VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Supponiamo che una variabile *Y* dipenda da un insieme di variabili esplicative X_2, ..., X_k come mostrato sopra, e supponiamo, per qualche ragione, di non disporre dati su X_2.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Come abbiamo già visto, una regressione di *Y* su X_3, ..., X_k porterebbe ad avere delle stime distorte dei coefficienti e degli standard error non validi (e quindi anche i test)

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

A volte, comunque, questi problemi possono essere ridimensionati o eliminati usando una variabile proxy al posto di X_2. Una variabile proxy è quella che viene ipotizzata essere linearmente legata alla variabile mancante. Nell'esempio sopra, Z potrebbe agire come proxy di X_2.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

La validità della relazione proxy deve essere giustificata sulla base della teoria, o senso comune, o esperienza. Essa non può essere controllata direttamente perché non si dispone di dati su X_2.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$Y = \beta_1 + \beta_2 (\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

Se una variabile proxy è stata identificata, il modello di regressione può essere riscritto come sopra.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

Così otteniamo un modello con tutte variabili osservabili. Se la relazione proxy è esatta, e stimiamo il modello molti risultati saranno salvati.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

1. **Le stime dei coefficienti di X_3, ..., X_k sarebbero le stesse di quelle ottenute regredendo Y su X_2, ..., X_k.**

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2\lambda) + \beta_2\mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

2. Gli standard error e le statistiche *t* dei coefficienti di X_3, ..., X_k sarebbero gli stessi di quelli ottenuti regredendo *Y* su X_2, ..., X_k.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2\lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

3. R^2 sarà lo stesso di quello ottenuto regredendo Y su X_2, ..., X_k.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

4. **Il coefficiente di *Z* sarà una stima di $\beta_2\mu$, e non sarà possibile ottenere una stima di β_2, a meno che non si disponga del valore di μ.**

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

5. Comunque la statistica *t* per *Z* sarebbe la stessa di quella ottenuta per X_2 se fosse stato possibile regredire *Y* su X_2, ..., X_k. Quindi in questo caso siamo in grado di valutare la significatività di X_2, anche se non siamo in grado di stimare il suo coefficiente.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

6. Non è possibile ottenere una stima di β_1 dal momento che l'intercetta nel modello rivisitato è (β_1+$\beta_2\lambda$), ma di solito β_1 è di scarso interesse.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

In generale, è più realistico ipotizzare che la relazione tra X_2 e Z sia un'approssimazione, piuttosto che esatta. In questo caso i risultati mostrati sopra sono validi in maniera approssimativa.

VARIABILI PROXY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

Comunque, se *Z* è una proxy (approssimativamente) di X_2, i risultati saranno soggetti ad errori di misurazione. Inoltre, è possibile che qualche altra variabile *X* potrebbe agire come proxy di X_2, e quindi ci potrebbe essere un problema derivante da variabile omessa.

VARIABILI PROXY

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

L'uso della variabile proxy sarà illustrato con un modello sull'istruzione scolastica. Supponiamo che l'istruzione scolastica dipenda congiuntamente dall'abilità cognitiva e dal background familiare.

VARIABILI PROXY

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

ASVABC **sarà usata come una misura della abilità cognitiva. Però, non disponiamo della variabile 'background familiare' nel data set. É difficile immaginare come tale variabile possa essere definita.**

VARIABILI PROXY

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

$$INDEX = \lambda + \mu_1 SM + \mu_2 SF$$

Allora, possiamo trovare una proxy. Una variabile ("ovvia") è il livello di istruzione della madre, *SM*. D'altronde, il livello di istruzione del padre, *SF*, potrebbe essere anche rilevante. Quindi possiamo ipotizzare che il background familiare possa dipendere da entrambi.

VARIABILI PROXY

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

$$INDEX = \lambda + \mu_1 SM + \mu_2 SF$$

$$\begin{aligned} S &= \beta_1 + \beta_2 ASVABC + \beta_3 (\lambda + \mu_1 SM + \mu_2 SF) + u \\ &= (\beta_1 + \beta_3 \lambda) + \beta_2 ASVABC + \beta_3 \mu_1 SM + \beta_3 \mu_2 SF + u \end{aligned}$$

Otteniamo così una relazione che esprime *S* come una funzione di *ASVABC*, *SM*, e *SF*.

VARIABILI PROXY

```
. reg S ASVABC SM SF
```

```
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  3,   536) =  104.30
       Model |  1181.36981     3  393.789935            Prob > F      =  0.0000
    Residual |  2023.61353   536  3.77539837            R-squared     =  0.3686
-------------+------------------------------           Adj R-squared =  0.3651
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.943

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1257087   .0098533    12.76   0.000     .1063528    .1450646
          SM |   .0492424   .0390901     1.26   0.208    -.027546    .1260309
          SF |   .1076825   .0309522     3.48   0.001       .04688    .1684851
       _cons |   5.370631   .4882155    11.00   0.000      4.41158    6.329681
------------------------------------------------------------------------------
```

Sopra riportiamo i risultati della regressione usando il Data Set 21 *EAEF*.

VARIABILI PROXY

```
. 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 sono riportati i risultati di *S* su *ASVABC*.

VARIABILI PROXY

```
. reg S ASVABC SM SF
```

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	.1257087	.0098533	12.76	0.000	.1063528	.1450646
SM	.0492424	.0390901	1.26	0.208	-.027546	.1260309
SF	.1076825	.0309522	3.48	0.001	.04688	.1684851
_cons	5.370631	.4882155	11.00	0.000	4.41158	6.329681

```
. 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

Un confronto tra le due regressioni indica che il coefficiente di *ASVABC* è distorto se non tentiamo di controllare per il background familiare.

VARIABILI PROXY

```
. reg S ASVABC SM SF
```

S	Coef.	Std. Err.
ASVABC	.1257087	.0098533
SM	.0492424	.0390901
SF	.1076825	.0309522
_cons	5.370631	.4882155

```
. cor ASVABC SM SF
(obs=570)
```

	ASVABC	SM	SF
ASVABC	1.0000		
SM	0.4202	1.0000	
SF	0.4090	0.6241	1.0000

```
. 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

***SM* e *SF* hanno un effetto positivo sull'istruzione, e sono correlati positivamente con *ASVABC*.**

VARIABILI PROXY

```
. reg S ASVABC SM SF LIBRARY SIBLINGS
```

Source	SS	df	MS		
Model	1191.57546	5	238.315093	Number of obs	540
Residual	2013.40787	534	3.77042672	F(5, 534)	63.21
Total	3204.98333	539	5.94616574	Prob > F	0.0000
				R-squared	0.3718
				Adj R-squared	0.3659
				Root MSE	1.9418

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	.1245327	.0099875	12.47	0.000	.104913	.1441523
SM	.0388414	.039969	0.97	0.332	-.0396743	.1173571
SF	.1035001	.0311842	3.32	0.001	.0422413	.1647588
LIBRARY	-.0355224	.2134634	-0.17	0.868	-.4548534	.3838086
SIBLINGS	-.0665348	.0408795	-1.63	0.104	-.1468392	.0137696
_cons	5.846517	.5681221	10.29	0.000	4.730489	6.962546

LIBRARY **(una variabile dummy uguale a 1 se qualcuno nella famiglia possiede una tessera per la biblioteca, considerando soltanto i membri della famiglia con un'età superiore a 14)** ***e SIBLINGS*** **(numero di fratelli e sorelle del rispondente) sono due variabili presenti nel data set che possono agire come proxy del background familiare.**

VARIABILI PROXY

```
. reg S ASVABC SM SF LIBRARY SIBLINGS

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  5,   534) =   63.21
       Model |  1191.57546     5  238.315093            Prob > F      =  0.0000
    Residual |  2013.40787   534  3.77042672            R-squared     =  0.3718
-------------+------------------------------           Adj R-squared =  0.3659
       Total |  3204.98333   539  5.94616574            Root MSE      =  1.9418

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1245327   .0099875    12.47   0.000      .104913    .1441523
          SM |   .0388414    .039969     0.97   0.332    -.0396743    .1173571
          SF |   .1035001   .0311842     3.32   0.001     .0422413    .1647588
     LIBRARY |  -.0355224   .2134634    -0.17   0.868    -.4548534    .3838086
    SIBLINGS |  -.0665348   .0408795    -1.63   0.104    -.1468392    .0137696
       _cons |   5.846517   .5681221    10.29   0.000     4.730489    6.962546
------------------------------------------------------------------------------
```

La variabile *LIBRARY* è una delle variabili incluse nell'indagine *National Longitudinal Survey of Youth* per cercare di cogliere l'influenza del background familiare sull'istruzione. Sorprendentemente, ha un effetto negativo, ma il coefficiente non è significativo.

VARIABILI PROXY

```
. reg S ASVABC SM SF LIBRARY SIBLINGS

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  5,   534) =   63.21
       Model |  1191.57546     5  238.315093            Prob > F      =  0.0000
    Residual |  2013.40787   534  3.77042672            R-squared     =  0.3718
-------------+------------------------------           Adj R-squared =  0.3659
       Total |  3204.98333   539  5.94616574            Root MSE      =  1.9418

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1245327   .0099875    12.47   0.000      .104913    .1441523
          SM |   .0388414    .039969     0.97   0.332    -.0396743    .1173571
          SF |   .1035001   .0311842     3.32   0.001     .0422413    .1647588
     LIBRARY |  -.0355224   .2134634    -0.17   0.868    -.4548534    .3838086
    SIBLINGS |  -.0665348   .0408795    -1.63   0.104    -.1468392    .0137696
       _cons |   5.846517   .5681221    10.29   0.000     4.730489    6.962546
------------------------------------------------------------------------------
```

C’è una tendenza per i genitori che sono ambiziosi nei confronti dei propri figli di limitare il loro numero, quindi *SIBLINGS* dovrebbe avere un coefficiente negativo. Ciò è quello che si ottiene, ma non è significativo.

VARIABILI PROXY

```
. reg S ASVABC SM SF LIBRARY SIBLINGS

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  5,   534) =   63.21
       Model |  1191.57546     5  238.315093            Prob > F      =  0.0000
    Residual |  2013.40787   534  3.77042672            R-squared     =  0.3718
-------------+------------------------------           Adj R-squared =  0.3659
       Total |  3204.98333   539  5.94616574            Root MSE      =  1.9418

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1245327   .0099875    12.47   0.000      .104913    .1441523
          SM |   .0388414    .039969     0.97   0.332    -.0396743    .1173571
          SF |   .1035001   .0311842     3.32   0.001     .0422413    .1647588
     LIBRARY |  -.0355224   .2134634    -0.17   0.868    -.4548534    .3838086
    SIBLINGS |  -.0665348   .0408795    -1.63   0.104    -.1468392    .0137696
       _cons |   5.846517   .5681221    10.29   0.000     4.730489    6.962546
------------------------------------------------------------------------------
```

Ci sono ulteriori variabili legate al background familiare che possono essere rilevanti per la variabile *S*: fede, gruppo etnico e regione di residenza. Queste variabili sono disponibili nel data set, e quindi siete lasciati liberi di far girare i diversi modelli.