IVB-paper

Manoel Galdino

novembro 08, 2024

Introduction

In this document, we derive the expression for the **Included Variable Bias (IVB)** caused by conditioning on a collider variable Z in a regression model. This bias is analogous to the well-known omitted variable bias (OVB) but arises when a collider is incorrectly included in the model.

Model Setup

We consider the following models:

Short Regression

The true relationship between the outcome y and the treatment D is given by:

$$y = \beta_0 + \beta_1 D + e$$

- β_0: Intercept term.
- β_1: True effect of D on y.
- e: Error term with $\mathbb{E}[e|D] = 0$.

Wrong Model (Long Regression with Collider Bias)

Including the collider variable Z leads to the following (misspecified) model:

$$y = \beta_0^\star + \beta_1^\star D + \beta_2^\star Z + e^\star$$

- $\beta_0^\star, \beta_1^\star, \beta_2^\star$: Estimated coefficients from the long regression.
- $e^\star$: Error term where $\mathbb{E}[e^\star|D] \neq 0$.

Collider Relationship

The collider variable Z is influenced by both D and y:

$$Z = \gamma_0 + \gamma_1 D + \gamma_2 y + u$$

- $\gamma_0, \gamma_1, \gamma_2$: Parameters of the collider equation.
- u: Error term with $\mathbb{E}[u|D] = 0$ and $\mathbb{E}[u|y] = 0$.

Step 1: Express Bias in Terms of Covariances

The bias in the estimated coefficient $\beta_1^\star$ arises due to the inclusion of the collider Z:

We know that $\beta_1 = \frac{\text{Cov}(D,Y)}{\text{Var}(D)}$.

We also know that $\beta_1^\star = \frac{\text{Cov}(D,Y) - \beta_2^\star \text{Cov}(D,Z)}{\text{Var}(D)}$.

If we define Bias $= \beta_1^\star - \beta_1$, then:

$$\text{Bias} = \beta_1^\star - \beta_1 = -\beta_2^\star \frac{\text{Cov}(D,Z)}{\text{Var}(D)}$$

- $\beta_2^\star$: Coefficient of Z from the long regression.
- $\text{Cov}(D,Z)$ and $\text{Var}(D)$: Covariance and variance terms involving observable variables.

Step 2: Avoiding γ_1 and γ_2

We substitute the expression for y into the collider equation to eliminate γ_1 and γ_2.

Substitute y into the Collider Equation

From the true model:

$$y = \beta_0 + \beta_1 D + e$$

Substitute y into Z:

$$\begin{aligned} Z &= \gamma_0 + \gamma_1 D + \gamma_2 y + u \\ &= \gamma_0 + \gamma_1 D + \gamma_2(\beta_0 + \beta_1 D + e) + u \\ &= (\gamma_0 + \gamma_2\beta_0) + (\gamma_1 + \gamma_2\beta_1)D + \gamma_2 e + u \end{aligned}$$

Define New Terms

Let:

$$\begin{aligned} \phi_0 &= \gamma_0 + \gamma_2\beta_0 \\ \phi_1 &= \gamma_1 + \gamma_2\beta_1 \\ \varepsilon &= \gamma_2 e + u \end{aligned}$$

So the collider equation becomes:

$$Z = \phi_0 + \phi_1 D + \varepsilon$$

- ϕ_1: Coefficient from regressing Z on D.
- ε: Composite error term.

Step 3: Express Covariance Using Observable Quantities

The covariance between D and Z is:

$$\text{Cov}(D, Z) = \phi_1 \, \text{Var}(D)$$

Thus, the bias becomes:

$$\text{Bias} = -\beta_2^\star \frac{\text{Cov}(D, Z)}{\text{Var}(D)} = -\beta_2^\star \phi_1$$

- Both $\beta_2^\star$ and ϕ_1 are estimable from data.

Final Expression of Included Variable Bias

The included variable bias is:

$$\text{Bias} = -\beta_2^\star \phi_1$$

Where:

- $\beta_2^\star$: Estimated coefficient of Z from the long regression.
- ϕ_1: Estimated coefficient from regressing Z on D.

Practical Implications

- **Estimating $\beta_2^\star$:** Run the long regression of y on D and Z.
- **Estimating ϕ_1:** Regress Z on D to obtain ϕ_1.
- **Calculating Bias:** Use the formula $\text{Bias} = -\beta_2^\star \phi_1$ to quantify the bias.

Conclusion

By expressing the included variable bias in terms of observable quantities ($\beta_2^\star$ and ϕ_1), we can quantify the bias introduced by incorrectly including a collider variable in the regression model. This derivation parallels the omitted variable bias formula but addresses the specific context of collider bias.

References

- **Econometrics Textbooks** for foundational knowledge on regression models and biases.
- **Causal Inference Literature** for understanding collider bias and its implications.