

Specification Errors

The General Model

consider

$$Y_t = \beta_0 + \beta_1 X_{1t} + \dots + \beta_k X_{kt} + U_t$$

The equation may be misspecified

it may

- omit some relevant variables
- or
- include some irrelevant variables

Omission of a Relevant Variable

1 - $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + U_t$ (correct model)

2 - $Y_t = \beta_0 + \beta_1 X_{1t} + w_t$ (misspecified model)

Consider:

A researcher estimates the misspecified model (2). From this model

$$E(w_t) = E(\beta_2 X_{2t} + U_t) = E(\beta_2 X_{2t}) \neq 0 \text{ iff } X_{2t} \text{ is a relevant variable}$$

$$\text{var}(w_t) = E(w_t - E(w_t))^2 = E(\beta_2 X_{2t} + U_t - \beta_2 X_{2t})^2 = E(U_t)^2 = \sigma^2$$

$$\text{Cov}(w_t, w_s) = E(w_t - E(w_t))(w_s - E(w_s)) = E(U_t U_s) = 0$$

The Slope Coefficient

(1) $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + U_t$ Correct model

(2) $Y_t = \beta_0 + \beta_1 X_{1t} + w_t$ misspecified model

The estimated slope coefficient from the misspecified model (2):

$$\hat{\beta}_1 = \frac{\sum x_{t1} y_t}{\sum x_{t1}^2}$$

$$\tilde{\beta}_1 = \frac{\sum a_{t1} Y_t}{\sum x_{t1}^2} \quad \text{where } a_{t1} = \frac{x_{t1}}{\sum x_{t1}^2}$$

$$\tilde{\beta}_1 = \sum a_{t1} (\beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + u_t)$$

$$\tilde{\beta}_1 = \underbrace{\beta_0 \sum a_{t1}}_{=0} + \underbrace{\beta_1 \sum a_{t1} x_{t1}}_{=1} + \underbrace{\beta_2 \sum a_{t1} x_{t2}}_{=0} + \sum a_{t1} u_t$$

$$\bullet \sum a_{t1} x_{t1} = \sum a_{t1} (x_{t1} + \bar{x}_1) = \underbrace{\sum a_{t1} x_{t1}}_{=1} + \bar{x}_1 \underbrace{\sum a_{t1}}_{=0}$$

$$\Rightarrow \sum a_{t1} x_{t1} = 1$$

$$\begin{aligned} \bullet \sum a_{t1} x_{t2} &= \sum a_{t1} (x_{t2} + \bar{x}_2) = \sum a_{t1} x_{t2} + \underbrace{\bar{x}_2 \sum a_{t1}}_{=0} \\ &= \sum a_{t1} x_{t2} \end{aligned}$$

$$\frac{\sum a_{t1} x_{t2}}{\sum x_{t1}^2} = \hat{c}_{21}$$

From:

$$x_{t2} = c_{20} + c_{21} x_{t1} + v_{t2}$$

$$\Rightarrow \tilde{\beta}_1 = \beta_1 + \beta_2 \underbrace{\sum a_{t1} x_{t2}}_{\hat{c}_{21}} + \sum a_{t1} u_t$$

BIAS

$$\tilde{\beta}_1 = \beta_1 + \beta_2 \hat{c}_{21} + \sum a_{t1} u_t$$

$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \hat{c}_{21}$
 the coefficient of the excluded variable $\leftarrow$ coefficient from the regression of the excluded variable on the included variable

$$E(\tilde{\beta}_1) = \beta_1 = \underbrace{\beta_2 \hat{c}_{21}}_{\text{BIAS}}$$

$\tilde{\beta}_1$ is biased

example:

Correct model: $Y_t = \beta_0 + \beta_1 K_t + \beta_2 L_t + U_t$

But: $Y_t = \beta_0 + \beta_1 K_t + w_t$

Suppose the result is: $Y_t = 10 + \underbrace{1.2}_{\text{biased}} K_t$

The direction of bias

The coefficient of the excluded var(t) $\neq c_{21}$

where $L_t = c_{20} + \underbrace{c_{21}}_+ K_t + U_t$

$\Rightarrow$ upward bias

Variance

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\sum x_{t1}^2} \quad \text{from incorrect model}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum \hat{u}_{t1}^2} \quad \text{from the correct model}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{t1}^2 (1 - r_{12}^2)}$$

where r_{12}^2 is the square of the simple correlation coef. between x_{t1} and x_{t2}

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{t1}^2 (1 - r_{12}^2)}$$

$$\text{Var}(\hat{\beta}_1) = \text{Var}(\tilde{\beta}_1) \quad \text{iff} \quad r_{12}^2 = 0$$

$$\text{Var}(\hat{\beta}_1) > \text{Var}(\tilde{\beta}_1) \quad \text{iff} \quad 0 < r_{12}^2 < 1$$

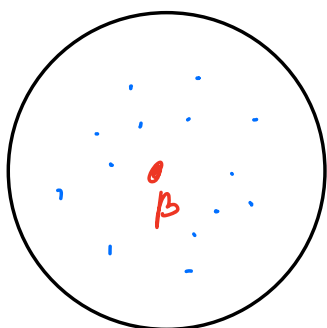
Results

* $\tilde{\beta}_1$: biased; bias = $\beta_2 \hat{c}_{21}$

* $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$

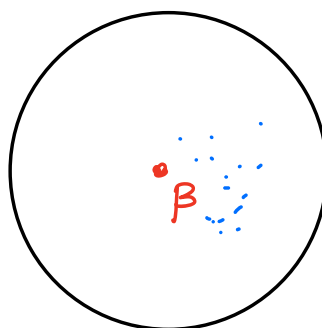
$\tilde{\beta}_1$ is biased but has a smaller variance

* $\tilde{\sigma}^2 = \frac{\sum \tilde{w}_t^2}{T-k-1}$ Biased
No valid inference



$\hat{\beta}$

unbiased but
a larger variance



$\tilde{\beta}$

biased but
smaller variance

Inclusion of an Irrelevant Variable

$$(1) Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t \quad \text{Misspecified Variable}$$

$$(2) Y_t = \beta_0 + \beta_1 X_{1t} + u_t \quad \text{Correct Model}$$

correct model satisfies GM assumptions
misspecified model

$$\tilde{\beta}_1 = \frac{\sum \hat{u}_{ti} y_t}{\sum \hat{u}_{ti}^2}$$

BIAS

$$\tilde{\beta}_1 = \frac{\sum \hat{u}_{ti} y_t}{\sum \hat{u}_{ti}^2} = \sum u_{ti} y_t$$

$$\begin{aligned} \tilde{\beta}_1 &= \sum a_{ti} (\beta_0 + \beta_1 X_{t1} + u_t) \\ &= \beta_0 \sum a_{ti} + \beta_1 \sum u_{ti} X_{t1} + \sum a_{ti} u_t \end{aligned}$$

$$\tilde{\beta}_1 = \beta_1 \underbrace{\sum a_{t1}}_{=1} X_t + \sum a_{t1} u_t$$

$$\tilde{\beta}_1 = \beta_1 + \sum a_{t1} u_t$$

$$E(\tilde{\beta}_1) = \beta_1 = 0 \quad \tilde{\beta}_1 \text{ is UNBIASED}$$

$$\tilde{\beta}_2 = \beta_1 \sum a_{t2} X_{t1} + \sum a_{t2} u_t$$

$$\tilde{\beta}_2 = \sum a_{t2} u_t = 0 \quad E(\tilde{\beta}_2) = 0$$

$\tilde{\beta}_2$: unbiased since the true (population) value of β_2 is 0

Variance

$$\text{Var}(\tilde{\beta}_i) = \frac{\sigma^2}{\sum u_{ti}^2} = \frac{\sigma^2}{\sum x_{ti}^2 (1 - r_{12}^2)}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{t1}^2}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{t1}^2}$$

$$\text{Var}(\hat{\beta}_1) \geq \text{Var}(\tilde{\beta}_1)$$

When an irrelevant (redundant) variable is included, then the variances of the slope coef. of the misspecified model decreases

$$\left. \begin{array}{l} \tilde{\sigma}^2 = \text{unbiased} \\ \tilde{\sigma}_{\tilde{\beta}_i}^2 = \text{unbiased} \end{array} \right\} \begin{array}{l} \text{the usual tests of} \\ \text{inference are valid} \end{array}$$

Testing for Specification Error

Consider

$$Y_t = \beta_0 + \beta_1 X_{1t} + \dots + \beta_k X_{kt} + u_t$$

The equation

- i) may exclude some relevant variables
- or
- ii) may contain some irrelevant/redundant variables
- under ii) straightforward: t and F tests
- i) if known: add to the regression and test for the significance. However: often unknown

unknown:

H_0 : The model is correctly specified

H_A : The model is NOT correctly specified

Under H_0 : $E(w_t) = 0$

Under H_A : $E(w_t)$ can be specified as a linear function of a set of variables

$$E(w_t) = \theta_0 + \theta_1 z_{1t} + \theta_2 z_{2t} + \dots + \theta_m z_{mt}$$

Then H_0 : $\theta_0 = \theta_1 = \theta_2 = \dots = \theta_m = 0$

Can be tested by estimating:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \dots + \beta_k X_{kt} + \theta_1 z_{1t} + \theta_2 z_{2t} + \dots + \theta_m z_{mt} + U_t$$

However,

z_t is often unknown

But we know that:

In the omitted variable case, the bias of the coefficients depend on the excluded and included variables

thus, the omitted variables should be correlated with the existing variables, if they are to be effective

A proxy for z_t : higher powers and cross-products of the explanatory variables. BUT: df problem

The same information is indeed summarized by the fitted values of Y_t

Regression Specification Error Test : RESET

$$(2) Y_t = \beta_0 + \beta_1 X_{t1} + \dots + \beta_k X_{tk} + \theta_1 \hat{Y}_t^2 + \dots + \theta_m \hat{Y}_t^{m+1} + v_t$$

and then

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_m = 0$$

$$Q_{\text{reset}} = \frac{S_R - S_u}{S_u} \cdot \frac{df_u}{p}$$

S_R : From (1) ; S_u from (2)

$m=1$ or $m=2 \rightarrow$ f reasonable choice