MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

Conseguenze di una mispecificazione			
		Modello Vero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modello Stimato	$\hat{Y} = b_1 + b_2 X_2$		
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		

In questa sequenza mostreremo le conseguenze di un modello non ben specificato rispetto alle variabili esplicative.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

Conseguenze di una mispecificazione			
		Modello Vero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modello Stimato	$\hat{Y} = b_1 + b_2 X_2$		
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		

Per mantenere l'analisi semplice, assumiamo che ci siano soltanto due possibilità: 1) *Y* dipende solo da X_2, o 2) *Y* dipende da X_2 e X_3.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

Conseguenze di una mispecificazione			
		Modello Vero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modello Stimato	$\hat{Y} = b_1 + b_2 X_2$	**Corretta specificazione, nessun problema!!!**	
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		

Se *Y* dipende solo da X_2, e stimiamo un modello di regressione non avremmo nessun tipo di problema (ovviamente assumendo che siano valide tutte le assunzioni del modello di regressione).

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

Conseguenze di una mispecificazione			
		Modello Vero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modello Stimato	$\hat{Y} = b_1 + b_2 X_2$	**Corretta specificazione, nessun problema!!!**	
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		**Corretta specificazione, nessun problema!!!**

Se *Y* dipende da X_2 e da X_3 allora non avremmo nessun problema a stimare il modello di regressione multiplo.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

Conseguenze di una mispecificazione			
		Modello Vero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modello Stimato	$\hat{Y} = b_1 + b_2 X_2$	**Corretta specificazione, nessun problema!!!**	
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		**Corretta specificazione, nessun problema!!!**

Adesso esamineremo le conseguenze della stima di una regressione semplice quando il modello vero è quello multiplo.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

Conseguenze di una mispecificazione			
		Modello Vero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modello Stimato	$\hat{Y} = b_1 + b_2 X_2$	**Corretta specificazione, nessun problema!!!**	
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		**Corretta specificazione, nessun problema!!!**

Successivamente si mostrerà l'opposto, ovvero le conseguenze della stima di un modello di regressione multiplo quando il modello vero è quello semplice.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

Conseguenze di una mispecificazione			
		Modello Vero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modello Stimato	$\hat{Y} = b_1 + b_2 X_2$	**Corretta specificazione, nessun problema!!!**	**I Coefficienti sono distorti (in generale). Standard error non sono "validi".**
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		**Corretta specificazione, nessun problema!!!**

L'omissione di una variabile esplicativa rilevante porta ad avere dei coefficienti di regressione distorti e gli standard error non sono validi.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

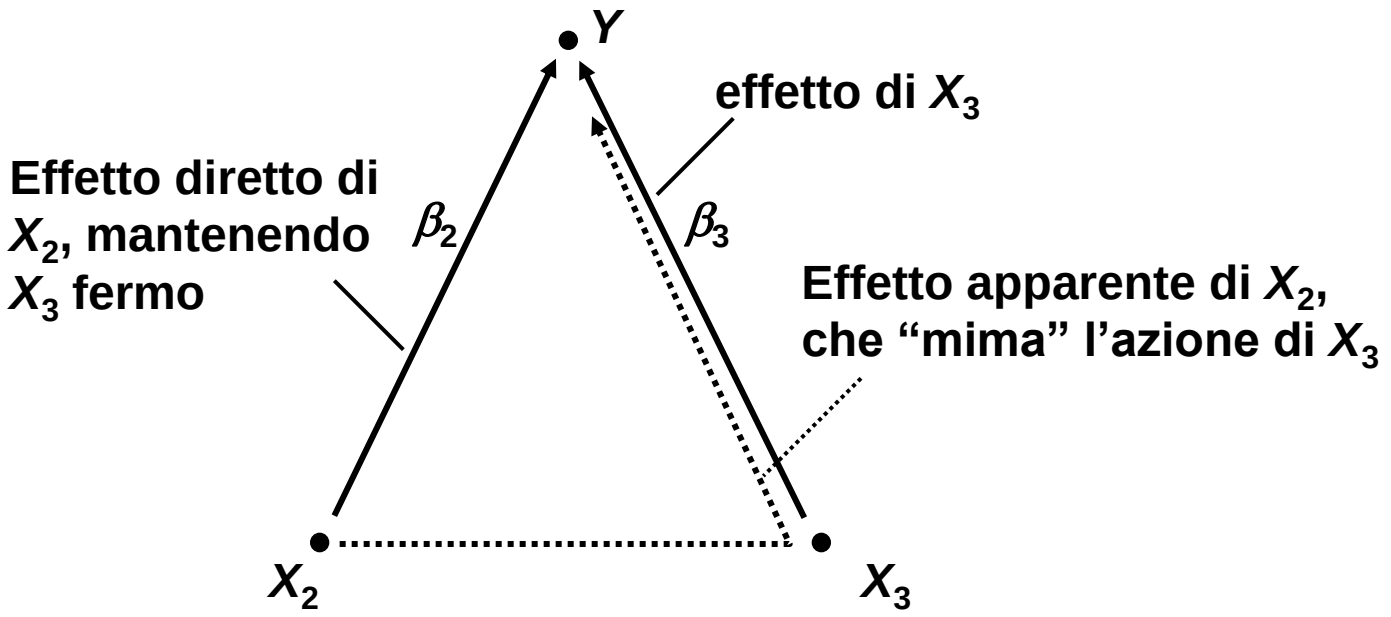
$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \overline{X}_2)(X_{3i} - \overline{X}_3)}{\sum (X_{2i} - \overline{X}_2)^2}$$

In questo caso, l'omissione di X_3 porta ad avere uno stimatore b_2 distorto (la distorsione è evidenziata in giallo). Spiegheremo ciò prima intuitivamente e poi verrà dimostrato matematicamente.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \overline{X}_2)(X_{3i} - \overline{X}_3)}{\sum (X_{2i} - \overline{X}_2)^2}$$

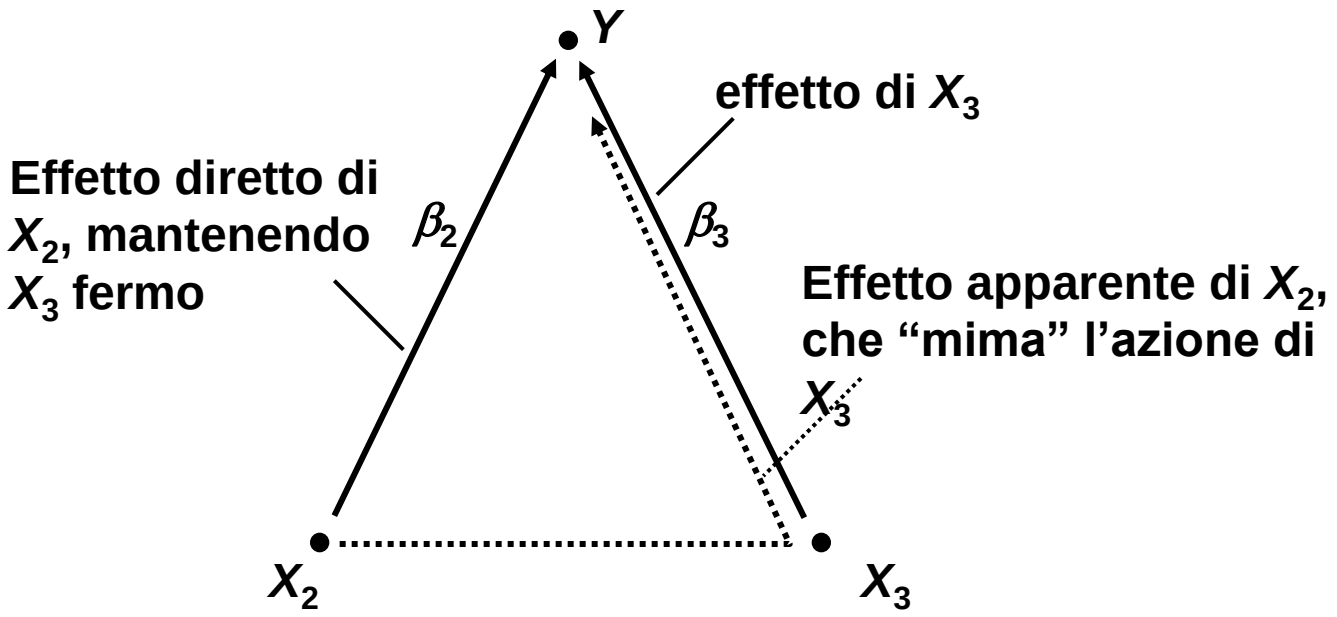


La ragione intuitiva è che, oltre all'effetto diretto β_2, X_2 ha un apparente effetto indiretto che agisce attraverso la variabile X_3 che è stata omessa.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \overline{X}_2)(X_{3i} - \overline{X}_3)}{\sum (X_{2i} - \overline{X}_2)^2}$$

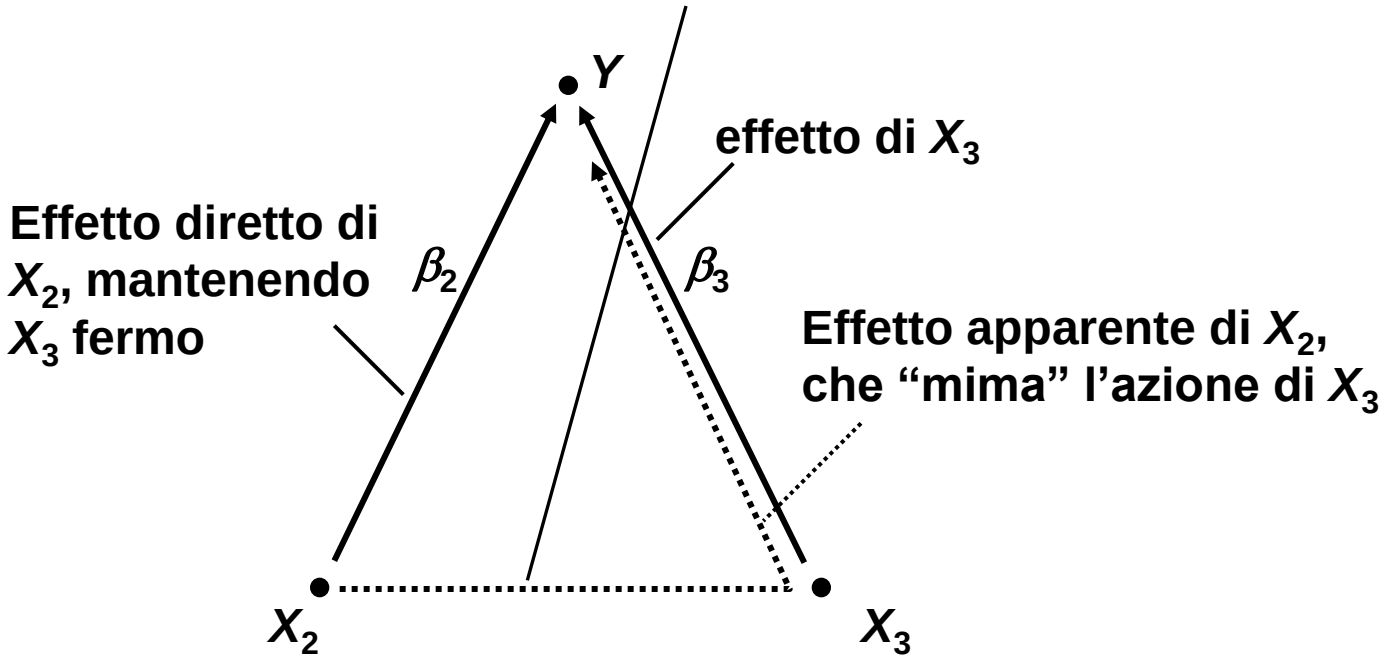


La forza dell'effetto proxy dipende da due fattori: la forza dell'effetto di X_3 su Y, ovvero β_3, e l'abilità di X_2 di mimare X_3.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$



L'abilità di X_2 di mimare X_3 viene determinata dal coefficiente angolare ottenuto quando X_3 viene regredita su X_2 (il termine evidenziato in giallo).

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$\begin{aligned} Y_i - \bar{Y} &= \left(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i\right) - \left(\beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \bar{u}\right) \\ &= \beta_2 \left(X_{2i} - \bar{X}_2\right) + \beta_3 \left(X_{3i} - \bar{X}_3\right) + u_i - \bar{u} \end{aligned}$$

Adesso deriviamo matematicamente l'espressione per la distorsione. È conveniente iniziare a derivare un'espressione per gli scarti di Y_i dalla media campionaria. Essa può essere espressa in termini di scarti di X_2, X_3 e u dalle loro medie campionarie.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$\begin{aligned} Y_i - \bar{Y} &= \left(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i\right) - \left(\beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \bar{u}\right) \\ &= \beta_2 \left(X_{2i} - \bar{X}_2\right) + \beta_3 \left(X_{3i} - \bar{X}_3\right) + u_i - \bar{u} \end{aligned}$$

$$b_2 = \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

Nonostante dipenda realmente da X_3 così come da X_2, commettiamo un errore se regrediamo semplicemente Y su X_2. Il coefficiente angolare viene riportato sopra.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$\begin{aligned} Y_i - \bar{Y} &= \left(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i\right) - \left(\beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \bar{u}\right) \\ &= \beta_2 \left(X_{2i} - \bar{X}_2\right) + \beta_3 \left(X_{3i} - \bar{X}_3\right) + u_i - \bar{u} \end{aligned}$$

$$b_2 = \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

$$= \frac{\sum \left[\beta_2 \left(X_{2i} - \bar{X}_2\right)^2 + \beta_3 \left(X_{2i} - \bar{X}_2\right)\left(X_{3i} - \bar{X}_3\right) + \left(X_{2i} - \bar{X}_2\right)\left(u_i - \bar{u}\right)\right]}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

$$= \beta_2 + \beta_3 \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(X_{3i} - \bar{X}_3\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2} + \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(u_i - \bar{u}\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

Adesso sostituiamo gli scarti di *Y* e semplifichiamo.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$\begin{aligned} Y_i - \bar{Y} &= \left(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i\right) - \left(\beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \bar{u}\right) \\ &= \beta_2 \left(X_{2i} - \bar{X}_2\right) + \beta_3 \left(X_{3i} - \bar{X}_3\right) + u_i - \bar{u} \end{aligned}$$

$$b_2 = \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

$$= \frac{\sum \left[\beta_2 \left(X_{2i} - \bar{X}_2\right)^2 + \beta_3 \left(X_{2i} - \bar{X}_2\right)\left(X_{3i} - \bar{X}_3\right) + \left(X_{2i} - \bar{X}_2\right)\left(u_i - \bar{u}\right)\right]}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

$$= \beta_2 + \beta_3 \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(X_{3i} - \bar{X}_3\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2} + \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(u_i - \bar{u}\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

Abbiamo dimostrato che b_2 è costituita da tre componenti.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$b_2 = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + \frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2}$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

Per verificare la correttezza o la distorsione, consideriamo il valore atteso di b_2. I primi due termini sono componenti deterministiche. Quindi ci possiamo focalizzare sul valore atteso del termine d'errore.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

$$E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right) = \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} E\left(\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u}) \right)$$

X_2 è non stocastica, quindi il denominatore del termine d'errore è non stocastico e può essere portato fuori dall'espressione del valore atteso.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

$$E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right) = \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} E\left(\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u}) \right)$$

$$= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \sum E\left\{ (X_{2i} - \bar{X}_2)(u_i - \bar{u}) \right\}$$

Al numeratore il valore atteso della somma è uguale alla somma dei valori attesi (prima regola del valore atteso).

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

$$E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right) = \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} E\left(\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u}) \right)$$

$$= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \sum E\left\{ (X_{2i} - \bar{X}_2)(u_i - \bar{u}) \right\}$$

$$= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \sum (X_{2i} - \bar{X}_2) E(u_i - \bar{u})$$

In ogni prodotto, il fattore che coinvolge X_2 può essere portato fuori del valore atteso perché X_2 non è stocastico.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

$$E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right) = \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} E\left(\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u}) \right)$$

$$= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \sum E\{(X_{2i} - \bar{X}_2)(u_i - \bar{u})\}$$

$$= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \sum (X_{2i} - \bar{X}_2) E(u_i - \bar{u})$$

$$= 0$$

Sulla base dell'Assunzione A.3, il valore atteso di *u* è 0. Segue che il valore atteso della media campionaria di *u* è anche 0. Quindi il valore atteso del termine d'errore è 0.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$

Quindi abbiamo dimostrato che il valore atteso di b_2 è eguale al vero valore più un termine d'errore. Nota: la definizione di distorsione è la differenza tra il il valore atteso dello stimatore e il vero valore del parametro che deve essere stimato.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$

Come conseguenza della mispecificazione, gli standard error, i test *t* e *F* test non sono validi.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM
```

```
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109            R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.963

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

Mostreremo un esempio di mispecificazione usando i dati *EAEF* . Volendo mantenere l'analisi semplice, assumiamo che il vero modello sia *S* funzione di *ASVABC* e *SM*. Sopra è riportato l'output usando *EAEF* Data Set 21.

MISPECIFICAZIONE: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM
```

```
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109            R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.963

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum \left(ASVABC_i - \overline{ASVABC}\right)\left(SM_i - \overline{SM}\right)}{\sum \left(ASVABC_i - \overline{ASVABC}\right)^2}$$

Facciamo girare la regressione una seconda volta omettendo *SM*. Prima di fare questo, cerchiamo di predire la direzione della distorsione del coefficiente di *ASVABC*.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM
```

```
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109            R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.963

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |    .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |    .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |    5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum \left(ASVABC_i - \overline{ASVABC}\right)\left(SM_i - \overline{SM}\right)}{\sum \left(ASVABC_i - \overline{ASVABC}\right)^2}$$

É ragionevole supporre che β_3 sia positivo. Questa assunzione è fortemente supportata dal fatto che la stima nella regressione multipla sia positiva ed altamente significativa.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM

      Source |       SS       df       MS
-------------+------------------------------
       Model |  1135.67473     2  567.837363
    Residual |  2069.30861   537  3.85346109
-------------+------------------------------
       Total |  3204.98333   539  5.94616574

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

```
. cor SM ASVABC
(obs=540)

             |       SM   ASVABC
-------------+------------------
          SM|   1.0000
      ASVABC|   0.4202   1.0000
```

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum \left(ASVABC_i - \overline{ASVABC}\right)\left(SM_i - \overline{SM}\right)}{\sum \left(ASVABC_i - \overline{ASVABC}\right)^2}$$

La correlazione tra *ASVABC* e *SM* è positiva, quindi il numeratore del termine di distorsione deve essere positivo. Il denominatore è sicuramente positivo dal momento che è dato dalla somma di quadrati. Quindi il *bias* dovrebbe essere positivo.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC
```

```
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  274.19
       Model |  1081.97059     1  1081.97059            Prob > F      =  0.0000
    Residual |  2123.01275   538  3.94612035            R-squared     =  0.3376
-------------+------------------------------           Adj R-squared =  0.3364
       Total |  3204.98333   539  5.94616574            Root MSE      =  1.9865

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |    .148084   .0089431    16.56   0.000     .1305165    .1656516
       _cons |   6.066225   .4672261    12.98   0.000     5.148413    6.984036
------------------------------------------------------------------------------
```

Sopra abbiamo l'output della regressione stimata omettendo *SM*.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------

. reg S ASVABC

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |    .148084   .0089431    16.56   0.000     .1305165    .1656516
       _cons |   6.066225   .4672261    12.98   0.000     5.148413    6.984036
------------------------------------------------------------------------------
```

Si può osservare che il coefficiente di *ASVABC* è più alto quando *SM* viene omessa. Parte della differenza può essere dovuta al puro caso, ma parte è attribuibile al *bias*.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S SM
```

```
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   80.93
       Model |  419.086251     1  419.086251            Prob > F      =  0.0000
    Residual |  2785.89708   538  5.17824736            R-squared     =  0.1308
-------------+------------------------------           Adj R-squared =  0.1291
       Total |  3204.98333   539  5.94616574            Root MSE      =  2.2756

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          SM |   .3130793   .0348012     9.00   0.000     .2447165    .3814422
       _cons |   10.04688   .4147121    24.23   0.000     9.232226    10.86153
------------------------------------------------------------------------------
```

$$E(b_3) = \beta_3 + \beta_2 \frac{\sum \left(ASVABC_i - \overline{ASVABC}\right)\left(SM_i - \overline{SM}\right)}{\sum \left(SM_i - \overline{SM}\right)^2}$$

Sopra è riportato l'output della regressione omettendo *ASVABC* al posto di *SM*. Ci aspettiamo che b_3 sia distorto. Sappiamo che β_2 è positivo e che sia il numeratore che il denominatore dell'altro fattore del *bias* sono positivi.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

```
. reg S SM

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          SM |   .3130793   .0348012     9.00   0.000     .2447165    .3814422
       _cons |   10.04688   .4147121    24.23   0.000     9.232226    10.86153
------------------------------------------------------------------------------
```

In questo caso il bias è abbastanza problematico. Il coefficiente di *SM* è più che raddoppiato. La ragione di questo effetto deriva dal fatto che la variazione in *SM* è molto più piccola di quella in *ASVABC*, mentre β_2 e β_3 sono simili nelle loro grandezze, giudicando dalle loro stime.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109           R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574           Root MSE      =   1.963
```

```
. reg S ASVABC
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  274.19
       Model |  1081.97059     1  1081.97059            Prob > F      =  0.0000
    Residual |  2123.01275   538  3.94612035           R-squared     =  0.3376
-------------+------------------------------           Adj R-squared =  0.3364
       Total |  3204.98333   539  5.94616574           Root MSE      =  1.9865
```

```
. reg S SM
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   80.93
       Model |  419.086251     1  419.086251            Prob > F      =  0.0000
    Residual |  2785.89708   538  5.17824736           R-squared     =  0.1308
-------------+------------------------------           Adj R-squared =  0.1291
       Total |  3204.98333   539  5.94616574           Root MSE      =  2.2756
```

Infine, osserviamo come si comporta R^2 quando la variabile viene omessa, nella regressione semplice *S* su *ASVABC*, R^2 è 0.34, e nella regressione semplice di *S* su *SM* , R^2 è 0.13.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109           R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574           Root MSE      =   1.963

. reg S ASVABC
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  274.19
       Model |  1081.97059     1  1081.97059            Prob > F      =  0.0000
    Residual |  2123.01275   538  3.94612035           R-squared     =  0.3376
-------------+------------------------------           Adj R-squared =  0.3364
       Total |  3204.98333   539  5.94616574           Root MSE      =  1.9865

. reg S SM
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   80.93
       Model |  419.086251     1  419.086251            Prob > F      =  0.0000
    Residual |  2785.89708   538  5.17824736           R-squared     =  0.1308
-------------+------------------------------           Adj R-squared =  0.1291
       Total |  3204.98333   539  5.94616574           Root MSE      =  2.2756
```

Questo implica che *ASVABC* spiega il 34% della varianza di *S* mentre per *SM* il 13%? No, perché la regressione multipla rivela che la capacità esplicativa congiunta è 0.35, e non 0.47.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109            R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.963

. reg S ASVABC
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  274.19
       Model |  1081.97059     1  1081.97059            Prob > F      =  0.0000
    Residual |  2123.01275   538  3.94612035            R-squared     =  0.3376
-------------+------------------------------           Adj R-squared =  0.3364
       Total |  3204.98333   539  5.94616574            Root MSE      =  1.9865

. reg S SM
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   80.93
       Model |  419.086251     1  419.086251            Prob > F      =  0.0000
    Residual |  2785.89708   538  5.17824736            R-squared     =  0.1308
-------------+------------------------------           Adj R-squared =  0.1291
       Total |  3204.98333   539  5.94616574            Root MSE      =  2.2756
```

Nella seconda regressione, *ASVABC* sta agendo come proxy per *SM*, e questo inflaziona la sua capacità esplicativa. In maniera simile, nella terza regressione, *SM* agisce parzialmente come proxy per *ASVABC*, e di nuovo questo inflaziona la sua capacità esplicativa.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

```
. reg LGEARN S EXP

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  100.86
       Model |  50.9842581     2   25.492129            Prob > F      =  0.0000
    Residual |  135.723385   537  .252743734            R-squared     =  0.2731
-------------+------------------------------           Adj R-squared =  0.2704
       Total |  186.707643   539   .34639637            Root MSE      =  .50274

------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .1235911   .0090989    13.58   0.000     .1057173     .141465
         EXP |   .0350826   .0050046     7.01   0.000     .0252515    .0449137
       _cons |   .5093196   .1663823     3.06   0.002     .1824796    .8361596
------------------------------------------------------------------------------
```

Comunque, è anche possibile avere una riduzione della capacità esplicativa nel caso di omissioni di variabili rilevanti

Questo verrà mostrato usando il modello di regressione logaritmo del guadagno orario su *S* e *EXP*.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

```
. reg LGEARN S EXP

      Source |       SS       df       MS
-------------+------------------------------
       Model |  50.9842581     2   25.492129
    Residual |  135.723385   537  .252743734
-------------+------------------------------
       Total |  186.707643   539   .34639637

------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .1235911   .0090989    13.58   0.000     .1057173     .141465
         EXP |   .0350826   .0050046     7.01   0.000     .0252515    .0449137
       _cons |   .5093196   .1663823     3.06   0.002     .1824796    .8361596
------------------------------------------------------------------------------
```

```
. cor S EXP
(obs=540)

             |        S      EXP
-------------+------------------
           S|   1.0000
         EXP|  -0.2179   1.0000
```

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (S_i - \bar{S})(EXP_i - \overline{EXP})}{\sum (S_i - \bar{S})^2}$$

Se omettiamo *EXP* dalla regressione, il coefficiente di *S* dovrebbe essere soggetto ad una distorsione negativa. β_3 è sicuramente positivo. Il numeratore del *bias* è negativo dal momento che *S* e *EXP* sono negativamente correlati. Il denominatore è positivo.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

```
. reg LGEARN S EXP

      Source |       SS       df       MS
-------------+------------------------------
       Model |  50.9842581     2   25.492129
    Residual |  135.723385   537  .252743734
-------------+------------------------------
       Total |  186.707643   539   .34639637

------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .1235911   .0090989    13.58   0.000     .1057173     .141465
         EXP |   .0350826   .0050046     7.01   0.000     .0252515    .0449137
       _cons |   .5093196   .1663823     3.06   0.002     .1824796    .8361596
------------------------------------------------------------------------------
```

```
. cor S EXP
(obs=540)

             |        S      EXP
-------------+------------------
           S|   1.0000
         EXP|  -0.2179   1.0000
```

$$E(b_3) = \beta_3 + \beta_2 \frac{\sum \left(EXP_i - \overline{EXP}\right)\left(S_i - \overline{S}\right)}{\sum \left(EXP_i - \overline{EXP}\right)^2}$$

Il coefficiente di *EXP* nella regressione semplice di *LGEARN* su *EXP* dovrebbe essere soggetto ad una distorsione negativa

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

```
. reg LGEARN S EXP
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    .1235911   .0090989    13.58   0.000     .1057173     .141465
         EXP |    .0350826   .0050046     7.01   0.000     .0252515    .0449137
       _cons |    .5093196   .1663823     3.06   0.002     .1824796    .8361596

. reg LGEARN S
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    .1096934   .0092691    11.83   0.000     .0914853    .1279014
       _cons |   1.292241   .1287252    10.04   0.000     1.039376    1.545107

. reg LGEARN EXP
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |    .0202708   .0056564     3.58   0.000     .0091595     .031382
       _cons |    2.44941   .0988233    24.79   0.000     2.255284    2.643537
```

Come si può vedere, i coefficienti di *S* e *EXP* sono più bassi nelle regressioni semplici.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

```
. reg LGEARN S EXP
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  100.86
       Model |  50.9842581     2  25.492129              Prob > F      =  0.0000
    Residual |  135.723385   537  .252743734             R-squared     =  0.2731
-------------+------------------------------           Adj R-squared =  0.2704
       Total |  186.707643   539   .34639637            Root MSE      =  .50274

. reg LGEARN S
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  140.05
       Model |  38.5643833     1  38.5643833              Prob > F      =  0.0000
    Residual |   148.14326   538  .275359219             R-squared     =  0.2065
-------------+------------------------------           Adj R-squared =  0.2051
       Total |  186.707643   539   .34639637            Root MSE      =  .52475

. reg LGEARN EXP
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   12.84
       Model |  4.35309315     1  4.35309315              Prob > F      =  0.0004
    Residual |   182.35455   538  .338948978             R-squared     =  0.0233
-------------+------------------------------           Adj R-squared =  0.0215
       Total |  186.707643   539   .34639637            Root MSE      =  .58219
```

Un confronto fra gli R^2 per le tre regressioni mostra come la somma di R^2 nelle regressioni semplici è in realtà inferiore al valore di R^2 nella regressione multipla.

MISPECIFICAZIONE I: OMISSIONE DI UNA VARIABILE RILEVANTE

```
. reg LGEARN S EXP
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  100.86
       Model |  50.9842581     2  25.492129              Prob > F      =  0.0000
    Residual |  135.723385   537  .252743734             R-squared     =  0.2731
-------------+------------------------------           Adj R-squared =  0.2704
       Total |  186.707643   539   .34639637            Root MSE      =  .50274

. reg LGEARN S
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  140.05
       Model |  38.5643833     1  38.5643833              Prob > F      =  0.0000
    Residual |   148.14326   538  .275359219             R-squared     =  0.2065
-------------+------------------------------           Adj R-squared =  0.2051
       Total |  186.707643   539   .34639637            Root MSE      =  .52475

. reg LGEARN EXP
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   12.84
       Model |  4.35309315     1  4.35309315              Prob > F      =  0.0004
    Residual |   182.35455   538  .338948978             R-squared     =  0.0233
-------------+------------------------------           Adj R-squared =  0.0215
       Total |  186.707643   539   .34639637            Root MSE      =  .58219
```

Questo perché la capacità esplicativa di *S* nella seconda regressione è stata inficiata dal bias negativo. Ciò accade anche per la terza equazione.