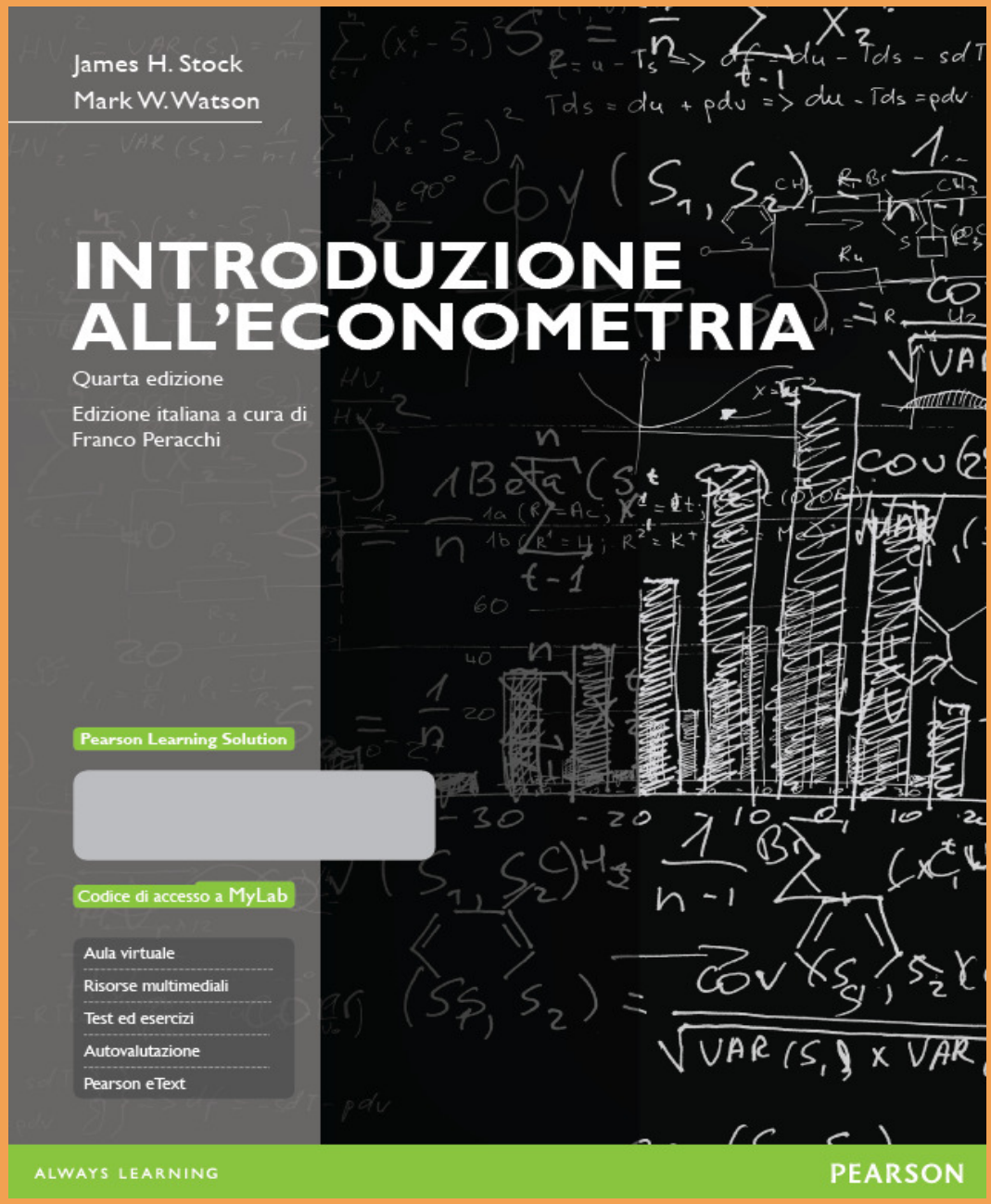
Capitolo 6

Regressione lineare con regressori multipli



Le assunzioni dei minimi quadrati per la regressione multipla (Paragrafo 6.5)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1,\ldots,n$$

1. La distribuzione di u condizionata alle X ha media nulla, cioè $E(u_i|X_{1i} = x_1,\ldots, X_{ki} = x_k) = 0$.

2. $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, sono i.i.d.

3. Gli outlier sono improbabili: $X_1,\ldots, X_k$, e Y hanno momenti quarti: $0<E(Y_i^4) < \infty, 0< E(X_{1i}^4) < \infty,\ldots, 0< E(X_{ki}^4) < \infty$.

4. Non vi è collinearità perfetta.

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

$$E(u|X_1 = x_1,\ldots, X_k = x_k) = 0$$

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

1. appartiene all'equazione (cioè è in *u*) ***e***
2. 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: $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, 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 (diagrammi a 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:

```
regress testscr str str, robust
Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
         str |  (dropped)
       _cons |   698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

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

- Nella regressione precedente, β_1 è l'effetto su *TestScore* 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* $\hat{\beta}_1$, $\hat{\beta}_2$,…, $\hat{\beta}_k$.

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

Sotto le quattro assunzioni dei minimi quadrati,

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

Concettualmente, non vi è nulla di nuovo!

Collinearità perfetta e imperfetta (Paragrafo 6.7)

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

Altri esempi di collinearità perfetta

1. Dal caso precedente: includete *STR* due volte,
2. Eseguite la regressione di *TestScore* su una costante, *D*, e *B*, dove: $D_i = 1$ se $STR \leq 20$, = 0 altrimenti; $B_i = 1$ se $STR > 20$, = 0 altrimenti, perciò $B_i = 1 - D_i$ e vi è collinearità perfetta.
3. Ci sarebbe collinearità perfetta se l'intercetta (costante) fosse esclusa da questa regressione? Questo esempio è un caso speciale di…

La trappola delle variabili dummy

Si supponga di avere un insieme di più variabili binarie (dummy) che sono mutuamente esclusive ed esaustive – cioè esistono più categorie e ogni osservazione ricade in una di esse e solo in una (Matricole, Studenti del secondo anno, Junior, Senior, Altri). Se includete tutte queste variabili dummy *e* una costante, avrete collinearità perfetta – si parla talvolta di ***trappola delle variabili dummy***.

- *Perché vi è collinearità perfetta in questo caso*?
- *Soluzioni alla trappola delle variabili dummy*:

1. omettere uno dei gruppi (per esempio Senior), oppure
2. omettere l'intercetta

- *Quali sono le implicazioni di (1) o (2) per l'interpretazione dei coefficienti?*

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.

Collinearità imperfetta

La collinearità imperfetta è ben diversa dalla collinearità perfetta, nonostante la somiglianza dei nomi.

La ***collinearità imperfetta*** si verifica quando due o più regressori sono altamente correlati.

•Perché si usa il termine "collinearità"? Se due regressori sono altamente correlati, allora il loro diagramma a nuvola apparirà molto simile a una retta – sono "co-lineari" – ma a meno che la correlazione sia esattamente ± 1, tale collinearità è imperfetta.

Collinearità imperfetta (continua)

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? Cfr. il volume stampato, Appendice 6.2

Prossimo argomento: test di ipotesi e intervalli di confidenza...