Selection bias and Heckman two-stage estimation

Prof. Dr. David Bendig

Jonathan Hoke

contact@statisticslab.org

April 2022

Disclaimer

- This presentation and its content represent a **work in progress** and are **still subject to peer review – your feedback is very welcome here**
- The basic assumption and examples are given in the context of **secondary panel data**
- Feel free to use our materials – if you decide to do so, **we ask you to cite our work**

Do not hesitate to contact us if you have any questions or feedback

contact@statisticslab.org | www.statisticslab.org

Contents

Introduction

- In his seminal paper, James J. Heckman (1979) pioneered a method that helps **identify and mitigate sample-induced endogeneity**: the Heckman two-stage estimation
- With the **rising relevance of selection bias** in the prevalent research and the **attention scientific journals pay** to the choices and implementation of econometric techniques, we notice an **increasing number of methodological errors** in applying the Heckman two-stage estimation
- Scholars and practitioners **may be challenged to stay in touch with this development**: For these reasons we developed a **practical and comprehensible step-by-step guide** with this presentation and the respective discussion paper
- These resources **aim to enable applying the Heckman two-stage estimation to any research model** in entrepreneurship, innovation, and other research streams

Sources: Certo et al., 2016; Heckman, 1979; Ullah et al., 2021; Wolfolds & Siegel, 2019

Problem

- **Selection bias** occurs when a sample is not randomly generated and, thus, **does not represent the population**
- An example is **corporate venture capital (CVC)** investments, where researchers can only observe a firm's final investment decision. However, the strategic decision of CVC may be based on characteristics that researchers cannot observe
- **Consequence**: "The problem of selection bias [...] arises when a rule other than simple random sampling is used to sample the underlying population that is the object of interest. The distorted representation of a true population as a consequence of a sampling rule is the essence of the selection problem." (Heckman, 2018, p. 12131)

Sources: Bushway et al., 2007; Greene, 2018; Heckman, 2018; Hill et al., 2021

Methodology

- The Heckman two-stage estimation supports **identifying and mitigating a potential selection bias**
- This technique **consists of two consecutively applied stages** that separate the selection process from the primary relationship of interest
  - In the **first stage**, the selection process of the underlying relationship is estimated
  - The **second stage** analyzes the primary relationship of interest
  - The connection between the two stages is a **unique selection parameter** induced from the first stage and inserted in the second-stage regression
  - The selection parameter **captures unobservable characteristics** found in the primary regression’s error term that lead to endogenous covariates

Sources: Certo et al. 2016; Clougherty et al., 2015; Hamilton & Nickerson, 2003; Heckman, 1979; Hill et al., 2021

Application in Stata (1|5)

First stage: Selection equation (1|2)

- The Selection Equation analyzes whether **observations from the population appear in the selected sample**
- A probit regression is performed, where a **binary selection variable** is chosen as the dependent variable
- In addition, **matching instruments** must be selected that **meet two requirements**: The instruments must
  1. **Influence the binary selection variable of the second stage**
  2. **Not influence the dependent variable of the second stage**

```
xtprobit Selection_Variable Independent_Variables Controls Instruments
```

Example for panel data: Depending on the data structure, a pooled probit regression may be useful

Sources: Clougherty et al., 2015; Hill et al., 2021; Wolfolds & Siegel, 2019

Application in Stata (2|5)

First stage: Selection equation (2|2)

- The Inverse Mills Ratio (IMR) correction variable is then determined to **retrieve the IMR as a selection parameter** and captures the significant unobserved characteristics that affect the underlying relationship
- The IMR can be calculated by **dividing the normal density function (PDF) by the normal cumulative distribution**

```
predict xb, xb
generate PDF = normalden(xb)
generate CDF = normal(xb)
generate IMR = PDF / CDF
```


Sources: Certo et al., 2016; Clougherty et al., 2015; Wolfolds & Siegel, 2019

Application in Stata (3|5)

Second stage: Outcome equation (1|3)

- The outcome equation is estimated **using a linear regression model (OLS)**
- The **Inverse Mills Ratio (IMR) correction variable** is used in the outcome equation
- The **standard least squares estimator may be downward biased** → One possible way to correct this biased variance may be to **bootstrap the standard errors** of the first and second stages

Sources: Clougherty et al., 2015; Hamilton & Nickerson, 2003; Heckman, 1979, Hill et al., 2003

Application in Stata (4|5)

Second stage: Outcome equation (2|3)

```
program heckman_2_stage
    xtprobit Selection Variable Independent_Variables Controls Instruments
    predict xb, xb
    gen PDF = normalden(xb)
    gen CDF = normal(xb)
    gen IMR = PDF / CDF
    xtreg Dependent_Variable Independent_Variables Controls IMR
    drop xb PDF CDF IMR
end

bootstrap: heckman_2_stage
```

Application in Stata (5|5)

Second stage: Outcome equation (3|3)

- The level of significance and sign of the IMR's beta coefficient suggests the magnitude of the correlation between the error terms of the selection equation and the outcome equation → represents the level of endogeneity present in the research model

➢ A **significantly positive** (**negative**) **beta coefficient** suggests that unobserved factors **positively** (**negatively**) **affect the estimated relationship**

➢ Note that an **insignificant Inverse Mills Ratio** at the second-stage level **does not entirely rule out a selection bias**; The power of the Heckman two-stage estimation of determining a selection bias is affected by the strength of the exclusion restriction and the sample size

Sources: Bushway et al., 2007; Certo et al., 2016; Hill et al., 2021

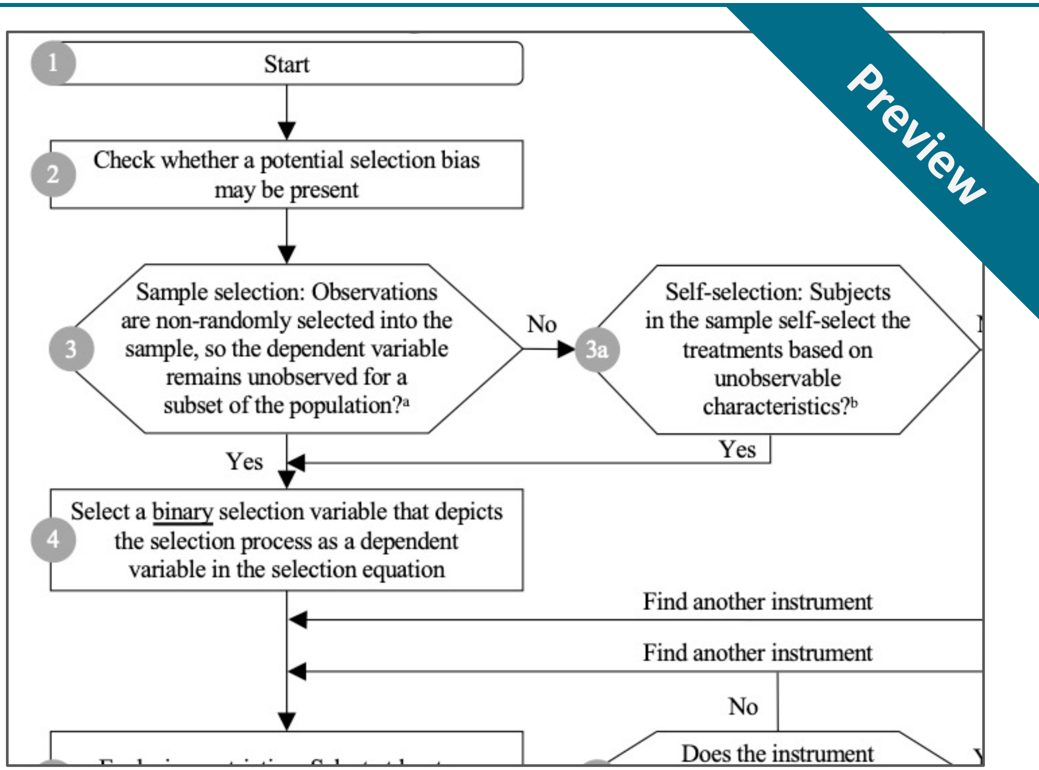
Critical factors and methodological pitfalls to avoid

- The regression analyses of the first and second stages should contain the **same independent and control variables** – do not forget time fixed-effects for panel data
- The **regression types used in the two stages are essential** for the Heckman two-stage estimation since the error terms of both stages should follow a bivariate normal distribution
  - The **first stage** must be a **probit regression**
  - The **second stage** should be **either a probit or an OLS regression**, and "since the derivation of the Heckman two-step method relies on the normality of errors, we are hesitant to suggest that the use of other estimation techniques is appropriate" (Wolfolds & Siegel, 2019, p. 452)
- The Heckman two-stage estimation is not suitable for count data as it **requires a full parametric specificity** → regression error specification test (RESET) by Ramsey (1969) to indicate whether the normality assumption can be found in the errors terms

Sources: Arregle et al., 2012; Bushway et al., 2007; Ramsey, 1969; Wolfolds & Siegel, 2019

Preview of our Heckman flowchart

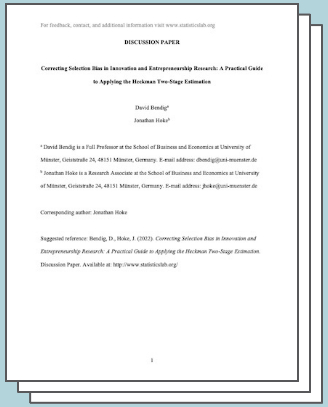
- We developed a **graphical representation** of the Heckman two-stage estimation in a flowchart
- The flowchart helps **better understand the theoretical assumptions and application** of the technique by **showing the process steps** and **ensuring that no step is omitted**
- Each step in the flowchart is **numbered to indicate the flow's direction**



➢ **Access a virtual version of the flowchart or download the PDF file on www.statisticslab.org**

Recommended literature

- Our discussion paper addresses the theoretical assumptions and methodological fundamentals of Heckman’s technique in more detail
  Download from www.statisticslab.org


For feedback, contact, and additional information visit www.statisticslab.org

DISCUSSION PAPER

Correcting Selection Bias in Innovation and Entrepreneurship Research: A Practical Guide to Applying the Heckman Two-Stage Estimation

David Bendig[a]
Jonathan Hoke[b]

[a] David Bendig is a Full Professor at the School of Business and Economics at University of Münster, Geiststraße 24, 48151 Münster, Germany. E-mail address: dbendig@uni-muenster.de
[b] Jonathan Hoke is a Research Associate at the School of Business and Economics at University of Münster, Geiststraße 24, 48151 Münster, Germany. E-mail address: jhoke@uni-muenster.de

Corresponding author: Jonathan Hoke

Suggested reference: Bendig, D., Hoke, J. (2022). *Correcting Selection Bias in Innovation and Entrepreneurship Research: A Practical Guide to Applying the Heckman Two-Stage Estimation*. Discussion Paper. Available at: http://www.statisticslab.org/

1

Access here

Further literature recommendations addressing the Heckman two-stage estimation:

- Certo, S. T., Busenbark, J. R., Woo, H.-S., & Semadeni, M. 2016. Sample selection bias and Heckman models in strategic management research. Strategic Management Journal, 37: 2639-2657.
- Bushway, S., Johnson, B. D., & Slocum, L. A. 2007. Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. Journal of Quantitative Criminology, 23: 151-178.
- Heckman, J. J. 1979. Sample selection bias as a specification error. Econometrica, 47: 153-161.
- Wolfolds, S. E., & Siegel, J. 2019. Misaccounting for endogeneity: The peril of relying on the Heckman two-step method without a valid instrument. Strategic Management Journal, 40: 432-462.

References

Arregle, J. L., Naldi, L., Nordqvist, M., & Hitt, M. A. (2012). Internationalization of family-controlled firms: A study of the effects of external involvement in governance. Entrepreneurship Theory and Practice, 36(6), 1115-1143. https://doi.org/10.1111/j.1540-6520.2012.00541.x

Bushway, S., Johnson, B. D., & Slocum, L. A. (2007). Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. Journal of Quantitative Criminology, 23(2), 151-178. https://doi.org/10.1007/s10940-007-9024-4

Cameron, A. C., & Trivedi, P. K. (2010). Microeconometrics using Stata (Vol. 2). College Station, TX: Stata press.

Certo, S. T., Busenbark, J. R., Woo, H.-S., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. Strategic Management Journal, 37(13), 2639-2657. https://doi.org/10.1002/smj.2475

Clougherty, J. A., Duso, T., & Muck, J. (2015). Correcting for self-selection based endogeneity in management research. Organizational Research Methods, 19(2), 286-347. https://doi.org/10.1177/1094428115619013

Greene, W. H. (2018). Econometric analysis (8th ed.). Pearson.

Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. Strategic Organization, 1(1), 51-78. https://doi.org/10.1177/1476127003001001218

Heckman, J. J. (1979). Sample selection bias as a specification error. Econometrica, 47(1), 153-161. https://doi.org/10.2307/1912352

Heckman, J. J. (2018). Selection bias and self-selection. In S. N. Durlauf & L. E. Blume (Eds.), The new palgrave dictionary of economics (pp. 12130-12147). Palgrave Macmillan. https://doi.org/10.1057/978-1-349-95121-5_1762-2

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, 47(1), 105-143. https://doi.org/10.1177/0149206320960533

Hill, R. C., Adkins, L. C., & Bender, K. A. (2003). Test statistics and critical values in selectivity models. In T. B. Fomby & R. Carter Hill (Eds.), Maximum likelihood estimation of misspecified models: Twenty years later (17th ed., pp. 75-105). Emerald Group Publishing. https://doi.org/10.1016/S0731-9053(03)17004-1

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society: Series B (Methodological), 31(2), 350-371. https://doi.org/10.1111/j.2517-6161.1969.tb00796.x

Ullah, S., Zaefarian, G., & Ullah, F. (2021). How to use instrumental variables in addressing endogeneity? A step-by-step procedure for non-specialists. Industrial Marketing Management, 96, A1-A6. https://doi.org/10.1016/j.indmarman.2020.03.006

Wolfolds, S. E., & Siegel, J. (2019). Misaccounting for endogeneity: The peril of relying on the Heckman two-step method without a valid instrument. Strategic Management Journal, 40(3), 432-462. https://doi.org/10.1002/smj.2995

You can find more information on our website

contact@statisticslab.org

www.statisticslab.org