MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

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!!!**

In questa sequenza mostreremo le conseguenze relativamente all'inclusione di una variabile irrilevante in un modello di regressione.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

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".**
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$	**I Coefficienti sono corretti (in generale), ma non effcienti. Standard error sono validi (in generale)**	**Corretta specificazione, nessun problema!!!**

Gli effetti sono differenti rispetto a quelli derivanti dall'omissione di una variabile rilevante. In questo caso i coefficienti restano corretti (in generale), ma sono inefficienti. Gli standard error restano validi, ma sono molto alti.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

Questi risultati possono essere mostrati molto velocemente.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

$$Y = \beta_1 + \beta_2 X_2 + 0 X_3 + u$$

Riscrivi il vero modello aggiungendo X_3 come variabile esplicativa, con coefficiente pari a 0. Quindi b_2 sarà uno stimatore corretto di β_2 e b_3 sarà uno stimatore corretto di 0.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

$$Y = \beta_1 + \beta_2 X_2 + 0 X_3 + u$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum \left(X_{2i} - \overline{X}_2 \right)^2} \times \frac{1}{1 - r_{X_2, X_3}^2}$$

Comunque, la varianza di b_2 sarà più grande da quello che si avrebbe se venisse considerato il vero modello, in quanto viene incluso il fattore 1 / (1 – r^2), dove r è il coefficiente di correlazione tra X_2 e X_3.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

$$Y = \beta_1 + \beta_2 X_2 + 0 X_3 + u$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum \left(X_{2i} - \overline{X}_2 \right)^2} \times \frac{1}{1 - r_{X_2, X_3}^2}$$

Lo stimatore b_2, usando il modello di regressione multiplo, sarà meno efficiente da quello che si otterrebbe considerando il modello di regressione semplice.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

$$Y = \beta_1 + \beta_2 X_2 + 0 X_3 + u$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum \left(X_{2i} - \overline{X}_2 \right)^2} \times \frac{1}{1 - r_{X_2, X_3}^2}$$

La ragione intuitiva è che il modello di regressione semplice parte dal vantaggio che la variabile X_3 non deve essere inserita nel modello di regressione, mentre nel modello di regressione multiplo si capisce della poca utilità della variabile X_3 dai risultati delle stime.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

$$Y = \beta_1 + \beta_2 X_2 + 0 X_3 + u$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum \left(X_{2i} - \overline{X}_2 \right)^2} \times \frac{1}{1 - r_{X_2, X_3}^2}$$

Gli standard error restano validi, perché il modello è correttamente (formalmente) specificato, ma saranno più elevati da quelli ottenuti considerando un modello di regressione semplice (perdita di efficienza).

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

$$Y = \beta_1 + \beta_2 X_2 + 0 X_3 + u$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum \left(X_{2i} - \overline{X}_2 \right)^2} \times \frac{1}{1 - r_{X_2, X_3}^2}$$

Questi risultati possono ovviamente essere generalizzati.

Nota che se X_2 e X_3 non sono correlati non ci sarà nessuna perdita di efficienza.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

```
. reg LGFDHO LGEXP LGSIZE

  Source |       SS       df       MS                  Number of obs =     868
---------+------------------------------                F(  2,   865) =  460.92
   Model |  138.776549     2  69.3882747                Prob > F      =  0.0000
Residual |  130.219231   865  .150542464                R-squared     =  0.5159
---------+------------------------------                Adj R-squared =  0.5148
   Total |  268.995781   867  .310260416                Root MSE      =    .388

------------------------------------------------------------------------------
  LGFDHO |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   LGEXP |   .2866813   .0226824     12.639   0.000       .2421622    .3312003
  LGSIZE |   .4854698   .0255476     19.003   0.000       .4353272    .5356124
   _cons |   4.720269   .2209996     21.359   0.000       4.286511    5.154027
------------------------------------------------------------------------------
```

Quanto detto verrà illustrato tramite un esempio usando *LGFDHO*, il logaritmo della spesa familiare annuale relativamente al cibo consumato a casa, su *LGEXP*, il logaritmo della spesa totale familiare annuale, e *LGSIZE*, il logaritmo del numero di persone che compongono una famiglia.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

```
. reg LGFDHO LGEXP LGSIZE

  Source |       SS       df       MS                  Number of obs =     868
---------+------------------------------                F(  2,   865) =  460.92
   Model |  138.776549     2  69.3882747                Prob > F      =  0.0000
Residual |  130.219231   865  .150542464                R-squared     =  0.5159
---------+------------------------------                Adj R-squared =  0.5148
   Total |  268.995781   867  .310260416                Root MSE      =    .388

------------------------------------------------------------------------------
  LGFDHO |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   LGEXP |   .2866813   .0226824     12.639   0.000       .2421622    .3312003
  LGSIZE |   .4854698   .0255476     19.003   0.000       .4353272    .5356124
   _cons |   4.720269   .2209996     21.359   0.000       4.286511    5.154027
------------------------------------------------------------------------------
```

I dati sono quelli del 1995 del data set US Consumer Expenditure Survey. La dimensione campionaria è di 868 famiglie.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

```
. reg LGFDHO LGEXP LGSIZE LGHOUS

  Source |       SS       df       MS                   Number of obs =     868
---------+------------------------------                 F(  3,   864) =  307.22
   Model |  138.841976     3  46.2806586                  Prob > F      =  0.0000
Residual |  130.153805   864  .150640978                  R-squared     =  0.5161
---------+------------------------------                 Adj R-squared =  0.5145
   Total |  268.995781   867  .310260416                  Root MSE      =  .38812

------------------------------------------------------------------------------
  LGFDHO |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   LGEXP |   .2673552   .0370782      7.211   0.000        .1945813     .340129
  LGSIZE |   .4868228   .0256383     18.988   0.000        .4365021    .5371434
  LGHOUS |   .0229611   .0348408      0.659   0.510       -.0454214    .0913436
   _cons |   4.708772   .2217592     21.234   0.000        4.273522    5.144022
------------------------------------------------------------------------------
```

Adesso aggiungiamo *LGHOUS*, il logaritmo della spesa familiare annuale per i servizi abitativi. É plausibile assumere che *LGHOUS* sia una variabile irrilevante, non sorprendentemente il suo coefficiente non è significativamente diverso da zero.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

```
. reg LGFDHO LGEXP LGSIZE LGHOUS

  Source |       SS       df       MS
---------+------------------------------
   Model |  138.841976     3  46.2806586
Residual |  130.153805   864  .150640978
---------+------------------------------
   Total |  268.995781   867  .310260416

------------------------------------------
  LGFDHO |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
   LGEXP |   .2673552   .0370782     7.211   0.000     .1945813     .340129
  LGSIZE |   .4868228   .0256383    18.988   0.000     .4365021    .5371434
  LGHOUS |   .0229611   .0348408     0.659   0.510    -.0454214    .0913436
   _cons |   4.708772   .2217592    21.234   0.000     4.273522    5.144022
------------------------------------------------------------------------------
```

```
. cor LGHOUS LGEXP LGSIZE
(obs=869)

         |   LGHOUS    LGEXP   LGSIZE
---------+---------------------------
  lGHOUS|   1.0000
   LGEXP|   0.8137   1.0000
  LGSIZE|   0.3256   0.4491   1.0000
```

Essa è altamente corelata con *LGEXP* (coefficiente di corelazione pari a 0.81) e leggermente correlata con *LGSIZE* (coefficiente di corelazione pari a 0.33).

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

```
. reg LGFDHO LGEXP LGSIZE
```

LGFDHO	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LGEXP	.2866813	.0226824	12.639	0.000	.2421622	.3312003
LGSIZE	.4854698	.0255476	19.003	0.000	.4353272	.5356124
_cons	4.720269	.2209996	21.359	0.000	4.286511	5.154027

```
. reg LGFDHO LGEXP LGSIZE LGHOUS
```

LGFDHO	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LGEXP	.2673552	.0370782	7.211	0.000	.1945813	.340129
LGSIZE	.4868228	.0256383	18.988	0.000	.4365021	.5371434
LGHOUS	.0229611	.0348408	0.659	0.510	-.0454214	.0913436
_cons	4.708772	.2217592	21.234	0.000	4.273522	5.144022

Comunque , la sua inclusione non porta ad avere degli stimatori distorti.

MISPECIFICAZIONE II: INCLUSIONE DI UNA VARIABILE IRRILEVANTE

```
. reg LGFDHO LGEXP LGSIZE
```

LGFDHO	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LGEXP	.2866813	.0226824	12.639	0.000	.2421622	.3312003
LGSIZE	.4854698	.0255476	19.003	0.000	.4353272	.5356124
_cons	4.720269	.2209996	21.359	0.000	4.286511	5.154027

```
. reg LGFDHO LGEXP LGSIZE LGHOUS
```

LGFDHO	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LGEXP	.2673552	.0370782	7.211	0.000	.1945813	.340129
LGSIZE	.4868228	.0256383	18.988	0.000	.4365021	.5371434
LGHOUS	.0229611	.0348408	0.659	0.510	-.0454214	.0913436
_cons	4.708772	.2217592	21.234	0.000	4.273522	5.144022

Però porta ad un incremento degli standard error, in modo particolare quello di *LGEXP*, come ci si aspettava (quindi si ha una perdita di efficienza).