METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 10: REGRESSIONE LINEARE (SECONDA PARTE)

A.A. 2024-2025

Prof. Martina Vettoretti

OUTLINE

- Gestione del problema della multicollinearità
- Variabili qualitative
- Confronto tra modelli in competizione
  - Adjusted R^2
  - F test
  - Indici di parsimonia

COLLINEARITA'

- **Collinearità**: forte relazione lineare tra due o più variabili esplicative X_j

  → le variabili esplicative non sono tra loro linearmente indipendenti.

  - **Collinearità tra 2 variabili**: due variabili fortemente correlate tra loro
    - Collinearità perfetta:

$$X_i = a \cdot X_j + b \quad i \neq j, a \neq 0$$

    - Collinearità imperfetta:

$$X_i \cong a \cdot X_j + b \quad i \neq j, a \neq 0$$

  - **Multicollinearità**: una variabile si spiega con una combinazione lineare di altre variabili
    - Multicollinearità perfetta:

$$X_i = a_{j_1} \cdot X_{j_1} + a_{j_2} \cdot X_{j_2} + \cdots + a_{j_m} \cdot X_{j_m} + b \quad i \neq j_k, a_{j_k} \neq 0, k = 1, \dots, m$$

    - Multicollinearità imperfetta:

$$X_i \cong a_{j_1} \cdot X_{j_1} + a_{j_2} \cdot X_{j_2} + \cdots + a_{j_m} \cdot X_{j_m} + b \quad i \neq j_k, a_{j_k} \neq 0, k = 1, \dots, m$$

IL PROBLEMA DELLA COLLINEARITA'

- In presenza di collinearità o multicollinearità perfetta, la stima dei coefficienti della regressione lineare multipla non è possibile.
  - $\boldsymbol{X}^T\boldsymbol{X}$ è matrice singolare, $\det(\boldsymbol{X}^T\boldsymbol{X}) = 0$
  - Non possiamo calcolare $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ e quindi la soluzione $\widehat{\boldsymbol{\beta}}$ non esiste.
- In presenza di collinearità o multicollinearità imperfetta, la stima dei coefficienti della regressione lineare multipla è problematica.
  - Difficoltà nello stimare l'effetto delle singole variabili (stimare correttamente i β_j), poiché è **difficile separare il contributo delle variabili collineari**.
    - Ampia **incertezza** nelle stime $\hat{\beta}_j$
    - SE_j e CV_j grandi per le variabili collineari
    - Piccole variazioni nei dati possono far variare notevolmente le stime $\hat{\beta}_j$

SOLUZIONI AL PROBLEMA DI MULTICOLLINEARITA'

- Soluzione: identificare la variabili collineari con altre presenti nel dataset ed escluderle dal modello.
- Metodi per identificare e gestire situazioni di collinearità:
  1. Analisi degli autovalori della matrice $\boldsymbol{X}^T\boldsymbol{X}$
  2. Analisi di correlazione
  3. Metodo VIF (Variance Inflation Factor)

1. ANALISI DEGLI AUTOVALORI DELLA MATRICE $X^T X$

- Autovalori della matrice $(X^T X)^{-1}$: $\lambda_1, \lambda_2, \dots, \lambda_m$
- Se multicollinearità → $X^T X$ è quasi singolare → almeno un autovalore è molto piccolo.
- Numero di condizionamento della matrice $X^T X$:

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

  - $\lambda_{min} = \min(\lambda_1, \lambda_2, \dots, \lambda_m)$
  - $\lambda_{max} = \max(\lambda_1, \lambda_2, \dots, \lambda_m)$
- $\kappa < 100$ → no multicollinearità
- $100 \leq \kappa < 1000$ → multicollinearità contenuta → non facciamo nulla
- $\kappa > 1000$ → multicollinearità importante → bisogna intervenire

ESEMPIO

- Dati del caso di studio sulla predizione del diametro della componente acetabolare della protesi d'anca.
- Autovalori di $X^T X$: 0.75, 578, 7981, 23852, 94054, $2.37 \cdot 10^7$

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} = 3.15 \cdot 10^7$$

- Presente importante collinearità, quali sono le variabili coinvolte?

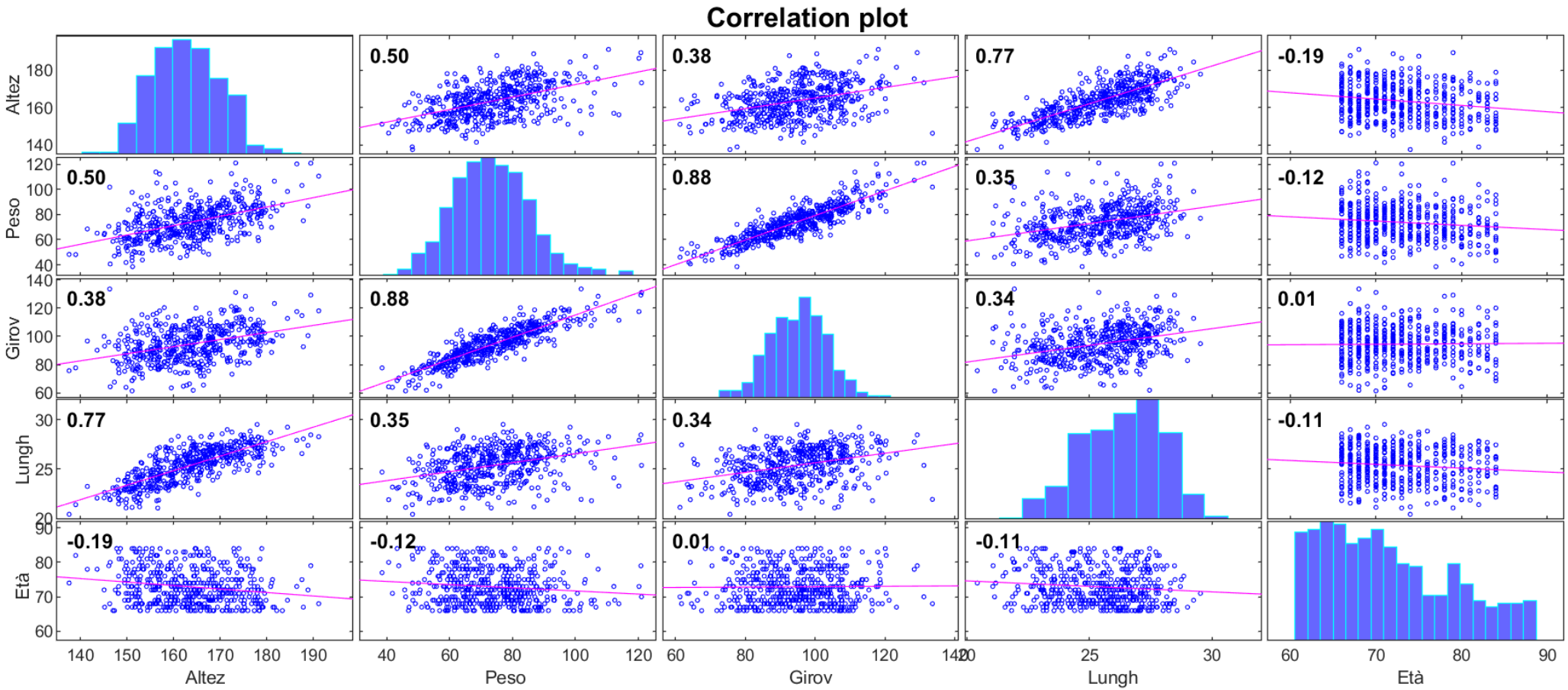
2. ANALISI DI CORRELAZIONE

1. Calcolo delle correlazioni lineari tra tutte le coppie di variabili esplicative.
2. Identificazione delle coppie aventi correlazione maggiore di una soglia $|r_{max}|$ (tipicamente $|r_{\max}| = 0.8$).
3. Per ciascuna di queste coppie, includere nel modello una sola variabile della coppia. Quale?
   - Quella che presenta minor correlazione con le altre variabili esplicative.
   - Quella che presenta minor numero di valori mancanti.
   - Quella più facile da raccogliere.

Domanda: secondo te questo approccio è in grado di rilevare situazioni di multicollinearità?

ESEMPIO

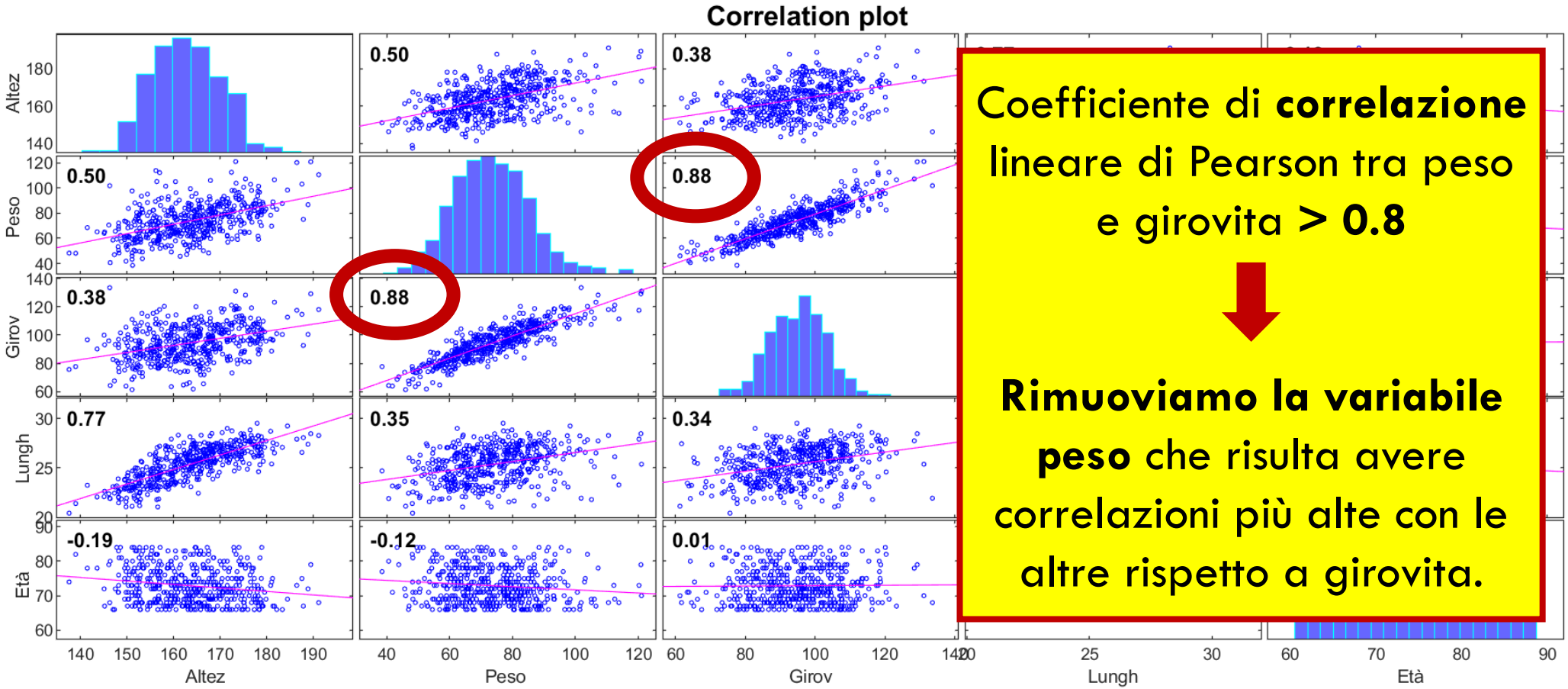
- Dati del caso di studio sulla predizione del diametro della componente acetabolare della protesi d'anca.



Quali variabili presentano un livello di correlazione critico?

ESEMPIO

- Dati del caso di studio sulla predizione del diametro della componente acetabolare della protesi d'anca.



Coefficiente di **correlazione** lineare di Pearson tra peso e girovita **> 0.8**

Rimuoviamo la variabile peso che risulta avere correlazioni più alte con le altre rispetto a girovita.

EFFETTO DELLA RIDUZIONE DELLA COLLINEARITA'

	Modello con variabile peso			Modello senza variabile peso		
Variabili	**Stime dei parametri** $\widehat{\beta}_j$	**Standard error** SE_j	**Coefficiente di variazione** CV_j	**Stime dei parametri** $\widehat{\beta}_j$	**Standard error** SE_j	**Coefficiente di variazione** CV_j
Intercetta	21.0441	1.6262	7.73%	18.6074	1.6543	8.89%
Altezza	0.0884	0.0120	13.63%	0.1212	0.0115	9.48%
Peso	0.0735	0.0109	14.89%	-	-	-
Girovita	**0.0069**	**0.0115**	**166.37%**	**0.0744**	**0.0058**	**7.75%**
Lunghezza piede	0.5155	0.0578	11.22%	0.4197	0.0132	13.94%
Età	-0.0111	0.0130	117.00%	-0.0310	0.0132	42.69%

Com'è variata l'incertezza sulle stime dei parametri dopo la rimozione della variabile peso?

3. METODO VIF (VARIANCE INFLATION FACTOR)

L'analisi di correlazione tra coppie di variabili non consente di rilevare situazioni di multicollinearità, cosa che invece ci consente di fare il metodo VIF.

Calcolo degli indici VIF (Variance Inflation Factor) per le variabili X_j:

- Identificare un modello di regressione lineare multipla che predice il valore di X_j usando come predittori le altre variabili esplicative.

$$X_j = \beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \cdots + \beta_m X_m + \varepsilon$$

- Calcolare il valore di R^2 per il modello di X_j: R_j^2
- Calcolare l'indice VIF di X_j come:

$$VIF_j = \frac{1}{1-R_j^2}$$

  - $VIF_j = 1$ $(R_j^2 = 0)$ → no multicollinearità, X_j linearmente indipendente dalle altre variabili esplicative.
  - $VIF_j > 1$ $(R_j^2 > 0)$ → X_j presenta un certo grado di dipendenza lineare dalle altre variabili, più alto è VIF_j maggiore è il grado di multicollinearità.

IL METODO VIF PER RIDURRE LA MULTICOLLINEARITÀ

1. Decidere una soglia T tale che:

   Se $VIF_j > T$ $\rightarrow$ il grado di multicollinearità di X_j è considerato importante

2. Calcolare l'indice VIF per ciascuna variabile esplicativa X_j
3. Se alcune variabili presentano indice VIF > T, escludere la variabile con indice VIF più alto.
4. Ricalcolare l'indice VIF per ciascuna delle variabili rimaste, usando le variabili a disposizione tranne quella rimossa allo step 3.
5. Se ci sono ancora variabili con indice VIF > T, rimuovere un'ulteriore variabile, quella con valore di VIF più alto.
6. …
7. L'algoritmo si ferma quando tutte le variabili rimaste hanno indice VIF < T.

Valori tipici per T sono 5 o 10.

ESEMPIO

- Dati del caso di studio sulla predizione del diametro della componente acetabolare della protesi d'anca.
- Applichiamo il metodo VIF e decidiamo di rimuovere le variabili con indice VIF > 5.
- Calcolo dell'indice VIF per tutte le variabili esplicative.
  - Esempio. L'indice VIF della variabile X_1 (altezza) si calcola come:

$$VIF_1 = 1/(1 - R_1^2)$$

dove R_1^2 è il coefficiente di determinazione del modello:

$$X_1 = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Variabili indipendenti:

- X_1: altezza [cm]
- X_2: peso [kg]
- X_3: girovita [cm]
- X_4: lunghezza del piede [cm]
- X_5: età [anni]

```
VIF di altezza: 3.1573
VIF di peso: 5.7709
VIF di girovita: 5.0969
VIF di lunghezza piede: 2.6481
VIF di età: 1.1033
```

Rimuoviamo la variabile peso e ricalcoliamo gli indici VIF

```
VIF di altezza: 2.6386
VIF di girovita: 1.1826
VIF di lunghezza piede: 2.4872
VIF di età: 1.0463
```

OK

VARIABILI QUALITATIVE – CASO BINARIO

Variabili qualitative possono essere inserite nel modello con un'opportuna codifica.

- **Variabili binarie:**
  - 2 soli valori possibili → rappresentati con i valori numerici 0 e 1.

VARIABILE IPERTENSIONE (Sì, No)	VARIABILE IPERTENSIONE CODIFICATA (1=Sì, 0=No)
Sì	1
No	0
No	0
No	0
Sì	1

VARIABILE CITTADINANZA (Italiana, Non italiana)	VARIABILE CITTADINANZA CODIFICATA (1=Italiana, 0=Non italiana)
Non italiana	0
Non italiana	0
Italiana	1
Non italiana	0
Italiana	1

VARIABILI QUALITATIVE – CASO N CATEGORIE

- **Variabili con più categorie:**
  - Variabile X con n>2 valori possibili → codificata con n-1variabili binarie, D_1 D_2, …, D_{n-1}, dette variabili dummy, che assumono valori 0 o 1.
  - Si sceglie una categoria di riferimento, supponiamo sia la n-esima (C_n).
    - $X = C_n \rightarrow D_j = 0 \; \forall \, j = 1, \dots, n-1$
    - $X = C_i \; i \neq n \rightarrow D_i = 1, D_j = 0 \; \forall \, j \neq i$

VARIABILE IMPIEGO (Non occupato, Lavoratore, Studente)	D1 - LAVORATORE (1=lavoratore, 0=non occupato o studente)	D2 - STUDENTE (1=studente, 0=lavoratore o non occupato)
Non occupato	0	0
Lavoratore	1	0
Lavoratore	1	0
Non occupato	0	0
Studente	0	1

Non occupato è categoria di riferimento.

ESERCIZIO

Codifica la variabile qualitativa STATO CIVILE avente 4 categorie, utilizzando 3 variabili dummy.

VARIABILE STATO CIVILE (Celibe/nubile, coniugato/a, vedovo/a, divorziato/a)	D_1 – celibe/nubile (1=celibe/nubile, 0=altrimenti)	D_2 – vedovo/a (1= vedovo/a, 0=altrimenti)	D_3 – divorziato/a (1= divorziato/a, 0= altrimenti)
Divorziato/a	0	0	1
Coniugato/a	0	0	0
Vedovo/a	0	1	0
Celibe/nubile	1	0	0
Divorziato/a	0	0	1
Coniugato/a	0	0	0
Celibe/nubile	1	0	0
Coniugato/a	0	0	0
Celibe/nubile	1	0	0

Categoria di riferimento: coniugato/a

ESEMPIO

- Dati del caso di studio sulla predizione del diametro della componente acetabolare della protesi d'anca.
- Aggiungiamo all'analisi 2 variabili qualitative:
  - X_6: sesso (femmina, maschio)
  - X_7: patologia (coxartrosi, frattura, necrosi)

VARIABILE SESSO ORIGINALE (maschio, femmina)	D_6 - VARIABILE SESSO CODIFICATA (1=maschio)
Maschio	1
Maschio	1
Femmina	0
Maschio	1
Femmina	0

VARIABILE PATOLOGIA ORIGINALE (coxartrosi, frattura, necrosi)	$D_{7,1}$ – FRATTURA (1=frattura, 0= altrimenti)	$D_{7,2}$ – NECROSI (1=necrosi, 0=altrimenti)
Coxartrosi	0	0
Necrosi	0	1
Coxartrosi	0	0
Frattura	1	0
Frattura	1	0

Coxartrosi è categoria di riferimento.

ESEMPIO

- Modello senza la variabile peso (X_2) e con l'aggiunta delle variabili sesso (D_6) e patologia ($D_{7,1}$, $D_{7,2}$).

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \boldsymbol{\beta_6} \cdot \boldsymbol{D_6} + \boldsymbol{\beta_{7,1}} \cdot \boldsymbol{D_{7,1}} + \boldsymbol{\beta_{7,2}} \cdot \boldsymbol{D_{7,2}}$$

Variabili	$\widehat{\boldsymbol{\beta}}_j$	$\boldsymbol{SE_j}$	$\boldsymbol{CV_j}$	$[\widehat{\boldsymbol{\beta}}_j - \boldsymbol{2SE_j}\ \ \widehat{\boldsymbol{\beta}}_j + \boldsymbol{2SE_j}]$	$\boldsymbol{z_j}^*$
Intercetta	15.7929	2.1028	13.31%	[11.5894 -19.9965]	7.5141
Altezza	0.1292	0.0120	9.29%	[0.1052 0.1533]	10.7611
Girovita	0.0776	0.0060	7.67%	[0.0657 0.0895]	13.0384
Lunghezza piede	0.4778	0.0622	13.03%	[0.3533 0.6023]	7.6769
Età	-0.0312	0.0133	42.50%	[-0.0577 -0.0047]	-2.3530
Sesso	-0.4939	0.2046	41.43%	[-0.9032 -0.0847]	-2.4137
Frattura	-0.2682	0.1578	58.85%	[-0.5838 0.0475]	-1.6992
Necrosi	0.1105	0.1592	144.04%	[-0.2078 0.4288]	0.6942

*Il valore critico con $\alpha = 0.05$ è 1.96.

DOMANDE

- Le variabili sesso e patologia hanno un impatto significativo sull'outcome assumendo un livello di significatività pari al 5%?
  - La variabile sesso ha un impatto significativo sull'outcome (|z-score| > 1.96)
  - La variabile patologia non ha un impatto significativo sull'outcome (|z-score| < 1.96) → possiamo pensare di escluderla dal modello?

- Come influisce la variabile sesso sull'outcome?
  - Il coefficiente associato a sesso è negativo. → negli individui di sesso maschile (D_6=1) il diametro della componente acetibolare è mediamente più basso rispetto agli individui di sesso femminile, se tutte le altre variabili sono costanti (a parità di altezza, girovita, lunghezza piede, età e patologia).

- Dopo aver rimosso la variabile peso e aver aggiunto le variabili sesso e patologia come è cambiato il ruolo della variabile età? Confrontare il modello di slide 19 con quello di slide 16 della parte 9.
  - Dopo aver rimosso la variabile peso dal modello, la variabile età diventa significativamente legata all'outcome. Il suo coefficiente è negativo → a parità di altezza, girovita, lunghezza piede, sesso e patologia, il diametro medio della componente acetibolare risulta maggiore per i soggetti più giovani.

CONFRONTO TRA MODELLI IN COMPETIZIONE

- Problema: abbiamo diversi modelli candidati per descrivere i nostri dati (es. considerando diversi set di variabili indipendenti) e vogliamo capire qual è il modello più adatto.
- Il modello più adatto sarà il modello che descrive sufficientemente bene i dati utilizzando il minor numero possibile di variabili → concetto di **parsimonia**.
- Importante **limitare la complessità del modello** evitando di tenere nel modello variabili non utili per la predizione dell'outcome.
  - Modelli complessi sono difficili da interpretare e da utilizzare nella pratica.
  - Più parametri dobbiamo stimare, maggiore sarà l'incertezza delle stime.

R^2 NON E' UNA BUONA METRICA PER IL CONFRONTO TRA MODELLI

Come facciamo a capire qual è il modello migliore?

- Potrei guardare l'R^2 e prendere il modello con R^2 maggiore.
  - Problema: l'R^2 è sensibile al numero di parametri nel modello. → se aggiungo variabili al modello l'R^2 aumenta → si finirà col selezionare sempre il modello con maggiore numero di variabili, ma questo è davvero il modello migliore? Non è detto!
- Non possiamo confrontare modelli che hanno un numero di parametri diverso utilizzando l'R^2.
- Infatti R^2 non tiene conto della complessità del modello.

ADJUSTED R^2

- R^2 *adjusted*: versione «aggiustata» dell'R^2 per tener conto del numero di variabili nel modello:

$$R^2_{adj} = 1 - \frac{n-1}{n-m-1} \cdot \frac{SSE}{SST}$$

$$R^2_{adj} = 1 - \frac{n-1}{n-m-1}(1-R^2)$$

  - m = numero di variabili indipendenti nel modello (considerando eventuali variabili dummy create per codificare le variabili qualitative).

- Per confrontare modelli che presentano un numero diverso di variabili indipendenti, usare l'R^2 *adjusted* al posto dell'R^2.

- Come facciamo però a capire se la differenza tra due modelli a confronto è statisticamente significativa?

F TEST PER IL CONFRONTO TRA MODELLI (1/2)

- **Modello completo** che include m predittori $X_1, X_2, \dots, X_m$, avente coefficiente di determinazione R_m^2.
- **Modello ridotto** che include un sottoinsieme di p<m predittori: $X_1, X_2, \dots, X_p$, avente coefficiente di determinazione R_p^2.
- Domanda: il modello ridotto è significativamente diverso da quello completo? Ovvero R_m^2 è significativamente maggiore di R_p^2?

  → F test per il confronto tra modelli

F TEST PER IL CONFRONTO TRA MODELLI (2/2)

- Sistema di ipotesi:
  - H_0: $\beta_{p+1} = \beta_{p+2} = \cdots = \beta_m = 0$ → il modello completo è equivalente a quello ridotto
  - H_1: almeno uno dei coefficienti $\beta_{p+1}, \beta_{p+2}, \ldots, \beta_m \neq 0$ → il modello completo non è equivalente a quello ridotto.
- Statistica del test:

$$F = \frac{(R_m^2 - R_p^2)/(m-p)}{(1-R_m^2)/(n-m-1)}$$

- Se vale H_0, F ha distribuzione F di Fisher con m-p e n-m-1 gradi di libertà.
- Regola decisionale:
  - Se $F > F_{\alpha,m-p,n-m-1}$ → rifiutiamo H_0
  - Se $F \leq F_{\alpha,m-p,n-m-1}$ → non possiamo rifiutare H_0

INDICI DI PARSIMONIA: AIC (1/2)

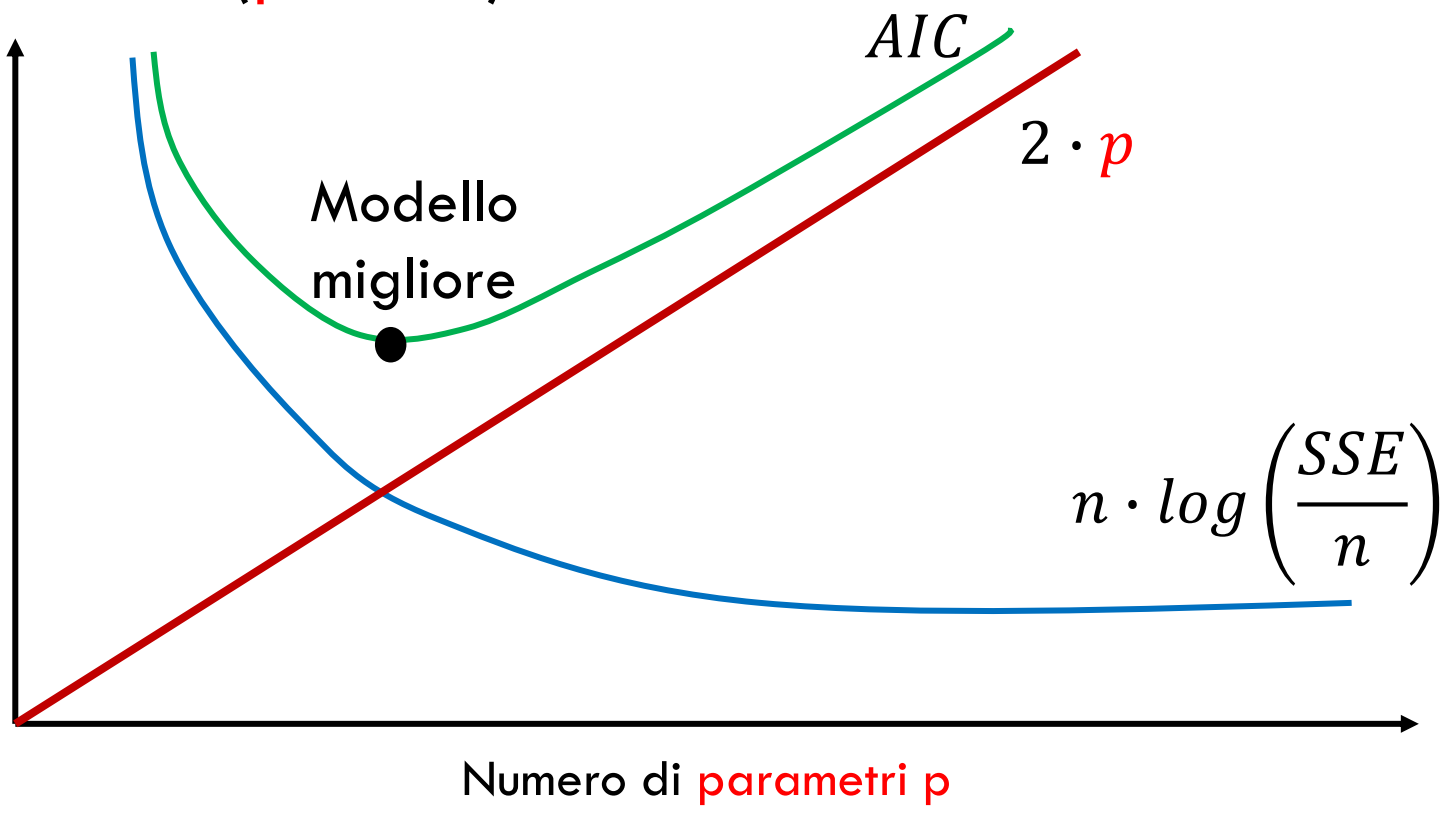
- Indici che bilanciano la capacità del modello di descrivere bene i dati e la complessità del modello.
- **Akaike Information Criterion (AIC):**

$$AIC = \underbrace{n \cdot log\left(\frac{SSE}{n}\right)}_{\text{Aderenza ai dati, capacità del modello di descrivere i dati}} + \underbrace{2 \cdot p}_{\text{Complessità del modello}}$$

p = numero di parametri nel modello, inclusa l'intercetta.

INDICI DI PARSIMONIA: AIC (2/2)

- **Idea:** tra più modelli in competizione scegliamo quello con indice AIC più basso, cioè quello che descrive bene i dati (SSE basso) con il minor numero di parametri possibili (p basso).



INDICI DI PARSIMONIA: BIC

- **Bayesian Information Criterion (BIC):**

$$BIC = \underbrace{n \cdot log\left(\frac{SSE}{n}\right)}_{\text{Aderenza ai dati, capacità del modello di descrivere i dati}} + \underbrace{p \cdot \log(n)}_{\text{Complessità del modello}}$$

- **Idea:** tra più modelli in competizione scegliamo quello con indice BIC più basso, cioè quello che descrive bene i dati (SSE basso) con il minor numero di parametri possibili (p basso).

AIC O BIC?

- BIC penalizza maggiormente modelli complessi rispetto ad AIC. BIC applica una penalità che aumenta con la numerosità del campione.
- In pratica:
  - AIC ci fa aggiungere variabili finché riusciamo a stimarne i parametri ragionevolmente bene; ci dice di smettere quando non riusciamo più.
  - BIC ci fa aggiungere variabili finché non abbiamo trovato gli effetti principali; siamo molto scettici prima di aggiungerne ancora.
- Preferiamo AIC quando lo scopo del modello è predittivo: ci interessa un modello che predica bene l'outcome.
- Preferiamo BIC quanto lo scopo del modello è descrittivo: ci interessa descrivere bene le relazioni tra le variabili esplicative e l'outcome.

ESEMPIO: CONFRONTO MEDIANTE R^2_{adj}

- Prendiamo il modello di slide 19 e proviamo a rimuovere la variabile patologia che non risultava avere un impatto significativo sull'outcome. Confrontiamo il modello ridotto con quello completo.

Metrica	**Modello con patologia**	**Modello senza patologia**
R^2	0.7194	0.7164
R^2_{adj}	0.7154	0.7135

- La differenza tra i valori di R^2 è statisticamente significativa?

ESEMPIO: F TEST

- Applichiamo il test F per confrontare il modello con e senza la variabile patologia (codificata dalle variabili dummy $D_{7,1}, D_{7,2}$).
- Sistema di ipotesi:
  - H_0: $\beta_{7,1} = \beta_{7,2} = 0$ → il modello completo è equivalente a quello ridotto
  - H_1: almeno uno dei coefficienti $\beta_{7,1}, \beta_{7,2} \neq 0$ → il modello completo non è equivalente a quello ridotto.
- Statistica del test:

$$F = \frac{(R_8^2 - R_6^2)/(8-6)}{(1-R_8^2)/(500-8-1)} = \frac{(0.7194 - 0.7164)/2}{(1-0.7194)/(491)} = 2.594$$

- $\alpha = 0.05, F_{0.05,2,491} = 3.0141$
- $F < F_{0.05,2,491}$ → non possiamo rifiutare H_0 → non possiamo dire che il modello senza la variabile patologia sia significativamente diverso da quello con la variabile patologia.

ESEMPIO: AIC E BIC

- Confrontiamo il modello con e senza la variabile patologia usando gli indici di parsimonia.

Indice di parsimonia	Modello con patologia	Modello senza patologia
AIC	385.80	387.06
BIC	419.52	412.35

- AIC preferisce il modello con patologia.
- BIC preferisce il modello senza patologia. BIC infatti tende a selezionare modelli più parsimoniosi rispetto ad AIC.

ESEMPIO: RIMOZIONE DI UNA VARIABILE IMPORTANTE

- Cosa succede invece se togliamo dal modello una variabile importante? Proviamo a rimuovere anche la lunghezza del piede.

Metrica	Modello con patologia	Modello senza patologia	Modello senza patologia e lunghezza piede
R^2	0.7194	0.7164	0.6837
R^2_{adj}	0.7154	0.7135	0.6811
AIC	385.80	387.06	439.67
BIC	419.52	412.35	460.74

- Test F per il confronto tra il modello senza patologia e il modello senza patologia e lunghezza piede: $F = 56.89, \quad F_{0.05,1,493} = 3.8604$ → rifiutiamo H_0