

Econometria | 2023/2024

Lezione 6: Distorsioni da variabili omesse

Lezione 7: Stima dei parametri

Giuseppe Ragusa

<https://gragusa.org>

Roma, marzo 2024



Struttura

- Distorsione da variabili omesse
- Causalità e analisi di regressione
- Regressione multipla e OLS
- Misure di adattamento
- Distribuzione campionaria dello stimatore OLS

La distorsione da variabili omesse (Paragrafo 6.1)

- L'errore u_i contiene variabili, che influenzano Y_i ma non sono incluse regressione.
- In applicazioni economiche sono tante le variabili che influenzano la variabile dipendente.
- Spesso l'omissione di queste variabili può portare a una distorsione dello stimatore OLS.

La distorsione da variabili omesse (continua)

La distorsione dello stimatore OLS che si verifica a seguito di un fattore, o variabile, omesso è detta **distorsione da variabile omessa**

- Affinché si verifichi tale distorsione, la variabile omessa, Z_i , deve soddisfare queste due condizioni:
 1. Z_i è una determinante di Y_i (cioè Z_i è parte di u_i); e
 2. Z_i è correlata con X_i (cioè $corr(Z_i, X_i) \neq 0$)

La distorsione da variabili omesse (continua)

La percentuale di bambini non madrelingua inglese $elpct$ influisce sui punteggi nei test standardizzati

- $elpct$ è un determinante di $testscr$

Le comunità di immigrati tendono ad essere meno abbienti e quindi hanno budget scolastici inferiori e probabilmente più studenti per insegnati

- $elpct$ è correlata con str

Di conseguenza, $\hat{\beta}_1$ è distorto. In quale direzione:

- che cosa suggerisce il buon senso?
- Se il buon senso vi fa difetto, c'è una formula

La distorsione da variabili omesse (continua)

Formula per la distorsione da variabili omesse

Dall'equazione

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \approx \frac{\frac{1}{n} \sum_{i=1}^n \nu_i}{\sigma_X^2}$$

dove $\nu_i = (X_i - \mu_X)u_i$.

Sotto la prima assunzione dei minimi quadrati:

$$E[(X_i - \mu_X)u_i] = \text{cov}(X_i, u_i) = 0.$$

Ma se $E[(X_i - \mu_X)u_i] = \text{cov}(X_i, u_i) = \sigma_{Xu} \neq 0$?

La distorsione da variabili omesse (continua)

Sotto le assunzioni dei minimi quadrati #2 e #3 (cioè anche se la **prima** assunzione dei minimi quadrati **non è vera**)

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_1 - \beta_1 \xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2} = \frac{\sigma_u}{\sigma_X} \times \frac{\sigma_{Xu}}{\sigma_X \sigma_u} = \frac{\sigma_u}{\sigma_X} \rho_{Xu}$$

dove $\rho_{Xu} = \text{corr}(X, u)$.

Se vale la prima assunzione, allora $\rho_{Xu} = 0$, altrimenti abbiamo la distorsione $\frac{\sigma_u}{\sigma_X} \rho_{Xu}$.

Formula della distorsione da variabili omesse:

$$\hat{\beta}_1 \approx \beta_1 + \frac{\sigma_u}{\sigma_X} \rho_{Xu}$$

dove

- ρ_{Xu} è la correlazione fra u e X ,
- σ_u standard deviation di u
- σ_X standard deviation di X

Se una variabile omessa Z è **contemporaneamente**:

- una determinante di Y (cioè se è contenuta in u); e
- correlata con X , allora $\rho(Xu) \neq 0$ e lo stimatore OLS $\hat{\beta}_1$ è **distorto** e **inconsistente**.

Dati California

Per esempio, i distretti scolastici con pochi studenti non di madrelingua

1. ottengono punteggi migliori nei test standardizzati
2. hanno classi più piccole (budget più elevati)

Ignorando i non madrelingua sovrastimeremmo l'effetto di str su $testscr$

Intuizione: oltre a catturare il proprio effetto, str cattura l'effetto di $elpct$ e visto che quando str è più grande più alto è il numero di non madrelingua il coefficiente che stimiamo sarà **distorto positivamente**.

Dati California

$$testscr_i = \beta_0 + \beta_1 str_i + u_i$$

```

OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  698.93      10.364   67.44 < 2.2e-16 ***
str          -2.28       0.519   -4.39 1.4467e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 18.5  Adj. R2: 0.04897

```

$$testscr_i = \alpha_0 + \alpha_1 elpct_i + e_i$$

```

OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  664.739     0.9740   682.5 < 2.2e-16 ***
elpct       -0.671     0.0321  -20.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 14.6  Adj. R2: 0.413496

```

- *elpct* e' correlato con *str*

```

      str elpct
str  1.000 0.267
elpct 0.267 1.000

```

Dati California

Possiamo calcolare la direzione della distorsione in $\hat{\beta}_1$.

- Assumiamo che $u = \gamma \times \text{elpct} + \varepsilon$ dove ε contiene altri fattori oltre a elpct e str che influenzano testscr che assumiamo non correlati con str

Quindi

$$\rho(Xu) = \text{correlazione fra elpct e str} = \gamma \times \text{corr}(\text{str}, \text{elpct})$$

- $\gamma < 0$ (perchè elpct è negativamente associato con testscr)
- $\text{corr}(\text{str}, \text{elpct}) > 0$ (scuole con str più alto hanno elpct più alto)

Quindi

$$\hat{\beta}_1 - \beta_1 < 0$$

La distorsione è negativa: la nostra stima ottenuta omettendo elpct è più piccola (nel nostro caso più negativa) di quella effettiva.

Causalità e analisi di regressione

- L'esempio dei punteggi nei test/str e percentuale di studenti non di madrelingua mostra che, se una variabile omessa soddisfa le due condizioni della distorsione da variabili omesse, allora lo stimatore OLS nella regressione che omette tale variabile è distorto e inconsistente.
- Perciò, anche se n è grande, $\hat{\beta}_1$ non sarà vicino a β_1 .
- Ciò fa sorgere una domanda più profonda: come definiamo β_1 ?

Ovvero, che cosa vogliamo stimare, precisamente, quando eseguiamo una regressione?

L'effetto **causale**!

Esperimento sulla dimensione delle classi:

- Si immagina un esperimento controllato casualizzato ideale per misurare l'effetto sui punteggi nei test della riduzione di STR
- In tale esperimento gli studenti sarebbero assegnati casualmente alle classi, che avrebbero dimensioni diverse.
- Poiché gli studenti sono assegnati casualmente, tutte le loro caratteristiche (e quindi gli u_i) sarebbero distribuiti in modo indipendente da str_i .
- Quindi, $E(u_i|str_i) = 0$ – cioè la prima assunzione dei minimi quadrati vale in un esperimento controllato casualizzato.

Evitare la distorsione da variabili omesse

1. Eseguire un esperimento controllato casualizzato in cui il trattamento (STR) sia assegnato casualmente: allora $elpct$ è ancora un determinante di $testscr$, ma $elpct$ è incorrelato con STR . (Questa soluzione è raramente praticabile.)
2. Stimare le regressioni in sottocampioni di str e $elpct$ in modo che all'interno di ogni gruppo tutte le classi abbiano lo stesso $elpct$ (ma presto si esauriranno i dati, e che dire di altri determinanti come il reddito familiare e il livello di istruzione dei genitori?)
3. Usare una regressione in cui la variabile omessa ($elpct$) non è più omessa: includere $elpct$ come regressore aggiuntivo in una regressione multipla.

Il modello di regressione multipla (Par. 6.2)

Si consideri il caso di due regressori:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{ki} + u_i$$

- Y è la **variabile dipendente**
- X_1, X_2, \dots, X_K sono le K **variabili indipendenti** (**regressori**)
- $Y_i, X_{1i}, \dots, X_{ki}$ denotano l' i -esima osservazione
- β_0 = intercetta della popolazione ignota
- β_1 = effetto su Y di una variazione in X_1 , tenendo X_2, \dots, X_K costanti
- β_2 = effetto su Y di una variazione in X_2 , tenendo tutte le altre X costanti
- u_i = errore di regressione (fattori omessi)

Modello a tre variabili

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

dove Y_i è la variabile dipendente e X_1, X_2 sono i regressori, u è il termine di errore.

- β_0 è l'intercetta: l'effetto medio sulla Y di tutte le variabili escluse dal modello (valore medio della Y quando sia X_1 che X_2 sono zero).
- β_1 e β_2 sono i **coefficienti parziali di regressione**.
- I coefficienti parziali di regressione vanno interpretati alla luce della condizione di **Ceteris paribus**

Coefficienti parziali di regressione: β_1

- β_1 misura il **cambiamento** nel valore medio di Y , $E(Y)$, dovuto al cambiamento di una unità di X_1 tenendo il valore di X_2 costante.

- Retta di regressione della popolazione **prima** della variazione:

$$E(Y_i | X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Retta di regressione della popolazione **dopo** la variazione:

$$E(Y_i | X_{1i} = x_1 + 1, X_{2i} = x_2) = \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2$$

- In altre parole, β_1 misura l'effetto della variazione di un'unità di X_1 sul valore medio di Y , al netto di qualsiasi effetto che X_2 può avere sulla media di Y (perché X_2 è costante).

Coefficienti parziali di regressione: β_2

- β_2 misura il **cambiamento** nel valore medio di Y , $E(Y)$, dovuto al cambiamento di una unità di X_2 tenendo il valore di X_1 costante.

- Retta di regressione della popolazione **prima** della variazione:

$$E(Y_i | X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Retta di regressione della popolazione **dopo** la variazione:

$$E(Y_i | X_{1i} = x_1, X_{2i} = x_2 + 1) = \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + 1)$$

- In altre parole, β_2 misura l'effetto di una variazione di un'unità di X_2 sul valore medio di Y , al netto di qualsiasi effetto che X_1 può avere sulla media di Y .

Minimi quadrati ordinari (OLS)

$$\min_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \right)^2$$

Le condizioni del primo ordine sono:

$$\frac{\partial \sum_{i=1}^n \hat{u}_i^2}{\partial \hat{\beta}_0} = 2 \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \right) (-1) = 0$$

$$\frac{\partial \sum_{i=1}^n \hat{u}_i^2}{\partial \hat{\beta}_1} = 2 \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \right) (-X_{1i}) = 0$$

$$\frac{\partial \sum_{i=1}^n \hat{u}_i^2}{\partial \hat{\beta}_2} = 2 \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \right) (-X_{2i}) = 0$$

Stimatori OLS

Risolvendo il precedente sistema (usiamo \tilde{Y} , \tilde{X} per definire le variabili in deviazione dalla loro media):

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

$$\hat{\beta}_1 = \frac{(\sum_{i=1}^n \tilde{Y}_i \tilde{X}_{1i})(\sum_{i=1}^n \tilde{X}_{2i}^2) - (\sum_{i=1}^n \tilde{Y}_i \tilde{X}_{2i})(\sum_{i=1}^n \tilde{X}_{1i} \tilde{X}_{2i})}{(\sum_{i=1}^n \tilde{X}_{1i}^2)(\sum_{i=1}^n \tilde{X}_{2i}^2) - (\sum_{i=1}^n \tilde{X}_{1i} \tilde{X}_{2i})^2}$$

$$\hat{\beta}_2 = \frac{(\sum_{i=1}^n \tilde{Y}_i \tilde{X}_{2i})(\sum_{i=1}^n \tilde{X}_{1i}^2) - (\sum_{i=1}^n \tilde{Y}_i \tilde{X}_{1i})(\sum_{i=1}^n \tilde{X}_{1i} \tilde{X}_{2i})}{(\sum_{i=1}^n \tilde{X}_{1i}^2)(\sum_{i=1}^n \tilde{X}_{2i}^2) - (\sum_{i=1}^n \tilde{X}_{1i} \tilde{X}_{2i})^2}$$

Stimatore OLS $\hat{\beta}_1$

Dato che:

$$\sigma_{yX_1} = \frac{1}{n-1} \sum_{i=1}^n \tilde{Y}_i \tilde{X}_{1i}, \quad \sigma_{yX_2} = \frac{1}{n-1} \sum_{i=1}^n \tilde{Y}_i \tilde{X}_{2i}$$

$$\sigma_{X_1}^2 = \frac{1}{n-1} \sum_{i=1}^n \tilde{X}_{1i}^2, \quad \sigma_{X_2}^2 = \frac{1}{n-1} \sum_{i=1}^n \tilde{X}_{2i}^2, \quad \text{e } \sigma_{X_1X_2} = \frac{1}{n-1} \sum_{i=1}^n \tilde{X}_{1i} \tilde{X}_{2i}$$

possiamo riscrivere $\hat{\beta}_1$ come

$$\hat{\beta}_1 = \frac{\sigma_{yX_1} \sigma_{X_2}^2 - \sigma_{yX_2} \sigma_{X_1X_2}}{\sigma_{X_1}^2 \sigma_{X_2}^2 - (\sigma_{X_1X_2})^2}$$

Stimatore OLS $\hat{\beta}_1$ nel modello a tre variabili

$$\hat{\beta}_1 = \frac{\sigma_{yX_1}\sigma_{X_2}^2 - \sigma_{yX_2}\sigma_{X_1X_2}}{\sigma_{X_1}^2\sigma_{X_2}^2 - (\sigma_{X_1X_2})^2}$$

Dividendo numeratore e denominatore per $\sigma_{X_1}^2$ e $\sigma_{X_2}^2$ otteniamo

$$\hat{\beta}_1 = \frac{\hat{\beta}_1^* - \hat{\beta}_{21}\hat{\beta}_2^*}{1 - \hat{\beta}_{21}\hat{\beta}_{12}}$$

dove:

- $\hat{\beta}_1^*$ è il coefficiente di una regressione di Y su X_1
- $\hat{\beta}_2^*$ è il coefficiente di una regressione di Y su X_2
- $\hat{\beta}_{21}$ è il coefficiente di una regressione di X_2 su X_1
- $\hat{\beta}_{12}$ è il coefficiente di una regressione di X_1 su X_2

Esempio: test della California

$$testscr_i = \beta_0 + \beta_1 str + u_i$$

```
1 lm1 <- feols(testscr~str, data = Caschool, vcov = "hetero")
2 lm1
```

```
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  698.93      10.364   67.44 < 2.2e-16 ***
str          -2.28       0.519   -4.39 1.4467e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 18.5  Adj. R2: 0.04897
```

$$testscr_i = \beta_0 + \beta_1 str + \beta_2 english + u_i$$

```
1 lm2 <- feols(testscr~str+elpct, data = Caschool, vcov = "hetero")
2 lm2
```

```
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  686.03      8.728   78.60 < 2.2e-16 ***
str          -1.10       0.433   -2.54 0.011309 *
elpct        -0.65       0.031  -20.94 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 14.4  Adj. R2: 0.42368
```

Misure di bontà dell'adattamento (Paragrafo 6.4)

- SER = standard deviation di \hat{u}_i (con correzione per gr. lib.)

$$SER = \sqrt{\frac{1}{n - k} \sum_{i=1}^n \hat{u}_i^2}$$

- \bar{R}^2 : R^2 aggiustato con una correzione per gradi di libertà che corregge per l'incertezza della stima

$$\bar{R}^2 = 1 - (1 - R^2) \frac{(n - 1)}{(n - k)}$$

Misure di bontà dell'adattamento nella regressione multipla

```
1 lm3 <- feols(testscr~str+elpct, data = Caschool, vcov = "hetero")  
2 lm3
```

```
OLS estimation, Dep. Var.: testscr  
Observations: 420  
Standard-errors: Heteroskedasticity-robust  
      Estimate Std. Error t value Pr(>|t|)  
(Intercept)  686.03      8.728   78.60 < 2.2e-16 ***  
str           -1.10      0.433   -2.54  0.011309 *  
elpct        -0.65      0.031  -20.94 < 2.2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
RMSE: 14.4  Adj. R2: 0.42368
```

Le assunzioni dei MQO (Paragrafo 6.5)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{ki} + u_i$$

1. La distribuzione di u condizionata alle X ha media nulla , cioè
 $E(u_i | X_{1i} = x_1, \dots, X_{ki} = x_k) = 0$.
2. $X_{1i}, \dots, X_{Ki}, Y_i, i = 1, \dots, n$ sono i.i.d.
3. Gli outlier sono improbabili: X_1, \dots, X_K e Y hanno momenti quarti finiti
 $E(X_{1i}^4) < \infty, E(X_{Ki}^4) < \infty, E(Y_i^4) < \infty$.
4. Non vi è collinearità perfetta.

Assunzione 1: la media condizionata di u date le X incluse è zero.

$$E(u_i | X_{1i} = x_1, \dots, X_{Ki} = x_k) = 0$$

Ha la stessa interpretazione del caso della regressione con un singolo regressore.

- La non validità di questa condizione porta a distorsione da variabili omesse ; nello specifico , se una variabile omessa
 - appartiene all'equazione (cioè è in u)
 - è correlata con una X inclusa

allora questa condizione non vale e vi è distorsione da variabili omesse.

- La soluzione migliore, se possibile, è quella di includere la variabile omessa nella regressione.
- Una seconda soluzione, correlata alla precedente, è quella di includere una variabile che controlli per la variabile omessa (cfr . Capitolo 7).

Assunzione 2: $X_{1i}, \dots, X_{ki}, Y_i, i = 1, \dots, n$
sono i.i.d.

È soddisfatta automaticamente se i dati sono raccolti mediante campionamento casuale semplice.

Assunzione 3: gli outlier sono rari (momenti quarti finiti)

È la stessa assunzione descritta per il caso di un regressore singolo.

Come in quel caso, l'OLS può essere sensibile agli outlier, perciò occorre controllare i dati (diagramma nuvola!) per assicurarsi che non vi siano valori “impazziti” (refusi o errori di codifica).

Assunzione 4: Non vi è collinearità perfetta

La collinearità perfetta si ha quando uno dei regressori è funzione lineare esatta degli altri.

Esempio: si supponga di includere due volte STR, per errore:

```
1 lm2 <- feols(testscr~str+str, data = Caschool, vcov = "hetero")
2 lm2
```

```
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  698.93      10.364   67.44 < 2.2e-16 ***
str          -2.28       0.519   -4.39 1.4467e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 18.5  Adj. R2: 0.04897
```

Collinearità perfetta

La **collinearità** perfetta si ha quando uno dei regressori è funzione lineare esatta degli altri.

- Nella regressione precedente, β_1 è l'effetto su *testscr* di una variazione unitaria in *STR*, tenendo *STR* costante (???)
- Torneremo alla collinearità perfetta (e imperfetta) tra breve, con altri esempi ...
- Con le assunzioni dei minimi quadrati , ora possiamo derivare la distribuzione campionaria di $\hat{\beta}_1, \dots, \hat{\beta}_K$.

La distribuzione degli stimatori OLS nella regressione multipla (Paragrafo 6.6)

Sotto le quattro assunzioni dei minimi quadrati:

1. La distribuzione campionaria di $\hat{\beta}_1$ ha media β_1
2. $var(\hat{\beta}_1)$ è **inversamente proporzionale** a n .
3. Al di là di media e varianza, la distribuzione esatta (n -finita) di $\hat{\beta}_1$ è molto complessa; ma per n grande:
 - a. $\hat{\beta}_1$ è consistente: $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (legge dei grandi numeri)
 - b. $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{var(\hat{\beta}_1)}}$ è approssimata da una $N(0, 1)$ (TLC)

Queste proprietà valgono per $\hat{\beta}_1, \dots, \hat{\beta}_k$.

Collinearità perfetta e imperfetta (Par. 6.7)

La **collinearità** si ha quando uno dei regressori è una funzione lineare esatta degli altri:

Due casi sono possibili:

1. **Collinearità perfetta** se $\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_K X_K = 0$:

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \dots - \frac{\lambda_K}{\lambda_2} X_{Ki}$$

2. **Collinearità imperfetta** se $\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_K X_K + v_i = 0$:

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \dots - \frac{\lambda_K}{\lambda_2} X_{Ki} - \frac{1}{\lambda_2} v_i$$

Conseguenze della collinearità

1. se la collinearità è **perfetta**:

a. i coefficienti della regressione sono **indeterminati**

b. gli standard error sono **infiniti**

2. se la collinearità è **imperfetta**:

a. i coefficienti della regressione sono **determinati**

b. gli standard error sono **grandi** (i coefficienti non possono essere stimati con precisione)

Conseguenze della collinearità

Consideriamo un esempio nel modello con X_1 e X_2 in deviazione dalla media:

$$\tilde{Y}_i = \beta_2 \tilde{X}_{1i} + \beta_3 \tilde{X}_{2i} + u_i$$

e quindi:

$$\hat{\beta}_1 = \frac{(\sum \tilde{Y}_i \tilde{X}_{1i})(\sum \tilde{X}_{2i}^2) - (\sum \tilde{Y}_i \tilde{X}_{2i})(\sum \tilde{X}_{1i} \tilde{X}_{2i})}{(\sum \tilde{X}_{1i}^2)(\sum \tilde{X}_{2i}^2) - (\sum \tilde{X}_{1i} \tilde{X}_{2i})^2}$$
$$\hat{\beta}_2 = \frac{(\sum \tilde{Y}_i \tilde{X}_{2i})(\sum \tilde{X}_{1i}^2) - (\sum \tilde{Y}_i \tilde{X}_{1i})(\sum \tilde{X}_{1i} \tilde{X}_{2i})}{(\sum \tilde{X}_{1i}^2)(\sum \tilde{X}_{2i}^2) - (\sum \tilde{X}_{1i} \tilde{X}_{2i})^2}$$

Assumiamo che $X_2 = \lambda_1 X_1$, quindi $(X_2 - \bar{X}_2) = \lambda_1 (X_1 - \bar{X}_1)$.

a. i coefficienti della regressione sono indeterminati

$$\hat{\beta}_1 = \frac{(\sum \tilde{Y}_i \tilde{X}_{1i})(\sum \lambda_1^2 \tilde{X}_{1i}^2) - (\sum \tilde{Y}_i \lambda_1 \tilde{X}_{1i})(\sum \tilde{X}_{1i} \lambda_1 \tilde{X}_{1i})}{(\sum \tilde{X}_{1i}^2)(\sum \lambda_1^2 \tilde{X}_{1i}^2) - (\sum \tilde{X}_{1i} \lambda_1^2 \tilde{X}_{1i})^2} = \frac{0}{0}$$

Perché otteniamo questo risultato?

- Ricordiamo il significato di $\hat{\beta}_1$: dà la variazione media in Y dovuto al cambiamento di una unità di X_1 **tenendo costante** X_2 .
- Ma se X_1 e X_2 sono perfettamente collineari non c'è modo di tenere costante X_2 quando X_1 cambia!

Collinearità perfetta (continua)

- La collinearità perfetta solitamente riflette un errore nelle definizioni dei regressori, o una stranezza nei dati
- Se avete collinearità perfetta, il software statistico ve lo farà sapere – bloccandosi, o mostrando un messaggio di errore, o “scaricando” arbitrariamente una delle variabili
- La soluzione alla collinearità perfetta consiste nel modificare l’elenco di regressori.

b. gli standard error sono grandi

La collinearità imperfetta implica che uno o più dei coefficienti di regressione sarà stimato in modo impreciso.

- L'idea: il coefficiente di X_1 è l'effetto di X_1 tenendo costante X_2 ; ma se X_1 e X_2 sono **altamente correlati**, vi è una ridottissima variazione in X_1 quando X_2 è mantenuta costante – perciò i dati non contengono molte informazioni su ciò che accade quando X_1 cambia e X_2 no. In questo caso, la varianza dello stimatore OLS del coefficiente di X_1 sarà grande.
- La collinearità imperfetta (correttamente) genera grandi errori standard per uno o più dei coefficienti OLS.
- La matematica? Appendice 6.2

Variabili Dummy

- Le variabili dummy sono usate in econometria per classificare i dati in categorie **mutualmente esclusive** relativi a **qualità** o **attributi** (come sesso, nazionalità, regioni geografiche, ecc).
- Per quantificare l'effetto di questi attributi possiamo costruire delle variabili artificiali che assumono valore 1 se la caratteristica è presente o 0 se è assente.
- Non è necessario che la variabile dummy assuma valori 0 o 1. Ad esempio potrebbe assumere valori 1 e 3, ogni altro valore che deriva da una trasformazione lineare come $Z = a + b * D$, dove D è la solita dummy 0-1 e a, b sono costanti.

Esempio: Current Population Surveys 1985

- Supponiamo di voler stimare la differenza nel salario orario dei lavoratori distinguendo per maschi e femmine.
- Stimiamo il modello:

$$wage_i = \beta_1 D_{fi} + \beta_2 D_{mi} + u_i,$$

dove $wage_i$: salario dell'individuo i ,

$$D_{fi} = \begin{cases} 1 & \text{se l'individuo } i \text{ è donna} \\ 0 & \text{se l'individuo } i \text{ è uomo} \end{cases},$$

$$D_{mi} = \begin{cases} 1 & \text{se l'individuo } i \text{ è uomo} \\ 0 & \text{se l'individuo } i \text{ è donna} \end{cases}.$$

Esempio: Current Population Surveys 1985

- Cosa ci dice il modello? Assumiamo che le proprietà degli OLS siano valide.
 - Il valore medio del salario orario delle donne è:

$$E(wage_i | D_{fi} = 1, D_{mi} = 0) = \beta_1$$

- Il valore medio del salario orario per gli uomini è:

$$E(wage_i | D_{fi} = 0, D_{mi} = 1) = \beta_2$$

Esempio: Current Population Surveys 1985

```
1 library(dplyr)
2 library(wooldridge)
3 data('cps78_85')
4
5 cps1985 <- cps78_85 |> filter(year==85) |> mutate(male=(1-female), wage=exp(lwage))
6
7 feols(wage~female+male-1, data = cps1985, vcov = "hetero") ## con -1 rimuoviamo l'intercetta
```

OLS estimation, Dep. Var.: wage

Observations: 534

Standard-errors: Heteroskedasticity-robust

	Estimate	Std. Error	t value	Pr(> t)
female	7.88	0.301	26.1	< 2.2e-16 ***
male	9.99	0.311	32.1	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

RMSE: 5.02461 Adj. R2: 0.03858

Trappola delle Dummy

- Nel modello precedente abbiamo incluso le due dummy D_f e D_m ma abbiamo rimosso la costante. Perché?
- Supponiamo di riscrivere il modello come:

$$wage_i = \beta_0 + \beta_1 D_{fi} + \beta_2 D_{mi} + u_i$$

Questa regressione non può essere stimata perché le due dummy e la costante sono **perfettamente collineari**.

```
1 lm2 <- feols(wage~female+male, data = cps1985, vcov = "hetero")
2 lm2
```

```
OLS estimation, Dep. Var.: wage
Observations: 534
Standard-errors: Heteroskedasticity-robust
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)    9.99      0.311    32.14 < 2.2e-16 ***
female        -2.12      0.433    -4.89 1.3667e-06 ***
... 1 variable was removed because of collinearity (male)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 5.02461  Adj. R2: 0.04038
```

- La stima può essere effettuata: **(a)** rimuovendo la costante; **(b)** rimuovendo una delle due dummy.

2. Rimuovere una delle due dummy

- L'intercetta assume il valore delle categoria **omessa**.
- La dummy inclusa diventa il **differenziale** rispetto alla categoria omessa.

```
1 lm3 <- feols(wage~female+male-1, data = cps1985, vcov = "hetero")
2 lm3
```

```
OLS estimation, Dep. Var.: wage
Observations: 534
Standard-errors: Heteroskedasticity-robust
      Estimate Std. Error t value Pr(>|t|)
female    7.88    0.301    26.1 < 2.2e-16 ***
male     9.99    0.311    32.1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 5.02461  Adj. R2: 0.03858
```

```
1 lm4 <- feols(wage~female, data = cps1985, vcov = "hetero")
2 lm4
```

```
OLS estimation, Dep. Var.: wage
Observations: 534
Standard-errors: Heteroskedasticity-robust
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.99    0.311    32.14 < 2.2e-16 ***
female     -2.12    0.433    -4.89 1.3667e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 5.02461  Adj. R2: 0.04038
```

2. Rimuovendo una delle due dummy

```
1 lm5 <- feols(wage~female+male-1, data = cps1985, vcov = "hetero")
2 lm5
```

```
OLS estimation, Dep. Var.: wage
Observations: 534
Standard-errors: Heteroskedasticity-robust
      Estimate Std. Error t value Pr(>|t|)
female      7.88      0.301    26.1 < 2.2e-16 ***
male        9.99      0.311    32.1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 5.02461  Adj. R2: 0.03858
```

```
1 lm6 <- feols(wage~male, data = cps1985, vcov = "hetero")
2 lm6
```

```
OLS estimation, Dep. Var.: wage
Observations: 534
Standard-errors: Heteroskedasticity-robust
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.88      0.301    26.13 < 2.2e-16 ***
male         2.12      0.433     4.89 1.3667e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 5.02461  Adj. R2: 0.04038
```

Trappola delle Dummy

- Quindi se la variabile qualitativa ha m categorie:
 - si può introdurre m dummy e rimuovere la costante
 - si può introdurre $m - 1$ dummy e tenere la costante.
- Usando il primo o il secondo metodo non cambia i risultati.
 - il significato dei coefficienti è però diverso: l'interpretazione è rispetto alla categoria omessa.