Endogeneity

Data Science and Causal Inference Workshop 2025

Dr Christian Engels

START RECORDING NOW

University of St Andrews | FOUNDED 1413

3 September 2025

What I hope you will get out of today

University of St Andrews

Learning objectives

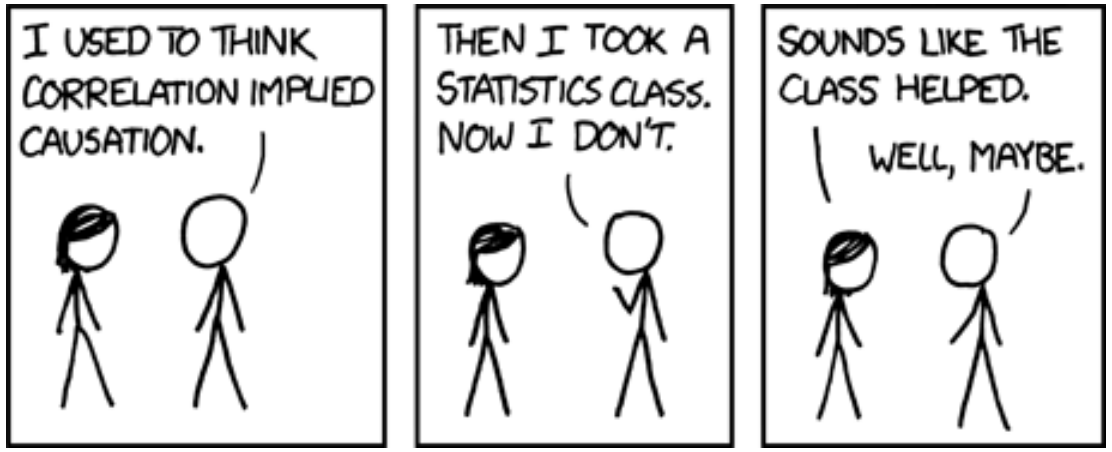
- Understand difference between econometrics for prediction and causal analysis
- Understand that endogeneity complicates deduction of causal statements
- Explain key feature of the OLS estimator regarding endogeneity
- Gain knowledge of sources of endogeneity
- Gain ability to speculate on and explain endogeneity concerns in empirical research

Outline

University of St Andrews

Endogeneity

1. Prediction ($\hat{y}_i$) vs causal effects ($\hat{\beta}$)
2. Model first and error term second
3. Introducing endogeneity
    1. Characterization
    2. Illustration
    3. Relation to Ordinary Least Squares (OLS)
4. Sources of endogeneity
    1. Omitted variable
    2. Simultaneity
    3. Selection
    4. Autocorrelation with lagged dependent variable
    5. Measurement error
5. Simulation in Stata



Studying endogeneity means understanding why correlation ≠ causation

Prediction ($\hat{y}$) vs causal effects ($\hat{\beta}$)

University of St Andrews

Simple linear model

Recall the simple linear regression equation

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

by which you seek to relate y_i to x_i where i is the unit index.

On the model side, taking (conditional) expectations leads to $E[y_i \mid x_i] = \alpha + \beta x_i$. This is the conditional expectation function (CEF).

On the estimation side, fitting the observations to data yields the estimated relationship

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

where the hat superscript indicates fitted values.

Prediction ($\hat{y}_i$)

Data Science

Prediction is concerned with providing the best guess about y_i when x_i is known. In other words, getting $\hat{y}_i$ right.

In this $\hat{y}_i$-world it does not matter whether the model you use describes reality correctly if your outcome is predicted well.

Causal effects ($\hat{\beta}$)

Academic Research

Causal analysis, on the other hand, is concerned with identifying the impact of x_i on y_i when x_i changes by a certain amount.

In this $\hat{\beta}$-world the aim is to learn about the underlying structure of the observable phenomena.

The error term follows the model

University of St Andrews

Simple linear model

Recall the simple linear regression equation

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

by which you seek to relate y_i to x_i where i is the unit index.

On the model side, taking (conditional) expectations leads to $E[y_i \mid x_i] = \alpha + \beta x_i$. This is the conditional expectation function (CEF).

On the estimation side, fitting the observations to data yields the estimated relationship

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

where the hat superscript indicates fitted values.

CEF and the role of the error term

The model error term can be expressed as

$$\varepsilon_i = y_i - E[y_i|x_i]$$

In this way you can see that the error term follows from the CEF – not the other way around.

First comes the model specification ($E[y_i|x_i]$) of the data process (y_i). From this flows the error term (ε_i): the component of the data our model does not explain.

If we do a good job, the error term has certain desirable properties. Most importantly, that it is independent of x_i: $E[x_i \varepsilon_i] = 0$.

Manifestation of endogeneity in linear models

University of St Andrews

Visual model representation

The simple linear regression model ($y_i = \alpha + \beta x_i + \varepsilon_i$) can be visualized in a causal diagram

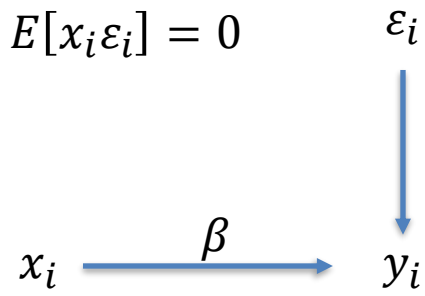
- Each component is represented
- Arrows indicate connected components
- Directions of arrows show nature of relationship

Note: simply drawing an error does not make the connection causal. Hypothesis development is important.

How endogeneity manifests

Endogeneity manifests as x_i not being independent of the error term ε_i. Simply put, in estimation $\hat{\beta}$ then absorbs this relation, leading to bias.

Desirable error term property



*Un*desirable error term property

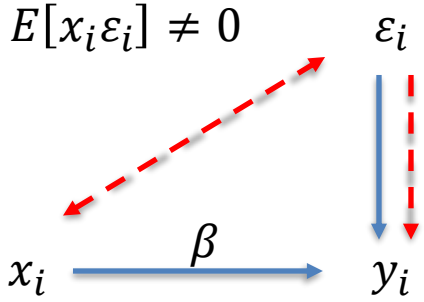
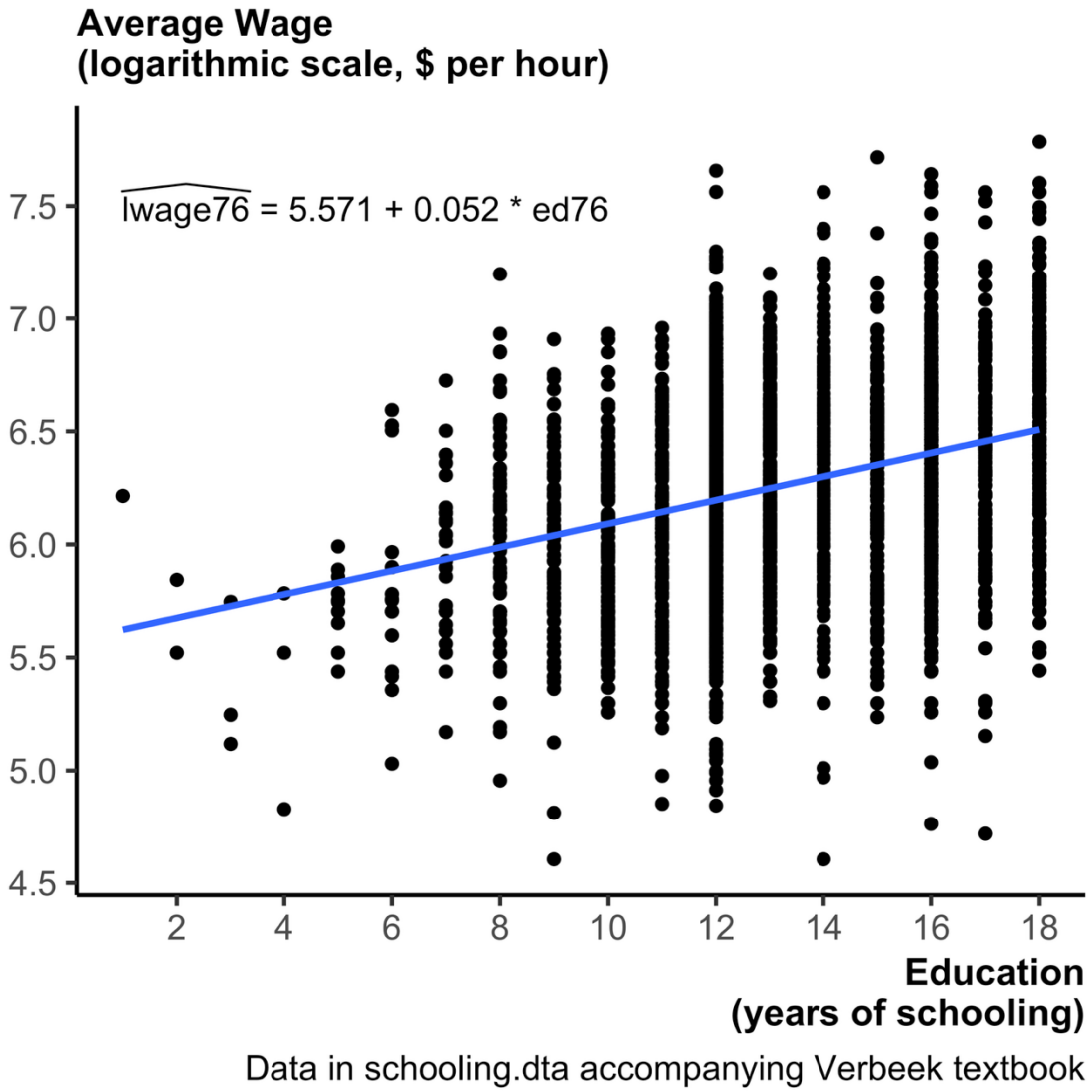


Illustration: education and income

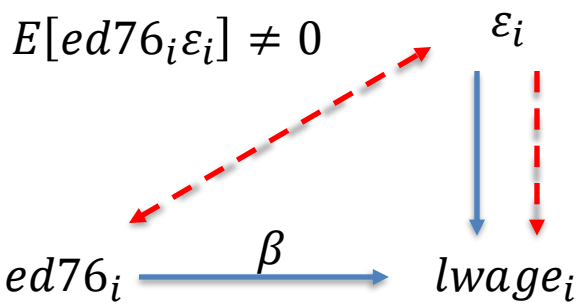


Data in schooling.dta accompanying Verbeek textbook

Schooling and hourly earnings

- Each black point is an observation on a person's education-wage combination
- The blue line describes the modeled relationship between wages (lwage76) and education (ed76)
- The coefficient of 0.052 on ed76 implies that an additional year of schooling increases hourly wages by 5.2%

Endogenous relationship



Relation to Ordinary Least Squares (OLS) estimation

University of St Andrews

Bias in the OLS estimator

Recall that the OLS estimator is unbiased if, on average, its estimates center on the respective true values.

In our simple linear model, this amounts to:

$$E[\hat{\beta}] = \beta$$

The expression derives from is the following:

$$E[\hat{\beta}] = \beta + \left(\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\sum_{i=1}^{n} E[x_i \varepsilon_i]$$

From this you can see that $E[\hat{\beta}] \neq \beta$ if $E[x_i \varepsilon_i] \neq 0$. Then the estimator is biased.

Why is this so important?

Even if it is indeed the case that $E[x_i \varepsilon_i] \neq 0$:

- Nothing prevents you from using OLS estimation and calculating estimates.
- It is up to you to ensure its conditions are met. Subject knowledge / the willingness to problematize is important.

But do not despair

It can be fun thinking about sources of endogeneity. You learn about the world.

A significant proportion of quantitative research is concerned with tackling endogeneity.

Omitted variable bias

University of St Andrews

Omitting an influential variable

Let us reconsider our linear regression model

$$y_i = \beta x_i + \varepsilon_i$$

The error term, the unexplained part of the model, now contains a variable z_i

$$\varepsilon_i = \gamma z_i + \nu_i$$

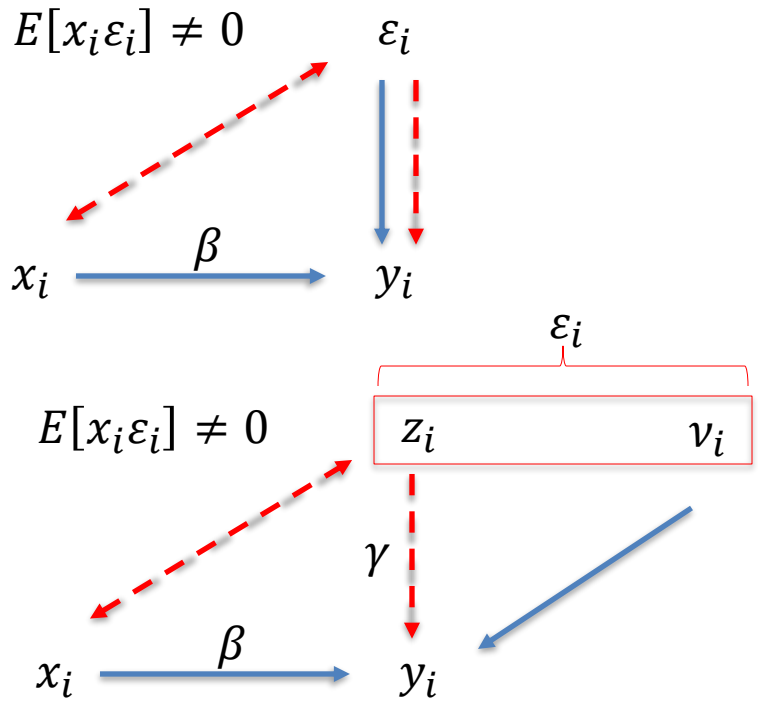
where:

- $E[x_i z_i] \neq 0$ (included and omitted are related)
- $E[z_i y_i] \neq 0$ (omitted and outcome are related)
- $E[x_i \nu_i] = E[z_i \nu_i] = 0$

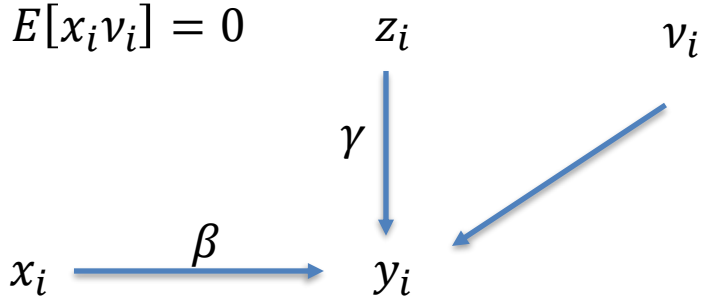
Omitting z_i from the main regression equation introduces: $E[x_i \varepsilon_i] = E[x_i(\gamma z_i + \nu_i)]$
$= \gamma E[x_i z_i] + E[x_i \nu_i] = \gamma E[x_i z_i] \neq 0$

This is omitted variable bias.

a) Omitted variable model: $y_i = \beta x_i + \varepsilon_i$



b) Correctly specified model: $y_i = \beta x_i + \gamma z_i + \nu_i$



Signing the bias

University of St Andrews

A specific case

Recall the simple model

- $y_i = \beta x_i + \varepsilon_i$
- $\varepsilon_i = \gamma z_i + \nu_i$

In this specific case it is possible to identify the sign of the bias

	$Corr(x_i, z_i) > 0$	$Corr(x_i, z_i) < 0$
$\beta > 0$	Positive bias	Negative bias
$\beta < 0$	Negative bias	Positive bias

This only holds in the case of one endogenous regressor and one omitted variable.

University of St Andrews

Noncognitive Abilities and Financial Distress: Evidence from a Representative Household Panel

Gianpaolo Parise

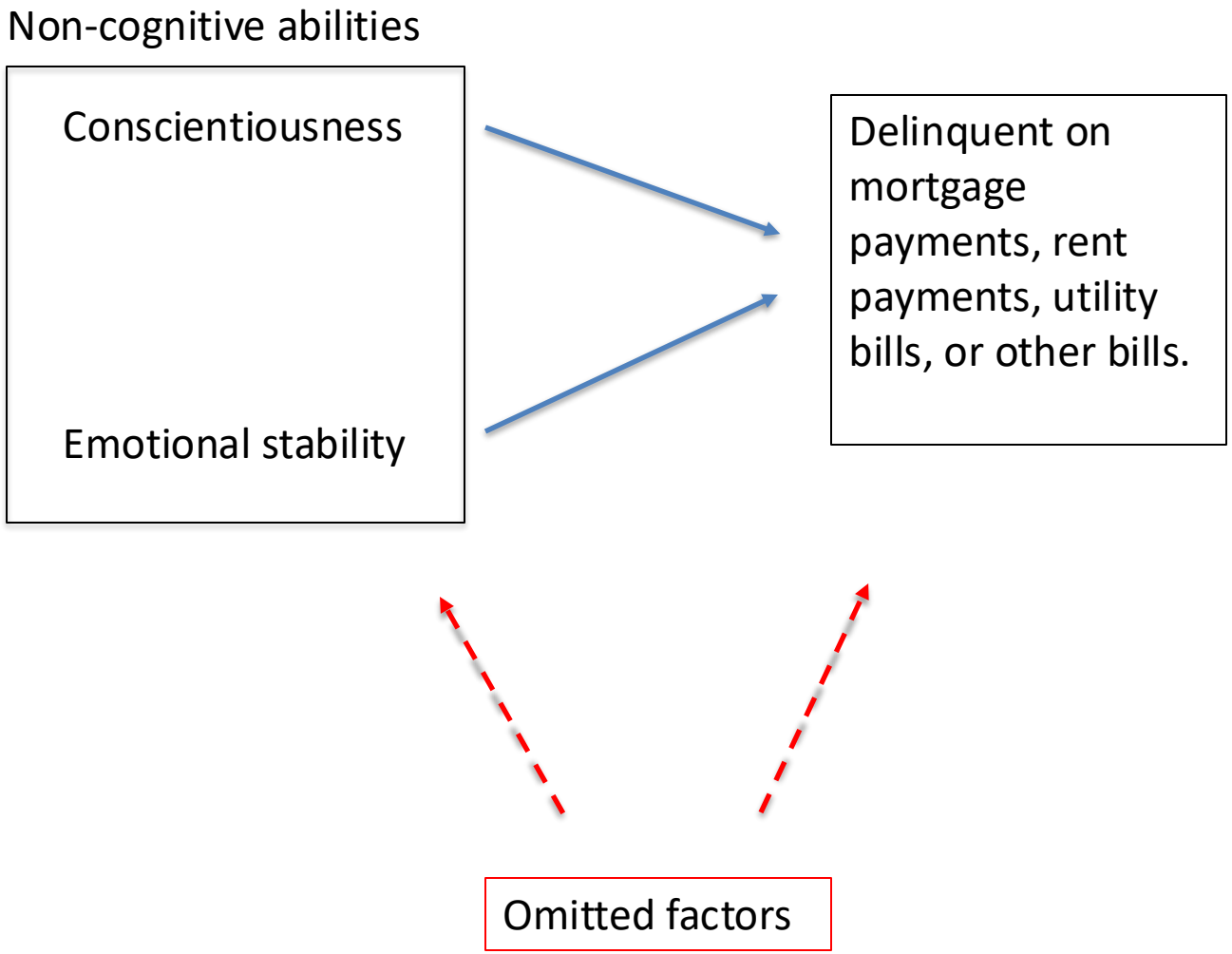
Kim Pejnenburg

Parise and Pejnenburg (2019, RFS)

University of St Andrews

"This paper provides evidence of how noncognitive abilities affect financial distress. In a representative panel of households, we find that people in the bottom quintile of noncognitive abilities are 10 times more likely to experience financial distress than those in the top quintile. We provide evidence that this relation largely arises from worse financial choices and lack of financial insight by low-ability individuals and reflects differential exposure to income shocks only to a lesser degree. We mitigate endogeneity concerns using an IV approach and an extensive set of controls. Implications for policy and finance research are discussed."

Parise and Pejnenburg (2019, RFS)



Parise and Pejnenburg (2019, RFS)

University of St Andrews

”We mitigate concerns about confounding effects by including a battery of additional controls. First, we add to our main specification several controls for the family background during childhood. These controls include proxies for the financial situation of the household during childhood, the exposure to financial distress in the household during childhood, the education level of the mother, and the education level of the father. Second, we control for […]. Third, to address concerns about confounding effects due to the external environment, […].”

Simultaneity bias

University of St Andrews

From y_i to x_i and x_i to y_i

Again, consider our simple linear model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

But now there is also an effect running from y_i to x_i:

$$x_i = \gamma + \lambda y_i + \nu_i$$

Not accounting for this introduces endogeneity:

$$E[x_i \varepsilon_i] = E[(\gamma + \lambda y_i + \nu_i)\varepsilon_i] = \lambda E[y_i \varepsilon_i] \neq 0$$

This is simultaneity or reverse causality bias.

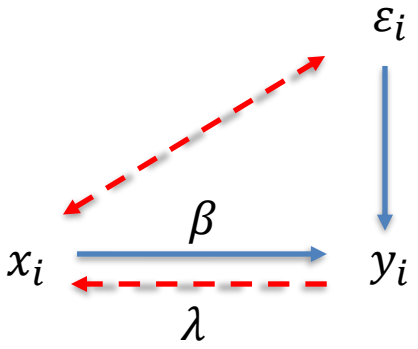
Illustration: error term propagation

A change in ε_i induces change in x_i:

$$\Delta\varepsilon_i \rightarrow \Delta y_i \rightarrow \Delta x_i$$

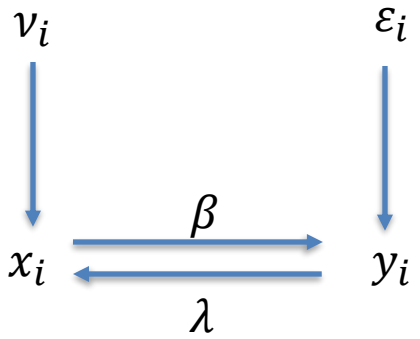
a) Misspecified model: $y_i = \beta x_i + \varepsilon_i$

$E[x_i \varepsilon_i] \neq 0$



b) Correctly specified model

$E[x_i \varepsilon_i] = 0$



University of St Andrews

Tracing out capital flows: How financially integrated banks respond to natural disasters

Kristle Romero Cortés

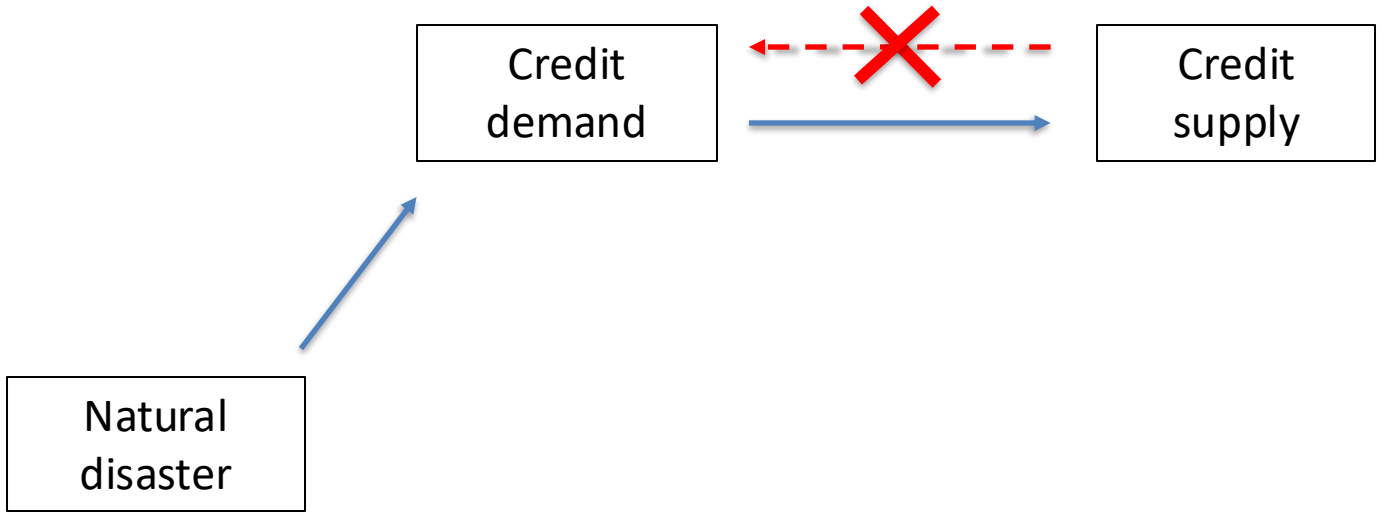
Philip E. Strahan

Cortes and Strahan (2017, JFE)

University of St Andrews

"Multi-market banks reallocate capital when local credit demand increases after natural disasters. Using property damage as an instrument for lending growth, we find credit in unaffected but connected markets declines by a little less than 50 cents per dollar of additional lending in shocked areas. However, banks shield their core markets because most of the decline comes from loans in areas where banks do not own branches. Moreover, banks increase sales of more-liquid loans and they bid up the rate on deposits in the connected markets. These actions help lessen the impact of the demand shock on credit supply."

Cortes and Strahan (2017, JFE)



Selection into sample bias

University of St Andrews

Missing data can matter

Again, consider our simple linear model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

However, we can only observe y_i if is greater than θ:

$$y_i = \begin{cases} y_i^* & if\ y_i > \theta \\ Missing\ if\ y_i \leq \theta \end{cases}$$

where y_i^* corresponds to the observed value of y_i. Thus the model we could feasibly estimate is:

$$y_i^* = \alpha + \beta x_i + \varepsilon_i$$

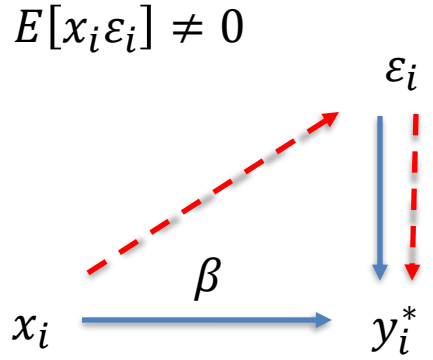
This will result in sample selection bias if x_i is indeed related to y_i $(\beta \neq 0)$.

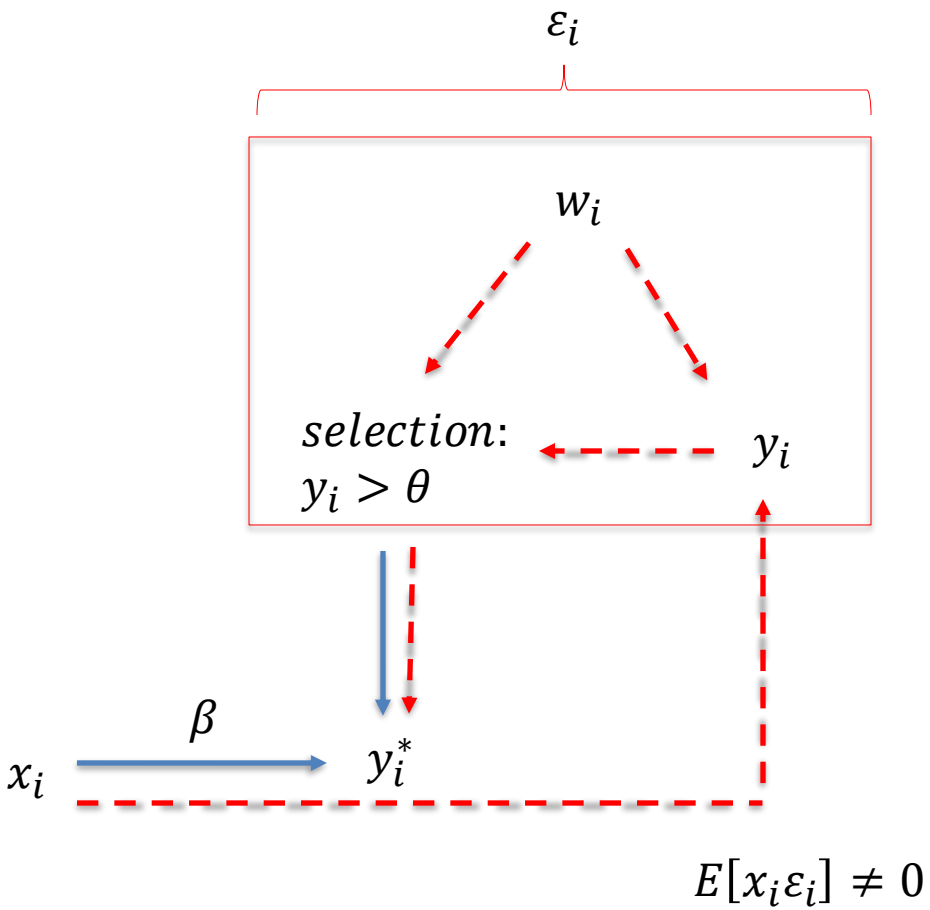
Verbal argument that $E[x_i \varepsilon_i] \neq 0$

- ε_i includes the process that determines whether y_i is observable (the greater the value of y_i, the higher the likelihood that it is observable).
- If x_i correlates with y_i, it will correlate with the likelihood that y_i is observable, the process included in ε_i , and therefore with ε_i itself.

Selection into sample

- $y_i = \alpha + \beta x_i + \varepsilon_i$
- $y_i = \begin{cases} y_i^* & if\ y_i > \theta \\ Missing\ if\ y_i \leq \theta \end{cases}$





Nguyen, Hagendorff and Eshragi (2018, RFS)

University of St Andrews

Does a CEO's Cultural Heritage Affect Performance under Competitive Pressure?

Duc Duy Nguyen

Jens Hagendorff

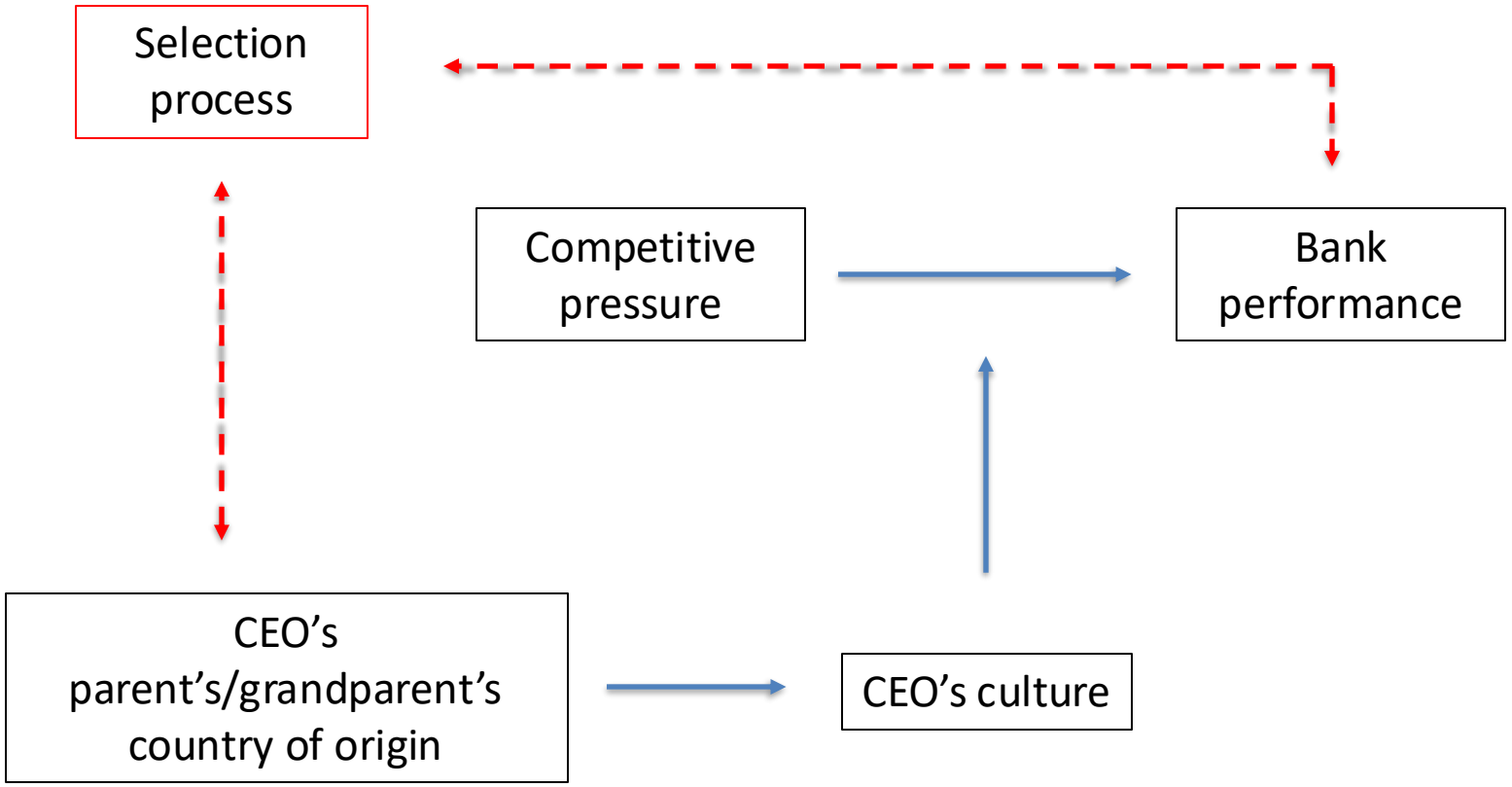
Arman Eshragi

Nguyen, Hagendorff and Eshragi (2018, RFS)

University of St Andrews

"We exploit variation in the cultural heritage across U.S. CEOs who are the children or grandchildren of immigrants to demonstrate that the cultural origins of CEOs matter for corporate outcomes. Following shocks to industry competition, firms led by CEOs who are second- or third-generation immigrants are associated with a 6.2% higher profitability compared with the average firm. This effect weakens over successive immigrant generations and cannot be detected for top executives apart from the CEO. Additional analysis attributes this effect to various cultural values that prevail in a CEO's ancestral country of origin."

Nguyen, Hagendorff and Eshragi (2018, RFS)



Selection into treatment bias

University of St Andrews

Receiving treatment

Consider the linear model with the dummy variable $D_i \in \{0,1\}$:

$$y_i = \alpha + \beta D_i + \varepsilon_i$$

The value 1 can be interpreted as receiving treatment and 0 as non-treatment.

Further say that D_i is determined by an additional factor w_i:

$$D_i = \mathbb{1}(\gamma + \lambda w_i + \nu_i > 0)$$

where $\mathbb{1}(\cdot)$ is the indicator function (More on this later.)

This is also known as endogenous selection into treatment. It will cause a bias in β if w_i is also related to y_i.

Example A

- D = stock market participation
- y = retirement wealth
- w = financial literacy

Example B

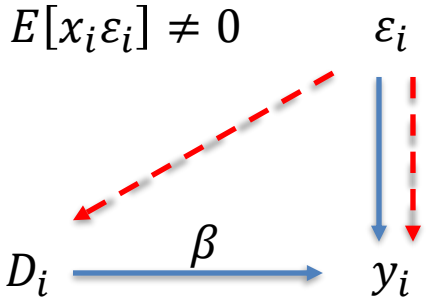
- D = financial education program
- y = financial distress (debt, arrears)
- w = psychological distress

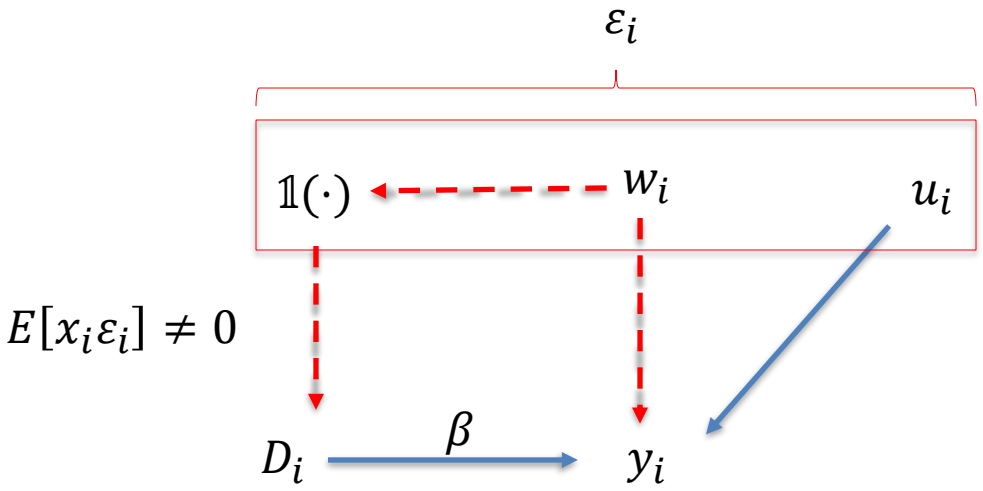
Example C

- D = MSc degree
- y = Wage growth
- w = Ability

Selection into treatment bias

- $D_i \in \{0,1\}$
- $y_i = \alpha + \beta D_i + \varepsilon_i$
- $D_i = \mathbb{1}(\gamma + \lambda w_i + \nu_i > 0)$





Autocorrelation with lagged dependent variable

University of St Andrews

The time series case

Consider the time series model

$$y_t = \alpha + \beta x_t + \lambda y_{t-1} + \varepsilon_t$$

where t denotes time. Further assume autocorrelated errors:

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$$

where $-1 < \rho < 1$.

The model can also be expressed as

$$y_{t-1} = \alpha + \beta x_{t-1} + \lambda y_{t-2} + \varepsilon_{t-1}$$

Therefore it can directly be seen that endogeneity arises because of the autocorrelated errors

$$E[y_{t-1}\varepsilon_t] = E[(\beta x_{t-1} + \lambda y_{t-2} + \varepsilon_{t-1})\varepsilon_t] \neq 0$$

Because $E[\varepsilon_{t-1}\varepsilon_t] \neq 0$.

Relation to panel context

You may come across two estimators, which tackle endogeneity in a panel setup:

- Arrelano-Bond estimator, also known as Difference GMM
- Arellano–Bover/Blundell–Bond estimator, also known as System GMM

In a panel setting, autocorrelation is highly likely (unobserved, time-invariant factors). This can cause problems.

Further information

ROODMAN, D. 2009. How to do xtabond2: An introduction to difference and system GMM in Stata. *The Stata Journal,* 9, 86-136.

Measurement error

University of St Andrews

In independent variable, x_i

Suppose the model is again given by

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

But we observe

$$x_i^* = x_i + u_i$$

where u_i is the measurement error.

Substituting this back into the model equation gives

$$y_i = \alpha + \beta(x_i^* - u_i) + \varepsilon_i$$

This simplifies to

$$y_i = \alpha + \beta x_i^* + \nu_i$$

where $\nu_i = \varepsilon_i - \beta u_i$.

It follows that $E[x_i^* \nu_i] \neq 0$ because x_i^* is related to u_i.

In dependent variable, y_i

In this case, bias only arises if the measurement error is related to x_i.

Specifically:

$$y_i = y_i^* + u_i = \alpha + \beta x_i + \varepsilon_i$$

So that

$$y_i^* = \alpha + \beta x_i + \nu_i$$

where $\nu_i = \varepsilon_i - u_i$. It is easy to see that $E[x_i \nu_i] \neq 0$ only if x_i is related to u_i.

University of St Andrews

Noncognitive Abilities and Financial Delinquency: The Role of Self-Efficacy in Avoiding Financial Distress

"An important caveat to our analysis is that the relationships we document between self-efficacy and financial choices may not be causal in nature. Although we control for observable differences and account for parental support in a siblings fixed effect analysis, omitted variables bias is certainly possible. For example, some limitations of the NLSY Child and Youth survey are that it fails to measure respondents' financial wealth and expectations for income growth, and its measure of time preferences is quite coarse."

Further reading

University of St Andrews

Indicative sections in textbooks

Verbeek (4e) Sections 5.1, 5.2

Wooldridge (6e) Sections 3-3b, 3-3c, 9-4, 9-5, 15-1, 16-2, 17-5a

Further resources

HILL, A. D., JOHNSON, S. G., GRECO, L. M., O'BOYLE, E. H. & WALTER, S. L. 2021. Endogeneity: A Review and Agenda for the Methodology-Practice Divide Affecting Micro and Macro Research. *Journal of Management*.*

NGUYEN, D. D., HAGENDORFF, J. & ESHRAGHI, A. 2018. Does a CEO's Cultural Heritage Affect Performance under Competitive Pressure? *The Review of Financial Studies,* 31, 97-141.