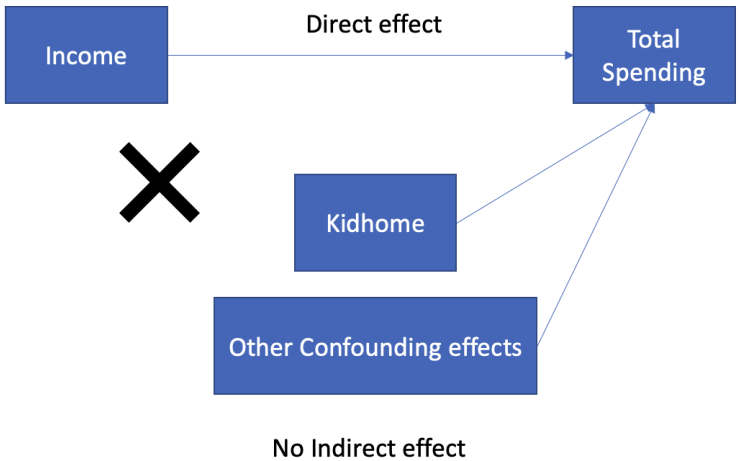
$$Spending = \beta_0 + \beta_1 Income + \epsilon$$

- In the above regression, are there still any confounding effects?

Comparison of RCT versus Secondary Data



- In non-experiment setting without randomisation, `Income` can be correlated with other unobserved confounding factors



- In experiment setting with randomisation, `Income` is randomised so should be uncorrelated with any other unobserved factors.

Section 2

Endogeneity: Omitted Variable Bias

Endogeneity

Endogeneity

Endogeneity refers to an econometric issue with OLS linear regression, in which a focal explanatory variable is correlated with the error term, such that the **Conditional Independence Assumption (CIA)** 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 omitted variable affects the dependent variable.
  - The omitted variable is correlated with the focal explanatory variable.[3]

[3] If the omitted variable is uncorrelated with X, then we do not have OVB problem, but the error term will have a larger noise and coefficients will have larger standard errors. Therefore, it's better to control these variables if possible.

Example I of OVB

- If we would like to understand the causal effect of `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?

Example II of OVB

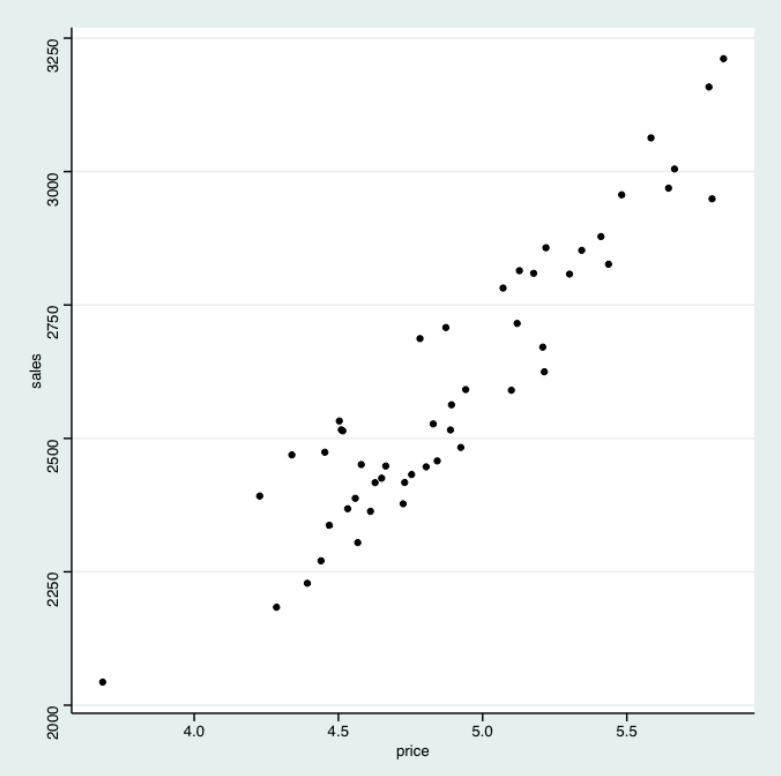
- When predicting unit sales from prices, the common practice in the industry is to regress the sales in each period on the price in each period (marketing mix modeling).

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

- However, is this regression correct?

Example II of OVB

- Very often, if we regress sales only on price, we get a positive coefficient for price.



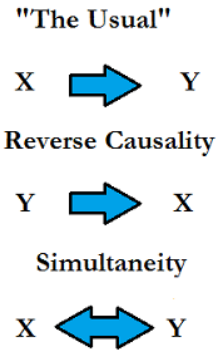
Section 3

Endogeneity: Reverse Causality

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.



Example I 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)

Example II 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 impact of network 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.

Section 4

Endogeneity: Measurement Error

Cause III: Measurement Error (Optional)

Suppose that a perfect measure of an independent variable is impossible. That is, instead of observing x^{real}, what is actually observed is $x^{observed} = x^{real} + \nu$ where ν is the measurement error with random "noise". In this case, a model given by

$$y_i = \alpha + \beta x_i^{observed} + \varepsilon_i$$

would not give us the coefficients from the regression we actually want to run

$$y_i = \alpha + \beta x_i^{real} + \varepsilon_i$$

This endogeneity issue is called **measurement error**.

When to Worry About Measurement Errors?

- This problem needs to be addressed when you expect a high measurement error in the **independent variable**, especially when using **proxy variables**. For example,
  - `grades` as a proxy for `Ability`
  - `ESGRating` as a proxy for firms' `ESGPerformance`
  - `audit fee` as a proxy for `audit quality`
- Meanwhile, if the measurement error is in the **dependent variable**, and the expectation of the measurement error is zero, then the OLS estimator is still unbiased.