GSE 522

Varieties of Selection Bias: Insights from James Heckman

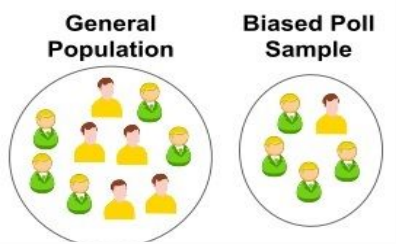
Anmol Lakhotia

01

Introduction to Selection Bias

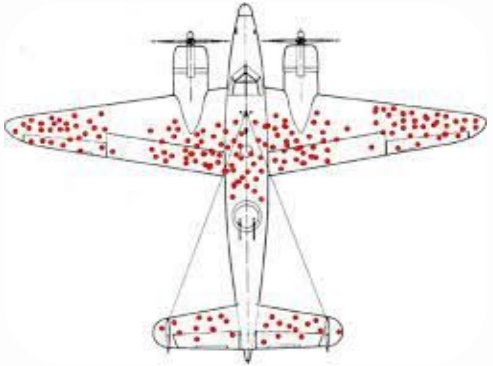
Understanding Selection Bias

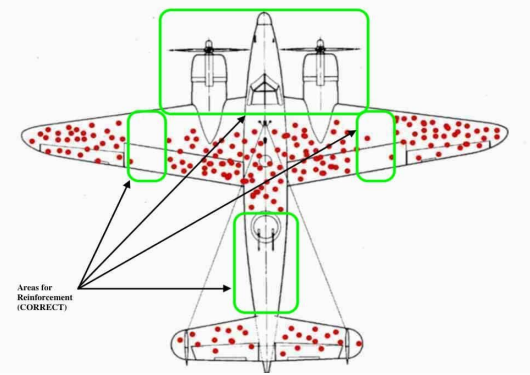
Selection bias occurs when the sample used in research does not accurately represent the population being studied, leading to skewed or biased results.



$$\hat{\beta} \neq \beta$$

In econometric analysis, this bias is particularly problematic, as it can significantly impact the estimation of parameters.





Importance of Dr. Heckman's Work

- In this paper, James Heckman's offering groundbreaking insights that have shaped subsequent research in the field.
- Dr. Heckman uses the **union wage differentials study** by H. Gregg Lewis as an example to explore the issue of selection bias.
- He created a comprehensive framework for addressing/mitigating bias, including innovative methods for identification and estimation with selection bias.
- He discussed uses involving both truncated and censored data.

Selection Bias in Unionization and Wages

Identification

- Workers **choose** to join unions based on factors that also affect their **wages:** skills, motivation, or access to union jobs.
- This **mixing** of **union choice** and **wage** causes selection bias.

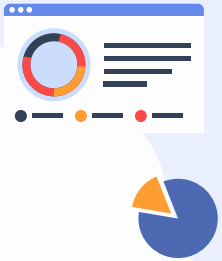
Implication

- It is difficult to isolate the true **effect of unionism on wages.**
- Failing to account for selection bias can lead to **misleading estimates** of the economic parameters of interest: union wage premium.

Innovation

- Dr. Heckman uses the union wage study as an example to explore the selection bias.
- He evaluates **nonparametric** and **functional form** approaches.
- He then argues for simpler methods: **IV** regression.

Splitting the Wages by Unionization Choice



1. **Union Wages (Y1)**: $Y1=X1*\beta1+U1$ where $E(U1)=0$

2. **Nonunion Wages (Y0):** $Y0=X0\beta0+U$ where $E(U0)=0$

3. **Choice Equation (D):** $I = Z_{gamma} + V$, where $E(V)=0$.
   I>0 implies choice of the union sector (D=1), otherwise D=0.

4. **Observed Wage (Y):** $Y=Y_1D+Y_0(1-D)=(X_1\beta_1)D+(X_0\beta_0)(1-D)+DU_1+(1-D)U_0$

02
Parameters of Interest and Identification

The Parameters of Interest

Experimental Treatment Average ($\alpha_1 - \alpha_0$): The effect of moving a nonunion worker to the union sector.

Roy model assuming $\beta_1 = \beta_0$.

Impact on the Unionized E(Y1–Y0 | D=1,Z): = $(a_1-a_0) + E(U_1-U_0 | D =1, Z)$

This estimate represents the gain for a unionized person moving from the nonunionized to the unionized sector, accounting for individual attributes X and Z. This is also known as average treatment effect among the treated, ATT.

Assumptions for Identification and its Purpose

1. Independence

The error terms (U_0, U_1, V) must be independent of the covariates (X,Z).

2. Continuous Distributions

The error terms are assumed to have continuous distributions.

3. Zero Conditional Mean:

$$E[U_0 \mid X, Z] = 0$$
$$E[U_1 \mid X, Z] = 0$$
$$E[V \mid X, Z] = 0$$

Purpose 1:

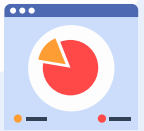
Ensures that the selection process can be separated from the outcome processes.

Purpose 2:

Allows using the observed data to infer the unobserved and identify parameters: error distributions and covariate effects.

Purpose 3:

Increases confidence due to normalization and centering.



03

Results

Utilizing Distributional Information and Identifying Estimates: “Two-Step”

- Examines variation in the distribution of observed outcomes based on changes in covariates and selection states
- Deduces the underlying structure of the error terms through probabilities. (logit/probit)
- Then decomposes the observed outcomes into components attributable to the treatment effect (union wage premium) and those due to selection bias.
- By modeling the selection process and its likelihood on the observed outcomes, Heckman is able to correct for the bias
- Heckman's approach is able to estimate the Union-wage counterfactual without making parametric assumptions.

Model Advancements

- Heckman discusses other innovative methods being used and researched especially with censoring: **kernel methods**, **series estimators**, and **density estimations**
- Traditional models often rely on linear assumptions, **semiparametric** estimation accommodate more **complex relationships and distributional forms.**
- These semiparametric methods follow Heckman's framework and **identification theorem** to efficiently estimate parameters.
- He also explains that if $E(U_1-U_0 \mid D=1, z, x_c)=0$ than IV regression will result in the correct **Experimental Treatment Average.**

Heckman's Contribution and Results

- Introduced robust nonparametric and semiparametric identification and estimation techniques so we are not bound by restrictive parametric assumptions
- Developed a theoretical framework to create consistency in addressing selection bias.
- Advocated for further research to deal with complex models of human behavior and market dynamics.