

Lectures in Quantitative International Relations Specification in International Relations

Kevin A. Clarke
University of Rochester

Dublin City University, May 2007

Lectures

- 1 Introduction and Introduction to Maximum Likelihood Estimation
- 2 Some Common MLE Models Used in International Relations
- 3 Comparative Theory Testing
- 4 **Choosing a Specification**

Overview of Lecture 4

- 1 The Logic of Control Variables
 - Familiar Results
 - Fear and Control Variables

Overview of Lecture 4

- 1 The Logic of Control Variables
 - Familiar Results
 - Fear and Control Variables
- 2 The Mathematical Argument
 - A Simple Example
 - Simulation Results
 - Efficiency

Overview of Lecture 4

- 1 The Logic of Control Variables
 - Familiar Results
 - Fear and Control Variables
- 2 The Mathematical Argument
 - A Simple Example
 - Simulation Results
 - Efficiency
- 3 Extension: GLMs
 - Results

Overview of Lecture 4

- 1 The Logic of Control Variables
 - Familiar Results
 - Fear and Control Variables
- 2 The Mathematical Argument
 - A Simple Example
 - Simulation Results
 - Efficiency
- 3 Extension: GLMs
 - Results
- 4 A New Logic
 - Implications
 - Possible Remedies

The Argument

Fear of omitted variable bias (O.V.B.) is the leading cause of control variables.

The Argument

Fear of omitted variable bias (O.V.B.) is the leading cause of control variables.

- The familiar O.V.B. result is one of those lessons we both learn and remember from graduate school.

The Argument

Fear of omitted variable bias (O.V.B.) is the leading cause of control variables.

- The familiar O.V.B. result is one of those lessons we both learn and remember from graduate school.
- Unfortunately, we are never in the position regarding O.V.B. that is discussed in most textbooks.

The Argument

Fear of omitted variable bias (O.V.B.) is the leading cause of control variables.

- The familiar O.V.B. result is one of those lessons we both learn and remember from graduate school.
- Unfortunately, we are never in the position regarding O.V.B. that is discussed in most textbooks.
- In any working situation, we know little, if anything, about the effects that control variables have on a coefficient of interest.

The Argument

Fear of omitted variable bias (O.V.B.) is the leading cause of control variables.

- The familiar O.V.B. result is one of those lessons we both learn and remember from graduate school.
- Unfortunately, we are never in the position regarding O.V.B. that is discussed in most textbooks.
- In any working situation, we know little, if anything, about the effects that control variables have on a coefficient of interest.
- We should replace the logic of control variables with a logic of research design.

Familiar Result 1

$$\begin{array}{ll} \text{DGP:} & \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \text{Assumed:} & \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^*, \quad \boldsymbol{\epsilon}^* = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \end{array}$$

Familiar Result 1

$$\begin{aligned} \text{DGP:} \quad & \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \text{Assumed:} \quad & \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^*, \quad \boldsymbol{\epsilon}^* = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \end{aligned}$$

$$E[\hat{\boldsymbol{\beta}}_1] = \boldsymbol{\beta}_1 + \mathbf{P}\boldsymbol{\beta}_2, \quad \text{where } \mathbf{P} = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2$$

Familiar Result 1

$$\begin{aligned} \text{DGP:} \quad & \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \text{Assumed:} \quad & \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^*, \quad \boldsymbol{\epsilon}^* = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \end{aligned}$$

$$E[\hat{\boldsymbol{\beta}}_1] = \boldsymbol{\beta}_1 + \mathbf{P}\boldsymbol{\beta}_2, \quad \text{where } \mathbf{P} = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2$$

Unless the omitted variables are orthogonal to the included variables, and thus $\mathbf{X}'_1\mathbf{X}_2$ is a zero matrix, or $\boldsymbol{\beta}_2$ is a zero vector, our estimator, $\hat{\boldsymbol{\beta}}_1$, is biased.

Familiar Result 2

$$\begin{array}{ll} \text{DGP:} & \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \text{Assumed:} & \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}^* \end{array}$$

Familiar Result 2

$$\begin{aligned} \text{DGP:} & \quad \mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon, \quad \epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \text{Assumed:} & \quad \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon^* \end{aligned}$$

$$E[\hat{\beta}_1] = \beta_1, \text{ but not minimum variance.}$$

Familiar Result 2

$$\begin{array}{ll} \text{DGP:} & \mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon, \quad \epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \text{Assumed:} & \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon^* \end{array}$$

$$E[\hat{\beta}_1] = \beta_1, \text{ but not minimum variance.}$$

The consequences of including irrelevant variables are “generally less serious than those pertaining to the exclusion of relevant variables” (Johnston and DiNardo 1997).

Possibly Less Familiar Result 3

Unrestricted: $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$

Restricted: $\mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon$

Possibly Less Familiar Result 3

$$\text{Unrestricted: } \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$$

$$\text{Restricted: } \mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon$$

$$\text{Var}(\hat{\beta}_1^U) - \text{Var}(\hat{\beta}_1^R) = \sigma^2(\mathbf{X}'_1\mathbf{M}_2\mathbf{X}_1)^{-1} - \sigma^2(\mathbf{X}'_1\mathbf{X}_1)^{-1},$$

where $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2$. The difference is positive semidefinite.

Possibly Less Familiar Result 3

$$\text{MSE}_U - \text{MSE}_R = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2 \{ \sigma^2 [\mathbf{X}'_2 \mathbf{X}_2 - \mathbf{X}'_2 \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2]^{-1} - \beta_2 \beta_2' \} \mathbf{X}'_2 \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1}.$$

is positive semidefinite if $\Sigma_{\hat{\beta}_2} - \beta_2 \beta_2'$ is positive semidefinite, where $\Sigma_{\hat{\beta}_2}$ is the covariance matrix for $\hat{\beta}_2$.

Possibly Less Familiar Result 3

$$\text{MSE}_U - \text{MSE}_R = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2 \{ \sigma^2 [\mathbf{X}'_2 \mathbf{X}_2 - \mathbf{X}'_2 \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2]^{-1} - \beta_2 \beta_2' \} \mathbf{X}'_2 \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1}.$$

is positive semidefinite if $\Sigma_{\hat{\beta}_2} - \beta_2 \beta_2'$ is positive semidefinite, where $\Sigma_{\hat{\beta}_2}$ is the covariance matrix for $\hat{\beta}_2$.

⇒ We might prefer the restricted model *even when* the unrestricted model is the DGP.

The Standard O.V.B. Lesson Doesn't Usually Apply

The Standard O.V.B. Lesson Doesn't Usually Apply

- We are never in the position of choosing between including the final relevant variable (or set of variables) or not.

The Standard O.V.B. Lesson Doesn't Usually Apply

- We are never in the position of choosing between including the final relevant variable (or set of variables) or not.
- We are faced with choosing between including an additional relevant variable (or set of variables) out of a larger set of relevant omitted variables that we either do not know or cannot measure.

The Standard O.V.B. Lesson Doesn't Usually Apply

- We are never in the position of choosing between including the final relevant variable (or set of variables) or not.
- We are faced with choosing between including an additional relevant variable (or set of variables) out of a larger set of relevant omitted variables that we either do not know or cannot measure.

The important question to ask is what is the effect of adding to a statistical specification some, but not all, of these relevant omitted variables.

Does Fear of O.V.B. Really Cause Control Variables?

Does Fear of O.V.B. Really Cause Control Variables?

KKV admonish us to “systematically look for omitted control variables” and note that if “relevant variables are omitted, our ability to estimate causal inferences correctly is limited.”

Does Fear of O.V.B. Really Cause Control Variables?

KKV admonish us to “systematically look for omitted control variables” and note that if “relevant variables are omitted, our ability to estimate causal inferences correctly is limited.”

Ansolabehere, Gerber, and Snyder (2002) include variables such as poverty, unemployment rates, median income, percentages of the population that are school-aged, black, and elderly, and population change “[t]o minimize the danger of omitted variables bias....”

Does Fear of O.V.B. Really Cause Control Variables?

KKV admonish us to “systematically look for omitted control variables” and note that if “relevant variables are omitted, our ability to estimate causal inferences correctly is limited.”

Ansolabehere, Gerber, and Snyder (2002) include variables such as poverty, unemployment rates, median income, percentages of the population that are school-aged, black, and elderly, and population change “[t]o minimize the danger of omitted variables bias....”

Bailey, Kamoie, and Maltzman (2005) write that “[t]o minimize the possibility that omitted variable bias affects our results, we control for a number of factors that may affect court voting.”

Does Fear of O.V.B. Really Cause Control Variables?

Does Fear of O.V.B. Really Cause Control Variables?

Hegre, Ellingsen, Gates, and Gleditsch (2001) identify a number of control variables “whose omission might bias the results for the regime change variable.”

Does Fear of O.V.B. Really Cause Control Variables?

Hegre, Ellingsen, Gates, and Gleditsch (2001) identify a number of control variables “whose omission might bias the results for the regime change variable.”

Rudolph and Evans (2005) control for a number of individual-level factors “[t]o reduce the risk of omitted variable bias.”

Does Fear of O.V.B. Really Cause Control Variables?

Hegre, Ellingsen, Gates, and Gleditsch (2001) identify a number of control variables “whose omission might bias the results for the regime change variable.”

Rudolph and Evans (2005) control for a number of individual-level factors “[t]o reduce the risk of omitted variable bias.”

Krause (2003) argues that leaving some variables out in an alternative specification “generated an omitted variable bias problem.”

Do Reviewers Play a Role By Any Chance?

“As critics we use omitted variables as the first line of attack, and as authors we know that controlling for more variables helps protect us from potential criticism; from this perspective, the more variables in X_2 the better.”

King, Honaker, Joseph, and Scheve (2001)

The Logic of Control Variables

The Logic of Control Variables

- Control variables have real effects so their absence, when correlated with included variables, may well cause bias and inconsistency in the variable of interest;

The Logic of Control Variables

- Control variables have real effects so their absence, when correlated with included variables, may well cause bias and inconsistency in the variable of interest;
- Control variables have real effects so their inclusion does not cause inefficiency, *ceteris paribus*;

The Logic of Control Variables

- Control variables have real effects so their absence, when correlated with included variables, may well cause bias and inconsistency in the variable of interest;
- Control variables have real effects so their inclusion does not cause inefficiency, *ceteris paribus*;
- The bias caused by omitted variables is an aggregation of the bias caused by each individual omitted variable;

The Logic of Control Variables

- Control variables have real effects so their absence, when correlated with included variables, may well cause bias and inconsistency in the variable of interest;
- Control variables have real effects so their inclusion does not cause inefficiency, *ceteris paribus*;
- The bias caused by omitted variables is an aggregation of the bias caused by each individual omitted variable;
- As the inclusion of all relevant, correlated variables is impossible, we should include as many as possible in order to reduce the bias.

The Mathematical Argument for Control Variables

DGP:
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

Model 1:
$$Y_i = \beta_{01} + \beta_{11} X_{i1} + \epsilon_{i1}$$

Model 2:
$$Y_i = \beta_{02} + \beta_{12} X_{i1} + \beta_{22} X_{i2} + \epsilon_{i2}$$

The Mathematical Argument for Control Variables

$$\text{DGP:} \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

$$\text{Model 1:} \quad Y_i = \beta_{01} + \beta_{11} X_{i1} + \epsilon_{i1}$$

$$\text{Model 2:} \quad Y_i = \beta_{02} + \beta_{12} X_{i1} + \beta_{22} X_{i2} + \epsilon_{i2}$$

$$\begin{aligned} |E[\hat{\beta}_{11}] - \beta_1| &> |E[\hat{\beta}_{12}] - \beta_1| \\ |b(\hat{\beta}_{11}, \beta_1)| &> |b(\hat{\beta}_{12}, \beta_1)| \end{aligned}$$

A Simple Example: The Bias in Model 1

$$\begin{array}{ll} \text{DGP:} & Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \\ \text{Model 1:} & Y_i = \beta_{01} + \beta_{11} X_{i1} + \epsilon_{i1} \end{array}$$

A Simple Example: The Bias in Model 1

$$\begin{aligned} \text{DGP:} \quad & Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \\ \text{Model 1:} \quad & Y_i = \beta_{01} + \beta_{11} X_{i1} + \epsilon_{i1} \end{aligned}$$

$$E[\hat{\beta}_{11}] = \beta_1 + \beta_2 \left(r_{12} \sqrt{\frac{V_2}{V_1}} \right) + \beta_3 \left(r_{13} \sqrt{\frac{V_3}{V_1}} \right)$$

A Simple Example: The Bias in Model 1

DGP: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$
Model 1: $Y_i = \beta_{01} + \beta_{11} X_{i1} + \epsilon_{i1}$

$$E[\hat{\beta}_{11}] = \beta_1 + \beta_2 \left(r_{12} \sqrt{\frac{V_2}{V_1}} \right) + \beta_3 \left(r_{13} \sqrt{\frac{V_3}{V_1}} \right)$$

$$b(\hat{\beta}_{11}, \beta_1) = (E[\hat{\beta}_{11}] - \beta_1) = \beta_2 \left(r_{12} \sqrt{\frac{V_2}{V_1}} \right) + \beta_3 \left(r_{13} \sqrt{\frac{V_3}{V_1}} \right)$$

A Simple Example: The Bias in Model 2

$$\begin{array}{ll} \text{DGP:} & Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \\ \text{Model 2:} & Y_i = \beta_{02} + \beta_{12} X_{i1} + \beta_{22} X_{i2} + \epsilon_{i2} \end{array}$$

A Simple Example: The Bias in Model 2

$$\begin{aligned} \text{DGP:} \quad & Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \\ \text{Model 2:} \quad & Y_i = \beta_{02} + \beta_{12} X_{i1} + \beta_{22} X_{i2} + \epsilon_{i2} \end{aligned}$$

$$E[\hat{\beta}_{12}] = \beta_1 + \beta_3 \left(\frac{(r_{13} - r_{12}r_{23})}{1 - r_{12}^2} \sqrt{\frac{V_3}{V_1}} \right)$$

A Simple Example: The Bias in Model 2

$$\begin{aligned} \text{DGP:} \quad & Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \\ \text{Model 2:} \quad & Y_i = \beta_{02} + \beta_{12} X_{i1} + \beta_{22} X_{i2} + \epsilon_{i2} \end{aligned}$$

$$E[\hat{\beta}_{12}] = \beta_1 + \beta_3 \left(\frac{(r_{13} - r_{12}r_{23})}{1 - r_{12}^2} \sqrt{\frac{V_3}{V_1}} \right)$$

$$b(\hat{\beta}_{12}, \beta_1) = (E[\hat{\beta}_{12}] - \beta_1) = \beta_3 \left(\frac{(r_{13} - r_{12}r_{23})}{1 - r_{12}^2} \sqrt{\frac{V_3}{V_1}} \right)$$

What We Want to Know

Under what conditions does the following inequality hold?

What We Want to Know

Under what conditions does the following inequality hold?

$$\left| \beta_2 \left(r_{12} \sqrt{\frac{V_2}{V_1}} \right) + \beta_3 \left(r_{13} \sqrt{\frac{V_3}{V_1}} \right) \right| > \left| \beta_3 \left(\frac{(r_{13} - r_{12}r_{23})}{1 - r_{12}^2} \sqrt{\frac{V_3}{V_1}} \right) \right|$$

We Could Solve the Inequality, but...

...it is messy and unilluminating.

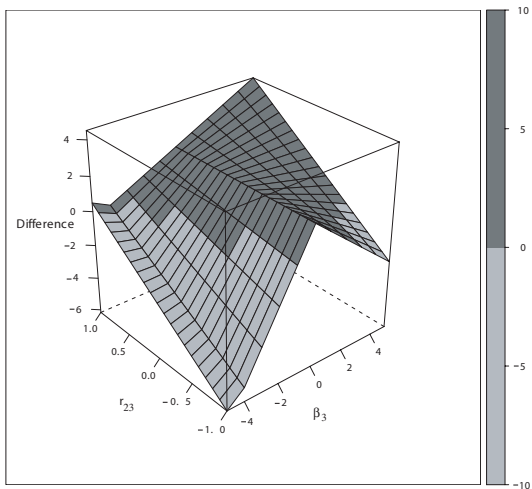
We Could Solve the Inequality, but...

...it is messy and unilluminating.

So we'll set some values:

- $V_1 = V_2 = V_3 = 1,$
- $\beta_1 = \beta_2 = 4,$
- $r_{12} = r_{13} = 0.5,$
- and $\beta_3 \in \{-5, \dots, 5\}$ and $r_{23} \in \{-1, \dots, 1\}.$

Results 1



Efficiency and the Phantom Menace

$$\text{DGP: } \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}_3\boldsymbol{\beta}_3 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

$$\text{Model 1: } \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_{11} + \boldsymbol{\epsilon}_1,$$

$$\text{Model 2: } \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_{12} + \mathbf{X}_2\boldsymbol{\beta}_{22} + \boldsymbol{\epsilon}_2$$

$$\text{where } \boldsymbol{\epsilon}_1 = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}_3\boldsymbol{\beta}_3 + \boldsymbol{\epsilon} \text{ and } \boldsymbol{\epsilon}_2 = \mathbf{X}_3\boldsymbol{\beta}_3 + \boldsymbol{\epsilon}$$

Efficiency and the Phantom Menace

$$\begin{aligned}\text{Var}(\hat{\beta}_{11})^{-1} - \text{Var}(\hat{\beta}_{12})^{-1} &= \frac{1}{\sigma^2} \mathbf{X}'_1 \mathbf{X}_1 - \frac{1}{\sigma^2} \mathbf{X}'_1 \mathbf{M}_2 \mathbf{X}_1 \\ &= \frac{1}{\sigma^2} \mathbf{X}'_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{X}_1,\end{aligned}$$

Efficiency and the Phantom Menace

$$\begin{aligned}\text{Var}(\hat{\beta}_{11})^{-1} - \text{Var}(\hat{\beta}_{12})^{-1} &= \frac{1}{\sigma^2} \mathbf{X}'_1 \mathbf{X}_1 - \frac{1}{\sigma^2} \mathbf{X}'_1 \mathbf{M}_2 \mathbf{X}_1 \\ &= \frac{1}{\sigma^2} \mathbf{X}'_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{X}_1,\end{aligned}$$

Therefore, $\text{Var}(\hat{\beta}_{12}) - \text{Var}(\hat{\beta}_{11})$ is positive semidefinite.

$$\begin{aligned}\text{MSE}(\hat{\beta}_{11}) &\geq \text{MSE}(\hat{\beta}_{12}) ? \\ \text{Var}(\hat{\beta}_{11}) + \text{Bias}(\hat{\beta}_{11})^2 &\geq \text{Var}(\hat{\beta}_{12}) + \text{Bias}(\hat{\beta}_{12})^2 ?\end{aligned}$$

Extension: GLMs

$$\text{DGP: } Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad Y_i = \begin{cases} 1, & \text{if } Y_i^* > 0 \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Model: } Y_i^* = \beta_0 + \beta_1 X_{i1} + \epsilon_i^*$$

$$\text{let } X_{i2} = \gamma_0 + \gamma_1 X_{i1} + \nu_i$$

$$\hat{\beta}_1 = \frac{c}{\sqrt{\sigma^2 + \beta_2^2 \sigma_\nu^2}} (\beta_1 + \gamma_1 \beta_2)$$

Extension: GLMs

$$\begin{aligned} \text{DGP:} \quad & Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \\ \text{Model 1:} \quad & Y_i^* = \beta_{01} + \beta_{11} X_{i1} + \epsilon_{i1}, \end{aligned}$$

$$\begin{aligned} \text{let} \quad & X_{i2} = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i3} + \nu_i \\ & X_{i3} = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \eta_i \end{aligned}$$

$$\hat{\beta}_{11} = \frac{c[\beta_1 - \beta_1 \gamma_2 \delta_2 + \beta_2(\gamma_1 + \gamma_2 \delta_1) + \beta_3(\delta_1 + \delta_2 \gamma_1)]}{\sqrt{\sigma^2 + \sigma^2 \gamma_2^2 \delta_2^2 + \sigma_\eta^2(\beta_2^2 \gamma_2^2 + \beta_3^2) + \sigma_\nu^2(\beta_2^2 + \beta_3^2 \delta_2^2)}}$$

Extension: GLMs

$$\text{DGP: } Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i,$$

$$\text{Model 2: } Y_i^* = \beta_{02} + \beta_{12} X_{i1} + \beta_{22} X_{i2} + \epsilon_{i2},$$

$$\text{let } X_{i3} = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \eta_i$$

$$\hat{\beta}_{12} = \frac{c}{\sqrt{\sigma^2 + \beta_3^2 \sigma_\eta^2}} (\beta_1 + \beta_3 \delta_1).$$

What We Want To Know

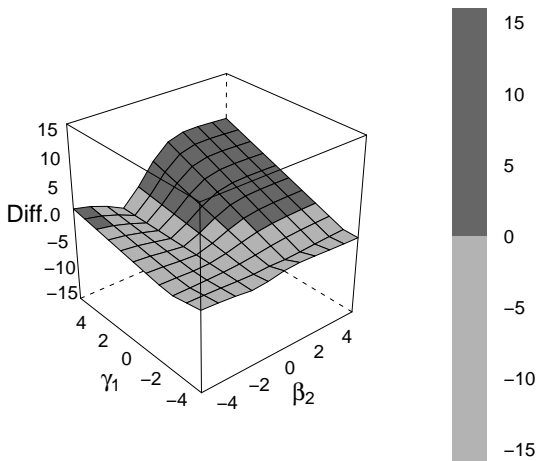
Under what conditions does the following inequality hold?

$$b(\hat{\beta}_{11}, \beta_1) > b(\hat{\beta}_{12}, \beta_1)$$

- $\beta_1, \beta_2, \beta_3, \gamma_1 \in \{-5, \dots, 5\}$,
- $\delta_1 = \delta_2 = \gamma_2 = 2$,
- $\sigma^2 = \sigma_\nu^2 = \sigma_\eta^2 = 1$,

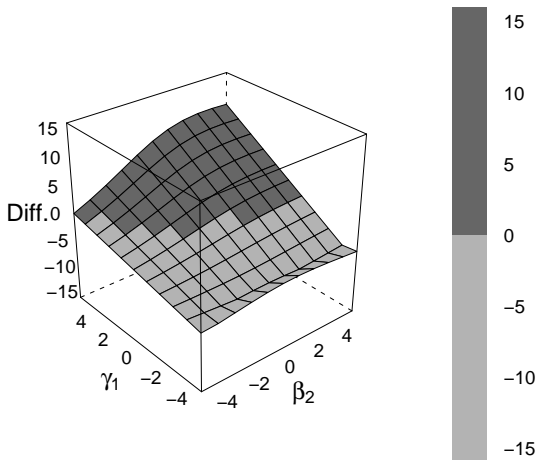
Results 2: $\beta_1 = -1, \beta_3 = 1$

$$\beta_1 = -1 \quad \beta_3 = 1$$

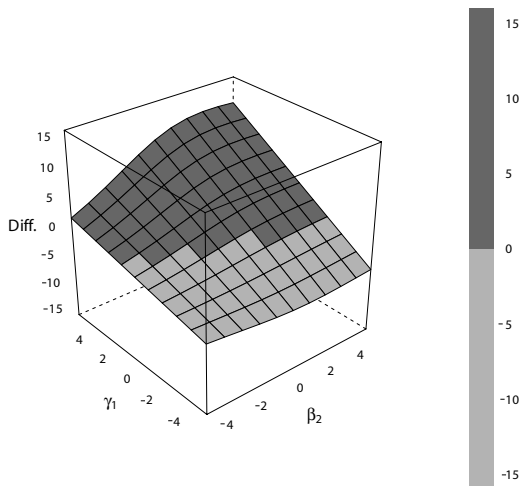


Results 2: $\beta_1 = -3, \beta_3 = 5$

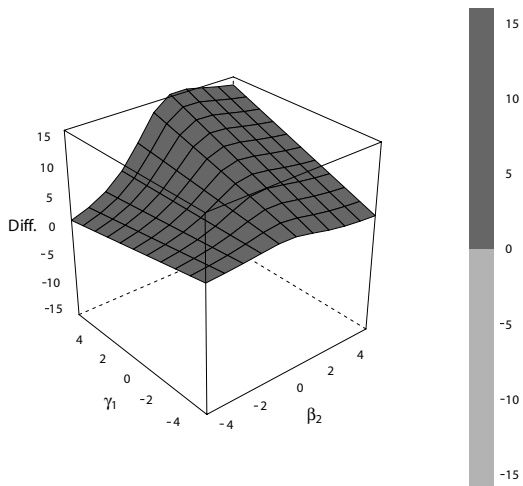
$$\beta_1 = -3 \quad \beta_3 = 5$$



Results 2: $\beta_1 = -5, \beta_3 = 5$



Results 2: $\beta_1 = -5, \beta_3 = 1$



Implications: Only You Can Prevent Control Variables

We cannot know the effect on the bias of a coefficient of interest of including an additional control variable, unless we know the complete and true specification (even if we cannot measure all the variables).

Implications: Only You Can Prevent Control Variables

We cannot know the effect on the bias of a coefficient of interest of including an additional control variable, unless we know the complete and true specification (even if we cannot measure all the variables).

Once the connection between omitted variable bias and control variables is gone, the main justification for using control variables is gone.

Supporting Documents....

Supporting Documents....

- Small amounts of measurement error in control variables “are magnified as more variables are added to the equation in an attempt to control for other possible sources of bias.” We may “kill the patient in our attempts to cure what may have been a rather minor disease originally” (Griliches 1977). Similar points are made by Maddala (1977), Welch (1975), and Achen (2005).

Supporting Documents....

- Small amounts of measurement error in control variables “are magnified as more variables are added to the equation in an attempt to control for other possible sources of bias.” We may “kill the patient in our attempts to cure what may have been a rather minor disease originally” (Griliches 1977). Similar points are made by Maddala (1977), Welch (1975), and Achen (2005).
- The kind of careful data analysis required to justify model specifications and assess fit is simply too hard with more than three variables (Achen 2002).

Supporting Documents....

- Small amounts of measurement error in control variables “are magnified as more variables are added to the equation in an attempt to control for other possible sources of bias.” We may “kill the patient in our attempts to cure what may have been a rather minor disease originally” (Griliches 1977). Similar points are made by Maddala (1977), Welch (1975), and Achen (2005).
- The kind of careful data analysis required to justify model specifications and assess fit is simply too hard with more than three variables (Achen 2002).
- “Usually, a regression equation based on a few variables will be more accurate...” (Breiman 1992 and any article by David Freedman).

What Can Be Done

All those things that Hanushek and Jackson wrote about back in 1977:

What Can Be Done

All those things that Hanushek and Jackson wrote about back in 1977:

- Use theory,

What Can Be Done

All those things that Hanushek and Jackson wrote about back in 1977:

- Use theory,
- Find natural experiments,

What Can Be Done

All those things that Hanushek and Jackson wrote about back in 1977:

- Use theory,
- Find natural experiments,
- Employ careful sample stratification.

What Can Be Done

All those things that Hanushek and Jackson wrote about back in 1977:

- Use theory,
- Find natural experiments,
- Employ careful sample stratification.

We might add explicit comparison of alternative theories and the choice of research hypothesis (perform narrow, controlled tests of broad theories (Rosenbaum 1999)).

What Can Be Done

All those things that Hanushek and Jackson wrote about back in 1977:

- Use theory,
- Find natural experiments,
- Employ careful sample stratification.

We might add explicit comparison of alternative theories and the choice of research hypothesis (perform narrow, controlled tests of broad theories (Rosenbaum 1999)).

The MLE case is more opaque....

The Logic of Research Design

The Logic of Research Design

- It is impossible to include all the relevant variables in a regression equation;

The Logic of Research Design

- It is impossible to include all the relevant variables in a regression equation;
- Omitted variable bias is therefore unavoidable;

The Logic of Research Design

- It is impossible to include all the relevant variables in a regression equation;
- Omitted variable bias is therefore unavoidable;
- The inclusion of a subset of relevant control variables may not ameliorate, and may increase, the bias caused by omitted variables;

The Logic of Research Design

- It is impossible to include all the relevant variables in a regression equation;
- Omitted variable bias is therefore unavoidable;
- The inclusion of a subset of relevant control variables may not ameliorate, and may increase, the bias caused by omitted variables;
- The inclusion of a subset of relevant control variables may also cause additional biases through measurement error;

The Logic of Research Design

- It is impossible to include all the relevant variables in a regression equation;
- Omitted variable bias is therefore unavoidable;
- The inclusion of a subset of relevant control variables may not ameliorate, and may increase, the bias caused by omitted variables;
- The inclusion of a subset of relevant control variables may also cause additional biases through measurement error;
- Experimental control can be achieved through careful research design.

Example: The Democratic Peace (what else?)

Example: The Democratic Peace (what else?)

Maoz and Russett (1992, 1993) limit their temporal domain to the Cold War period.

Example: The Democratic Peace (what else?)

Maoz and Russett (1992, 1993) limit their temporal domain to the Cold War period.

- No need to include a variable to control for the number of states in the international system.

Example: The Democratic Peace (what else?)

Maoz and Russett (1992, 1993) limit their temporal domain to the Cold War period.

- No need to include a variable to control for the number of states in the international system.
- No need to include a variable to control for the changing meaning of democracy or the strength of democratic norms.

Example: The Democratic Peace (what else?)

Maoz and Russett (1992, 1993) limit their temporal domain to the Cold War period.

- No need to include a variable to control for the number of states in the international system.
- No need to include a variable to control for the changing meaning of democracy or the strength of democratic norms.

While not definitive, their results are more convincing than many studies that begin at 1816 and include a raft of control variables.

The End Is Near....

The End Is Near....

- The effect of including a subset of omitted variables as controls in a regression...depends.

The End Is Near....

- The effect of including a subset of omitted variables as controls in a regression...depends.
- Knowing whether the controls help or hurt requires knowing much more than we typically do in practice.

The End Is Near....

- The effect of including a subset of omitted variables as controls in a regression...depends.
- Knowing whether the controls help or hurt requires knowing much more than we typically do in practice.
- We need an approach to achieving convincing experimental control that has fewer debilitating side effects than control variables...such as research design.

The End Is Near....

- The effect of including a subset of omitted variables as controls in a regression...depends.
- Knowing whether the controls help or hurt requires knowing much more than we typically do in practice.
- We need an approach to achieving convincing experimental control that has fewer debilitating side effects than control variables...such as research design.
- Convincing evidence, even if it is not definitive, is the foundation of a compelling science.