Metodi statistici per la ricerca sociale

Capitolo 11. Regressione Multipla e Correlazione

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM–88 SOCIOLOGIA E RICERCA SOCIALE

Relazioni Multivariate

- Lo studio di un fenomeno collettivo richiede in genere la rilevazione, su ogni singola unità statistica, di molteplici caratteri che devono a essere analizzati simultaneamente
- Analisi delle relazioni che intercorrono tra le variabili che caratterizzano il fenomeno
- Variabili di controllo

Variabili di controllo

- Obiettivo: Studiare l'associazione tra una variabile Y e una variabile X_1
- Studiare l'associazione tra Y e X_1 *controllando* per le altre variabili: Come cambia l'associazione tra X_1 e Y tenendo conto della presenza di altre variabili?
- Controllare (aggiustare) per altre variabili significa rimuovere l'influenza di tali variabili fissando il loro valore
- Studi sperimentali: I fattori di controllo (di stratificazione) sono fissati dallo sperimentatore e mantenuti costanti
- Studi osservazionali: Raggruppare unità con valori uguali (o simili) delle variabili di controllo

Variabili di controllo: Esempi

Esempio 1: Studio sperimentale

- Obiettivo: Valutare l'effetto della dimensione delle classi in asili sul punteggio a un test cognitivo
- Unità di analisi: Insegnante
- Variabile esplicativa di interesse (Trattamento): Classe piccola (13-17 bambini) versus classe regolare (22-25 bambini)
- Variabile di controllo (Fattore di stratificazione): Asilo
- Variabile risultato (variabile al livello di classe): media dei punteggi al test cognitivo dei bambini della classe a cui un insegnante è assegnato

Esempio 2: Studio osservazionale

- Studio della relazione tra reddito e gli anni di istruzione
- Il reddito può essere influenzato dal genere
- Variabile di controllo: Genere
- Si può controllare per il genere analizzando l'associazione tra il reddito e gli anni di istruzione separatamente per ogni livello della variabile di controllo

Il modello di regressione lineare multipla

- I dati consistono di n osservazioni su una variabile dipendente (risposta), Y, e k variabili esplicative $X_1, \ldots, X_k$

Osservazione	Risposta	Variabili esplicative			
u_i	Y_i	x_{i1}	x_{i2}	$\cdots$	x_{ik}
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
3	y_3	x_{31}	x_{32}	$\cdots$	x_{3k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

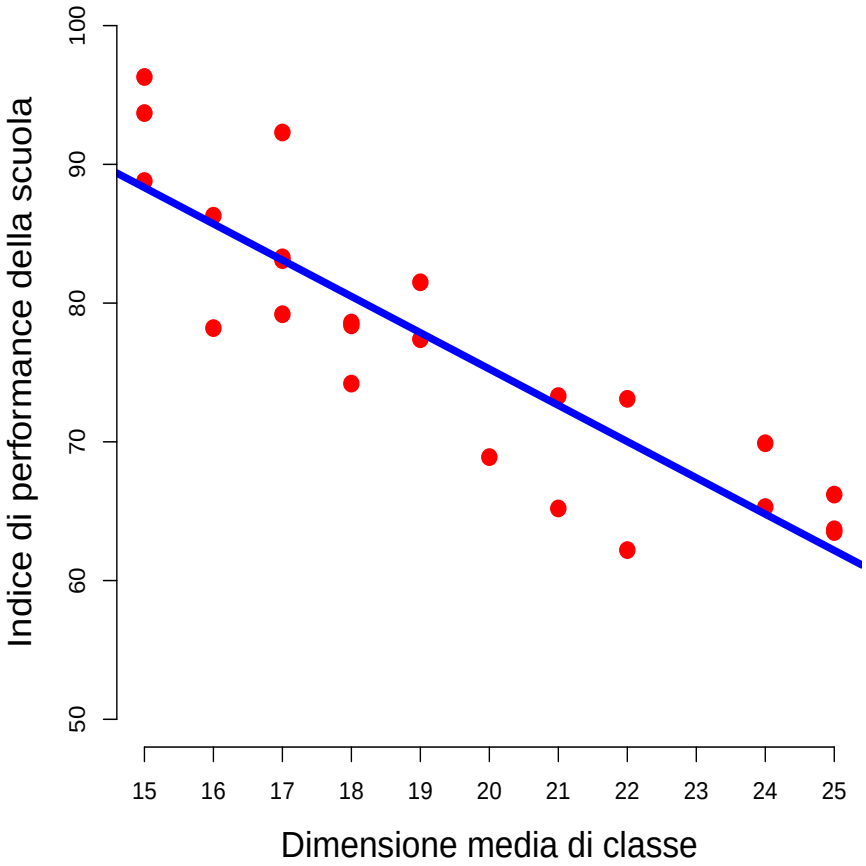
Performance delle scuole elementari

u_i	y_i	x_{i1}	x_{i2}	x_{i3}	u_i	y_i	x_{i1}	x_{i2}	x_{i3}
1	65.3	24	88	26	13	73.1	22	85	15
2	66.2	25	83	17	14	74.2	18	78	21
3	86.3	16	87	19	15	88.8	15	84	14
4	83.3	17	81	20	16	92.3	17	87	16
5	62.2	22	74	22	17	68.9	20	75	24
6	69.9	24	82	24	18	96.3	15	93	9
7	93.7	15	90	13	19	77.4	19	71	23
8	79.2	17	69	16	20	63.5	25	76	21
9	78.6	18	80	18	21	78.4	18	79	13
10	78.2	16	77	19	22	65.2	21	72	23
11	83.1	17	83	20	23	63.7	25	73	27
12	81.5	19	80	12	24	73.3	21	81	17

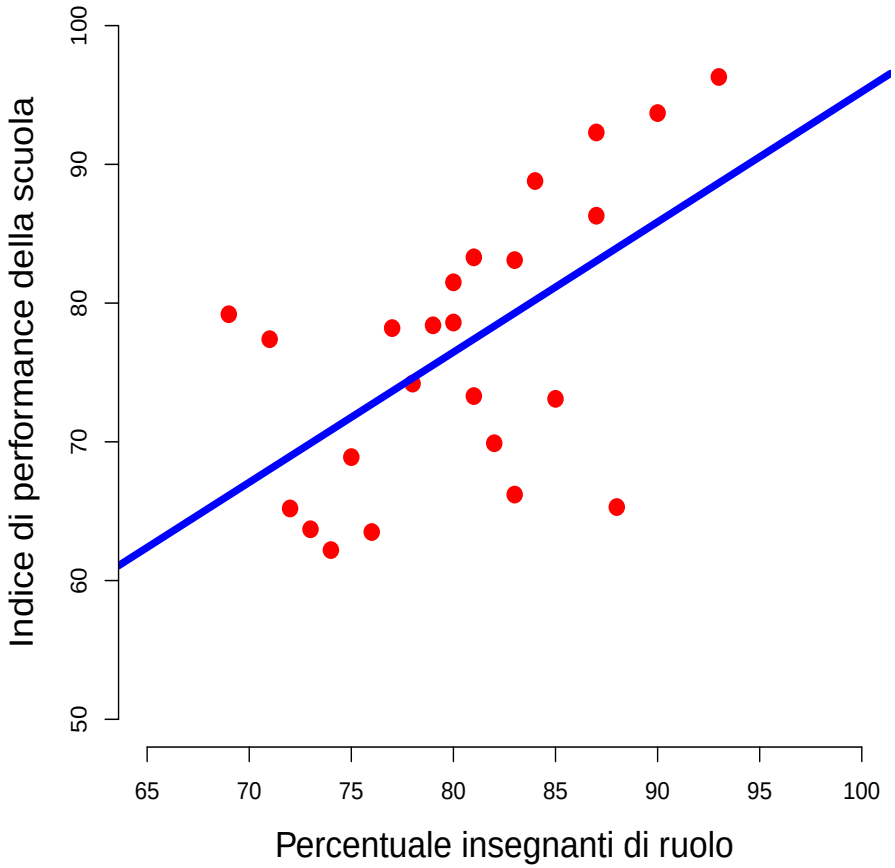
u = Scuola elementare ($n = 24$)
Y = Indice di performance accademiche della scuola
X_1 = Dimensione media di classe
X_2 = Percentuale di insegnanti di ruolo
X_3 = Percentuale di studenti le cui famiglie hanno un reddito al di sotto della soglia di povertà

Performance delle scuole elementari

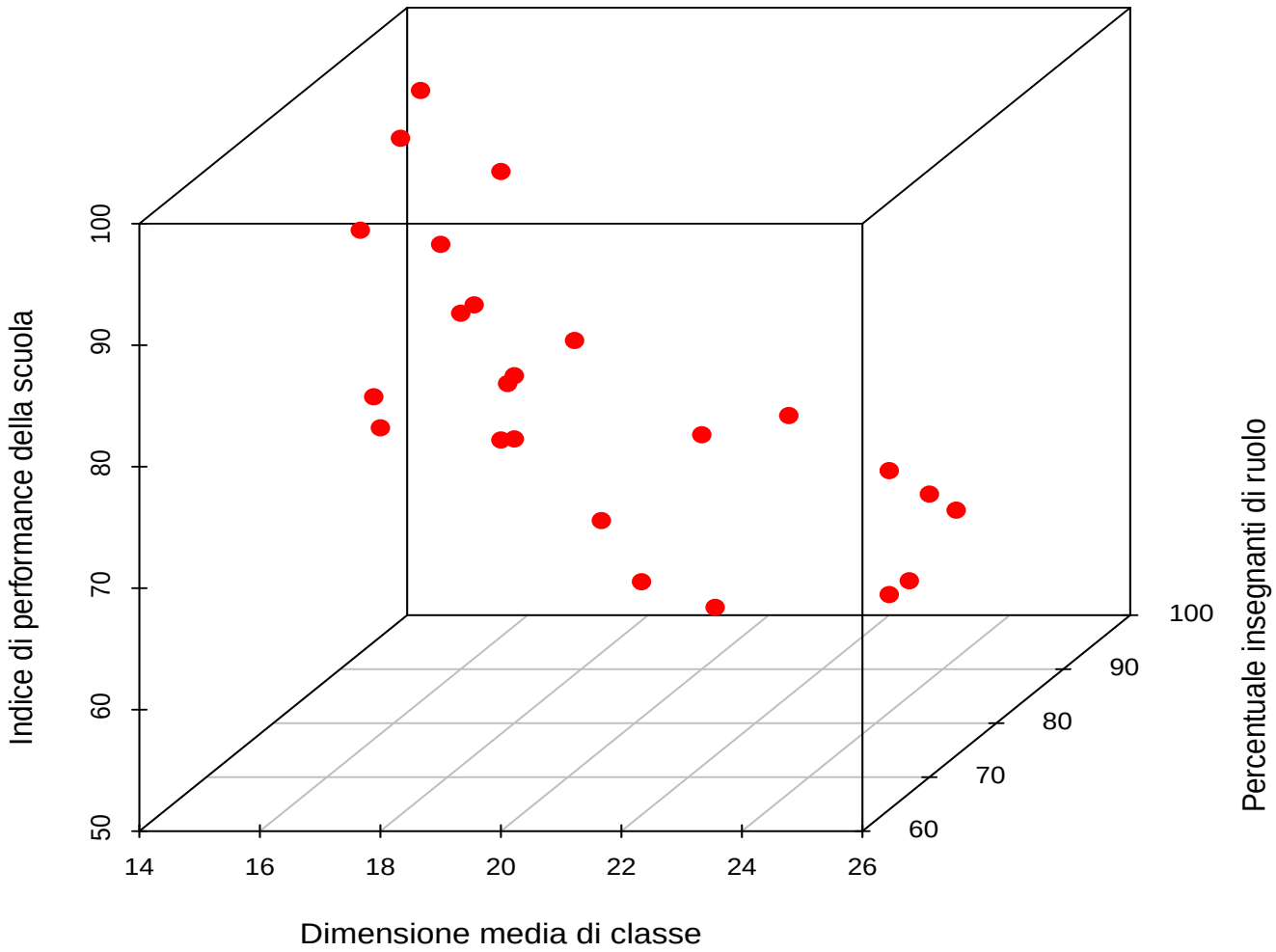
$$\hat{Y}_i = 127.56 - 2.62\, x_{i1}$$



$$\hat{Y}_i = 1.39 + 0.94\, x_{i2}$$



Performance delle scuole elementari: Diagramma a dispersione tridimensionale



Il modello di regressione lineare multipla

Le osservazioni $y_1, \ldots, y_i, \ldots, y_n$ sono realizzazioni di variabili aleatorie $Y_1, \ldots, Y_i, \ldots, Y_n$ Normali indipendenti aventi media che è funzione lineare delle variabili esplicative $X_1, \ldots, X_k$, e varianza costante indipendentemente dal valore delle variabili esplicative $X_1, \ldots, X_k$:

Per $i = 1, \ldots, n$

$$Y_i | X_{i1} = x_{i1}, \ldots X_{ik} = x_{ik} \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \sigma^2) \quad \text{indipendenti}$$

Quindi, per ogni $i = 1, \ldots, n$

$$\mathbb{E}(Y_i | X_{i1} = x_{i1}, \ldots X_{ik} = x_{ik}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

e

$$Var(Y_i | X_i = x_i) = \sigma^2 \qquad \textbf{(ipotesi di omoschedasticità)}$$

Il modello di regressione lineare multipla: Assunzioni

Assunzione 1. Per ogni unità $i = 1, \ldots, n$

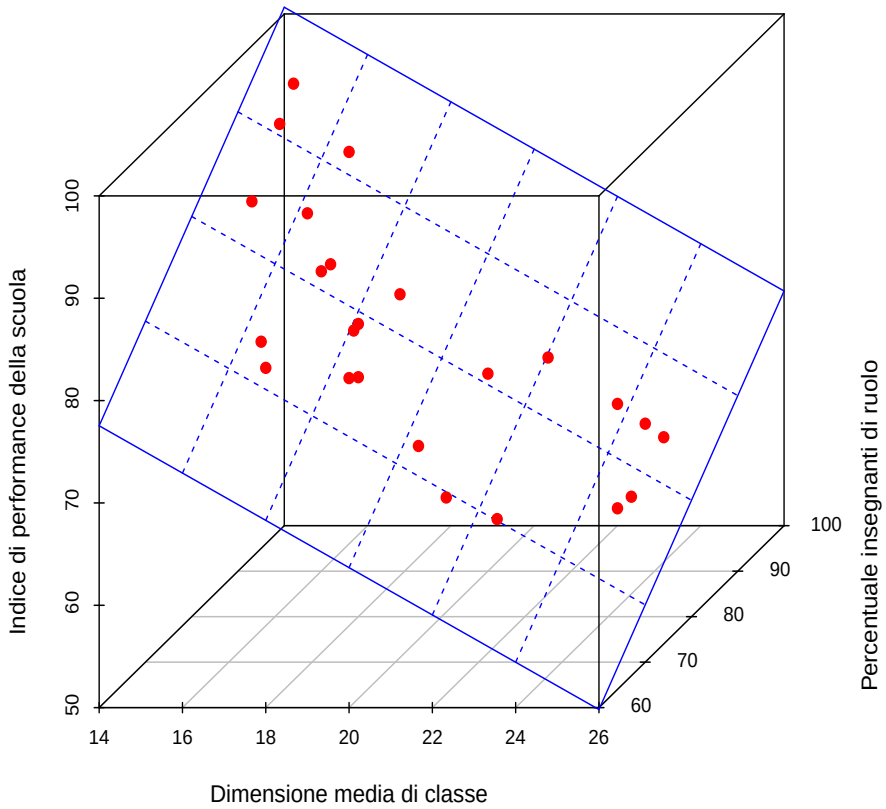
$$Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_k \cdot x_{ik} + \epsilon_i$$

Assunzione 2. Gli errori ϵ_i, $i = 1, \ldots, n$, sono variabili aleatorie indipendenti aventi media nulla e varianza costante indipendentemente dal valore delle variabili esplicative $X_1, \ldots, X_k$:

$$\text{Per} \quad i = 1, \ldots, n \qquad \epsilon_i \sim N(0, \sigma^2) \qquad \text{indipendenti}$$

Il modello di regressione lineare multipla
Esempio: Performance delle scuole elementari

$$\hat{\mathbb{E}}\left(Y_i | X_{i1} = x_{i1}, X_{i2} = x_{i2}\right) = 75.21 - 2.31\, x_{i1} + 0.58\, x_{i2}$$



Interpretazione dei coefficienti di regressione

Modello di regressione lineare multipla

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, \ldots, n$$

- I parametri di regressione $\beta_1, \ldots, \beta_k$ sono chiamati coefficienti di regressione parziale
- Il parametro β_0 (constante) rappresenta il valore atteso della variabile risposta Y per $X_1 = X_2 = \cdots = X_k = 0$

$$\mathbb{E}(Y_i | X_{i1} = 0, X_{i2} = 0, \ldots, X_{ik} = 0) = \beta_0 + \beta_1 \cdot 0 + \cdots + \beta_k \cdot 0 = \beta_0$$

- Il parametro β_j (coefficiente di regressione parziale) descrive come la variabile risposta Y varia **in media** per ogni incremento unitario della variabile esplicativa X_j quando tutte le altre variabili sono tenute costanti
  - ✓ La grandezza della variazione non dipende dai valori a cui sono fissate le altre variabili esplicative
- Il parametro β_j rappresenta l'effetto (il contributo) della variabile esplicativa X_j *controllando per* le altre variabili

Interpretazione dei coefficienti di regressione
Esempio: Performance delle scuole elementari

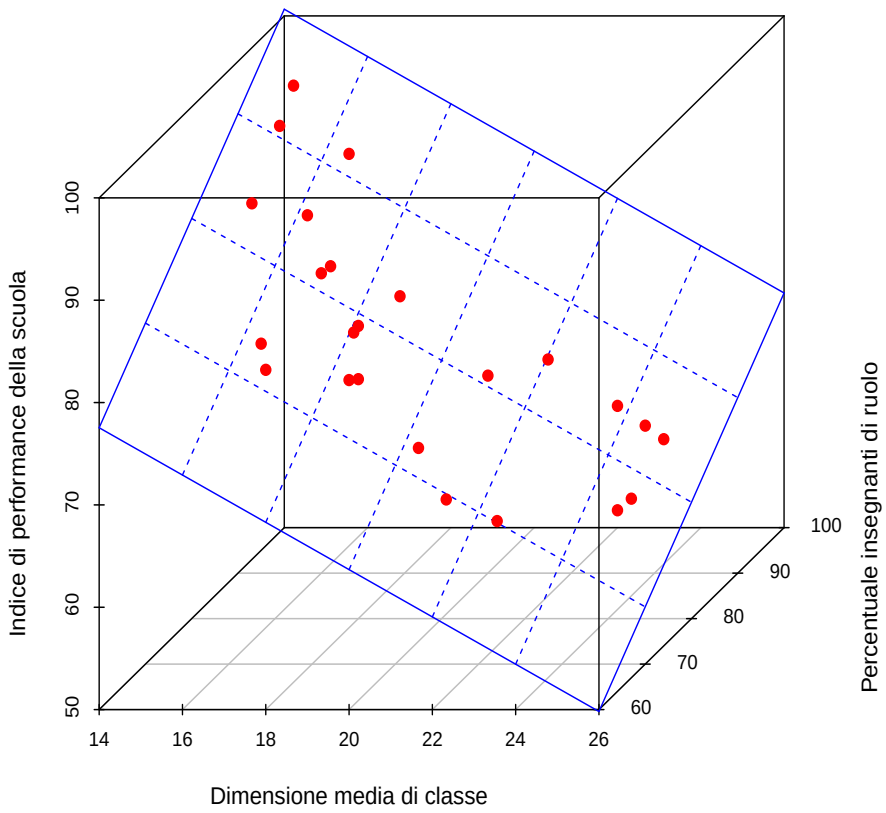
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

$$\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} = 75.21 - 2.31 x_{i1} + 0.58 x_{i2}$$

- Le performance attese di scuole elementari con dimensione media di classe nulla e percentuale di insegnanti di ruolo pari a zero sono 75.21
- Per un incremento unitario della dimensione media di classe le performance attese delle scuole elementari diminuiscono di 2.31 punti, fissato il valore della percentuale di insegnanti di ruolo, ossia controllando per la percentuale di insegnanti di ruolo
- Per un incremento unitario nella percentuale di insegnanti di ruolo, le performance delle scuole elementari aumentano in media di 0.58 punti, fissato il valore della dimensione media di classe, ossia controllando per dimensione media di classe

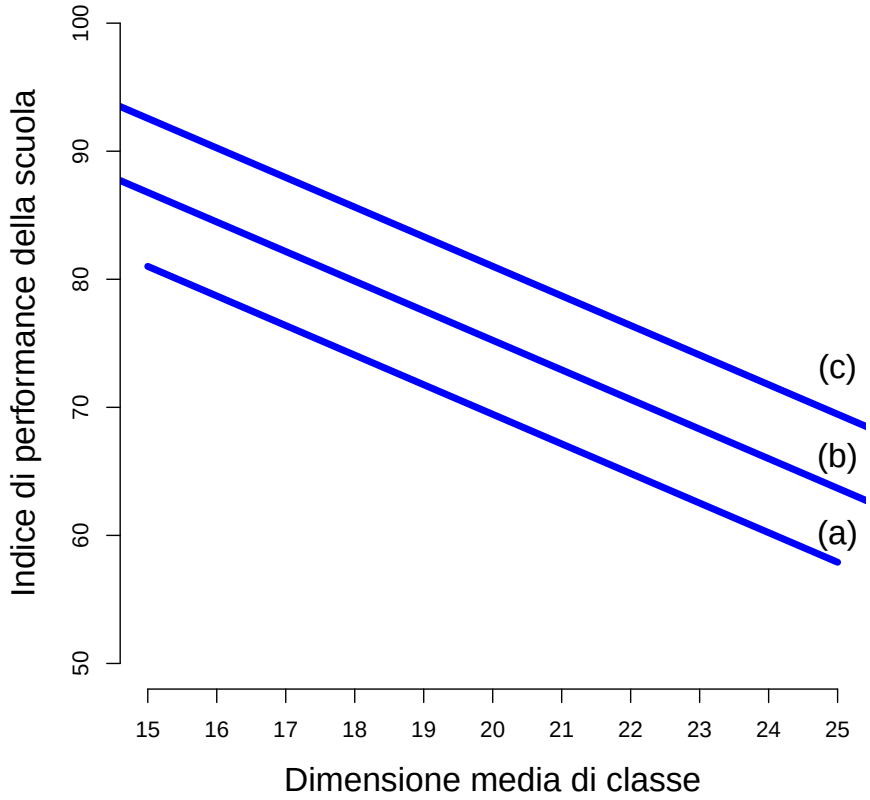
Interpretazione dei coefficienti di regressione
Esempio: Performance delle scuole elementari

$$\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} = 75.21 - 2.31 x_{i1} + 0.58 x_{i2}$$



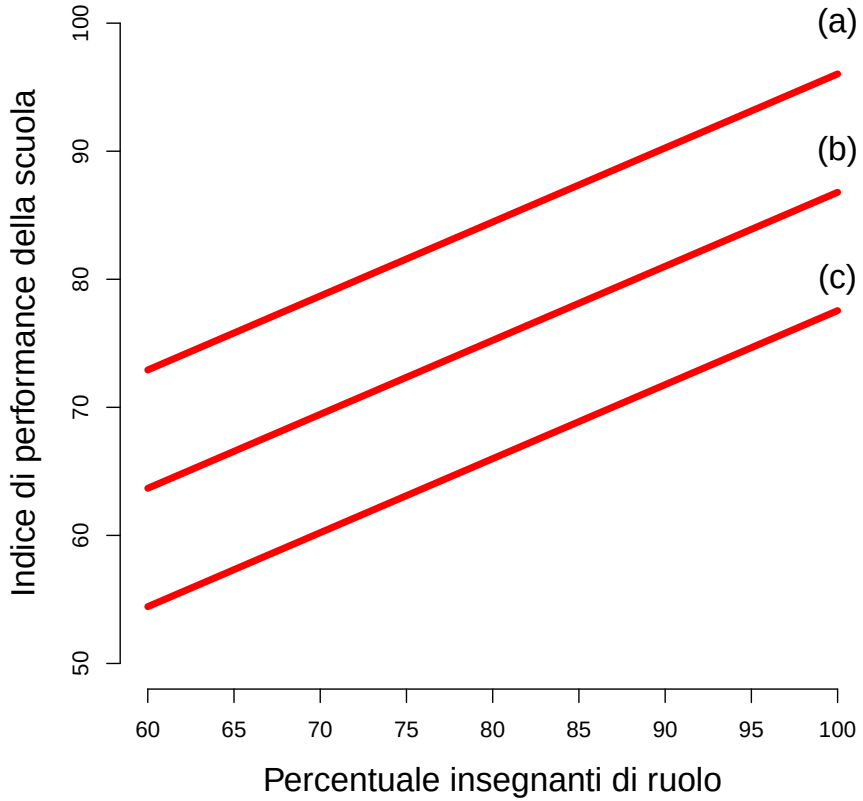
Interpretazione dei coefficienti di regressione: Regressioni parziali
Esempio: Performance delle scuole

$$\hat{Y}_i = 75.21 - 2.31\, x_{i1} + 0.58\, x_{i2}^* = \begin{cases} 115.66 - 2.31\, x_{i1} & \text{se } x_2^* = 70 \quad (a) \\ 121.44 - 2.31\, x_{i1} & \text{se } x_2^* = 80 \quad (b) \\ 127.22 - 2.31\, x_{i1} & \text{se } x_2^* = 90 \quad (c) \end{cases}$$



Interpretazione dei coefficienti di regressione: Regressioni parziali
Esempio: Performance delle scuole

$$\hat{Y}_i = 75.21 - 2.31\, x_{i1}^* + 0.58\, x_{i2} = \begin{cases} 38.25 + 0.58\, x_{i2} & \text{se } x_1^* = 16 \quad (a) \\ 29.01 + 0.58\, x_{i2} & \text{se } x_1^* = 20 \quad (b) \\ 19.76 + 0.58\, x_{i2} & \text{se } x_1^* = 24 \quad (c) \end{cases}$$



Stima puntuale dei coefficienti di regressione
Il metodo dei minimi quadrati

Il metodo dei minimi quadrati consiste nel ricercare quei valori di $\beta_0, \beta_1, \ldots, \beta_k$ che rendono minima la funzione

$$G(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2$$

- Valori Teorici

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_{i1} + \cdots + \widehat{\beta}_k \cdot x_{ik} \quad i = 1, \ldots, n$$

- I valori teorici sono i valori di Y forniti dal modello di regressione stimata in corrispondenza dei valori osservati $x_{i1}, \ldots, x_{ik}$, $i = 1, \ldots, n$
- La funzione di regressione stimata fornisce una stima della media di Y al variare di $X_1, \ldots, X_k$
- Residui di regressione

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 \cdot x_{i1} - \cdots - \widehat{\beta}_k \cdot x_{ik} \quad i = 1, \ldots, n$$

Valori teorici e residui di regressione
Esempio: Performance delle scuole elementari

$$\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} = 75.21 - 2.31 x_{i1} + 0.58 x_{i2}$$

Scuola	x_1	x_2	y_i	$\widehat{y}_i$	$\hat{e}_i$
1	24	88	65.3	70.62	-5.32
2	25	83	66.2	65.42	0.78
3	16	87	86.3	88.52	-2.22
4	17	81	83.3	82.74	0.56
5	22	74	62.2	67.15	-4.95
6	24	82	69.9	67.15	2.75
7	15	90	93.7	92.56	1.14
8	17	69	79.2	75.81	3.39
9	18	80	78.6	79.86	-1.26
10	16	77	78.2	82.74	-4.54
11	17	83	83.1	83.90	-0.80
12	19	80	81.5	77.55	3.95
13	22	85	73.1	73.50	-0.40
14	18	78	74.2	78.70	-4.50
15	15	84	88.8	89.10	-0.30
16	17	87	92.3	86.21	6.09
17	20	75	68.9	72.35	-3.45
18	15	93	96.3	94.30	2.00
19	19	71	77.4	72.34	5.06
20	25	76	63.5	61.37	2.13
21	18	79	78.4	79.28	-0.88
22	21	72	65.2	68.30	-3.10
23	25	73	63.7	59.64	4.06
24	21	81	73.3	73.50	-0.20

$$\begin{aligned} \widehat{Y}_{10} &= 75.21 - 2.31 \cdot 16 + 0.58 \cdot 77 \\ &= 82.74 \end{aligned}$$

$$\begin{aligned} \widehat{e}_{24} &= Y_{24} - \widehat{Y}_{24} = 73.3 - 73.50 \\ &= 73.3 - (75.21 - 2.31 \cdot 21 + 0.58 \cdot 81) \\ &= -0.20 \end{aligned}$$

Alcune proprietà

- Il valore atteso della variabile risposta in corrispondenza dei valori $(\overline{x}_1, \ldots, \overline{x}_k)$ è $\overline{y}$

$$\widehat{\beta}_0 + \widehat{\beta}_1 \cdot \overline{x}_1 + \ldots + \widehat{\beta}_k \cdot \overline{x}_k = \overline{y}$$

- La somma (media) dei valori teorici è uguale alla somma (media) dei valori osservati

$$\sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} y_i \quad \Longleftrightarrow \quad \overline{\widehat{y}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{y}_i = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}$$

- La somma dei residui è uguale a zero

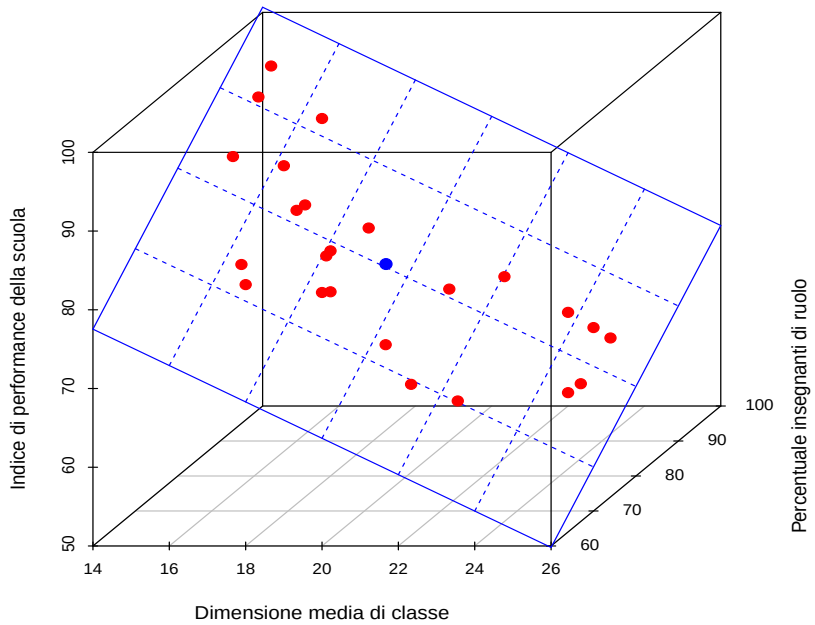
$$\widehat{e}_1 + \cdots + \widehat{e}_n = \sum_{i=1}^{n} \widehat{e}_i = 0$$

Esempio: Performance delle scuole elementari

$$\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} = 75.21 - 2.31 x_{i1} + 0.58 x_{i2}$$

$$\overline{x}_1 = \frac{466}{24} = 19.42 \qquad \overline{x}_2 = \frac{1928}{24} = 80.33 \qquad \overline{y} = \frac{1842.6}{24} = 76.775$$

$$\widehat{\beta}_0 + \widehat{\beta}_1 \cdot \overline{x}_1 + \widehat{\beta}_2 \cdot \overline{x}_2 = 75.21 - 2.31 \cdot 19.42 + 0.58 \cdot 80.3376.775 = \overline{y}$$



$$\sum_{i=1}^{24} \hat{y}_i = 1842.6 = \sum_{i=1}^{24} y_i \qquad \text{e} \qquad \sum_{i=1}^{24} \hat{e}_i = 0$$

Scomposizione della somma dei quadrati

$$\begin{array}{ccccc} \text{Somma dei Quadrati Totale} & = & \text{Somma dei Quadrati della Regressione} & + & \text{Somma dei Quadrati degli Errori} \end{array}$$

$$\Updownarrow$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Scomposizione della somma dei quadrati

- Somma dei Quadrati Totale $\Longleftrightarrow$ Devianza Totale

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n \cdot \overline{y}^2$$

- Somma dei Quadrati di Regressione $\Longleftrightarrow$ Devianza di Regressione (Devianza Spiegata): Devianza dei valori teorici

$$RSS = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

- Somma dei Quadrati dei Residui $\Longleftrightarrow$ Devianza Residua

$$SSE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}(\widehat{e}_i)^2$$

Indice di determinazione multipla

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{SSE}{TSS}$$

- Indica la proporzione di variabilità totale di Y spiegata dalla variabili esplicative $X_1, \ldots, X_k$ attraverso il modello di regressione
- R^2 non dipende dall'unità di misura delle variabili
- $0 \leq R^2 \leq 1$
- $R^2 = 0$ ($RSS = 0$ e $SSE = TSS$)

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik} = \widehat{\beta}_0 = \overline{y} \qquad \text{per ogni } i$$

- $R^2 = 1$ ($RSS = TSS$ e $SSE = 0$): tutti i residui sono nulli (il modello passa per tutti i punti campionari)
- R^2 NON può diminuire se si aggiungono variabili esplicative
- $R^2 \geq R^2_{YX_j}$ per ogni $j = 1, \ldots, k$

Indice di determinazione multipla
Esempio: Performance delle scuole elementari

Scuola	y_i	$\widehat{y_i}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$\hat{e}_i^2$
1	65.3	70.62	131.68	37.93	28.26
2	66.2	65.42	111.83	129.01	0.61
3	86.3	88.52	90.73	137.97	4.93
4	83.3	82.74	42.58	35.62	0.31
5	62.2	67.15	212.43	92.70	24.47
6	69.9	67.15	47.27	92.66	7.57
7	93.7	92.56	286.46	249.32	1.29
8	79.2	75.81	5.88	0.93	11.50
9	78.6	79.86	3.33	9.49	1.58
10	78.2	82.74	2.03	35.61	20.63
11	83.1	83.90	40.01	50.75	0.64
12	81.5	77.55	22.33	0.59	15.64
13	73.1	73.50	13.51	10.70	0.16
14	74.2	78.70	6.63	3.70	20.25
15	88.8	89.10	144.60	151.85	0.09
16	92.3	86.21	241.03	89.03	37.08
17	68.9	72.35	62.02	19.62	11.87
18	96.3	94.30	381.23	307.07	4.01
19	77.4	72.34	0.39	19.63	25.56
20	63.5	61.37	176.23	237.26	4.53
21	78.4	79.28	2.64	6.26	0.77
22	65.2	68.30	133.98	71.80	9.62
23	63.7	59.64	170.96	293.67	16.50
24	73.3	73.50	12.08	10.71	0.04
Totale	1842.6	1842.60	2341.80	2093.90	247.91

Indice di determinazione multipla
Esempio: Performance delle scuole elementari

$$TSS = 2341.81 \qquad RSS = 2093.90 \qquad SSE = 247.91$$

$$TSS = 2341.81 = 2093.90 + 247.91 = RSS + SSE$$

$$R^2 = \frac{RSS}{TSS} = \frac{2093.90}{2341.81} = 0.894$$

$$\Updownarrow$$

$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{247.91}{2341.81} = 1 - 0.106 = 0.894$$

L' 89.4% della variabilità totale delle performance delle scuole è spiegata dal modello di regressione lineare

- Si noti che

$$R^2 = 0.894 > R^2_{Y,X_1} = 0.777 \qquad \text{e} \qquad R^2 = 0.894 > R^2_{Y,X_2} = 0.338$$

Coefficiente di correlazione multipla

- Il coefficiente di correlazione multipla è il coefficiente di correlazione campionario tra i valori osservati della variabile risposta e i valori teorici

$$R = r_{Y,\hat{Y}} = \frac{s_{Y,\hat{Y}}}{s_Y \cdot s_{\hat{Y}}} = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) \cdot (\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}}$$

- $0 \leq R \leq 1$
- Si può dimostrare che

$$R = +\sqrt{(R^2)}$$

Coefficiente di correlazione multipla
Esempio: Performance delle scuole elementari

Scuola	y_i	$\widehat{y_i}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$\hat{e}_i^2$
1	65.3	70.62	131.68	37.93	28.26
2	66.2	65.42	111.83	129.01	0.61
3	86.3	88.52	90.73	137.97	4.93
4	83.3	82.74	42.58	35.62	0.31
5	62.2	67.15	212.43	92.70	24.47
6	69.9	67.15	47.27	92.66	7.57
7	93.7	92.56	286.46	249.32	1.29
8	79.2	75.81	5.88	0.93	11.50
9	78.6	79.86	3.33	9.49	1.58
10	78.2	82.74	2.03	35.61	20.63
11	83.1	83.90	40.01	50.75	0.64
12	81.5	77.55	22.33	0.59	15.64
13	73.1	73.50	13.51	10.70	0.16
14	74.2	78.70	6.63	3.70	20.25
15	88.8	89.10	144.60	151.85	0.09
16	92.3	86.21	241.03	89.03	37.08
17	68.9	72.35	62.02	19.62	11.87
18	96.3	94.30	381.23	307.07	4.01
19	77.4	72.34	0.39	19.63	25.56
20	63.5	61.37	176.23	237.26	4.53
21	78.4	79.28	2.64	6.26	0.77
22	65.2	68.30	133.98	71.80	9.62
23	63.7	59.64	170.96	293.67	16.50
24	73.3	73.50	12.08	10.71	0.04
Totale	1842.6	1842.60	2341.80	2093.90	247.91

Coefficiente di correlazione multipla
Esempio: Performance delle scuole elementari

$$R = r_{Y,\hat{Y}} = \frac{91.04}{\sqrt{101.82} \cdot \sqrt{91.04}} = 0.946$$

$$\Updownarrow$$

$$R = +\sqrt{R^2} = +\sqrt{0.894} = 0.946$$

Indice di determinazione multipla corretto

$$R^2_{Adjusted} = 1 - \frac{SSE/(n-k-1)}{TSS/(n-1)}$$

- $R^2_{Adjusted} \leq R^2$
- Introducendo una nuova variabile esplicativa nel modello $R^2_{Adjusted}$ può aumentare o diminuire a seconda di quanto si riduce la somma dei quadrati dei residui (SSE)
- Può assumere valori negativi

Indice di determinazione multipla corretto
Esempio: Performance delle scuole elementari

$$TSS = 2341.81 \quad RSS = 2093.90 \quad SSE = 247.91 \quad n = 24 \quad k = 2$$

Quindi

$$R^2_{Adjusted} = 1 - \frac{247.91/(24-2-1)}{2341.81/(24-1)} = 1 - \frac{11.805}{101.818} = 1 - 0.116 = 0.884$$

Modello	R^2	$R^2_{Adjusted}$
Solo x_1	0.777	0.766
Solo x_2	0.338	0.308
$x_2 + x_1$	0.894	0.884

Stima della varianza degli errori

$$s^2 = \frac{SSE}{n-k-1} = \frac{\widehat{e}_1^2 + \cdots + \widehat{e}_n^2}{n-k-1} = \frac{\sum_{i=1}^n \widehat{e}_i^2}{n-k-1} = \frac{\sum_{i=1}^n \widehat{e}_i^2}{gdl}$$

- È uno stimatore non distorto della varianza degli errori σ^2
- Errore Standard di regressione

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n \widehat{e}_i^2}{n-k-1}}$$

- s è una misura della variabilità degli scostamenti dei valori osservati dai valori teorici: misura la dispersione media dello stimatore intorno al suo valore atteso

Stima della varianza degli errori
Esempio: Performance delle scuole elementari

$$\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} = 75.21 - 2.31 x_{i1} + 0.58 x_{i2}$$

e

$$SSE = 247.91 \qquad n = 24$$

Quindi

$$s^2 = \frac{SSE}{n-k-1} = \frac{247.91}{24-2-1} = 11.81 \qquad \text{e} \qquad s = \sqrt{s^2} = \sqrt{11.81} = 3.44$$

Stimatori dei minimi quadrati

Nel modello di regressione multipla le caratteristiche della popolazione che interessa studiare sono i coefficienti $\beta_0, \beta_1, \ldots, \beta_k$

- Indichiamo con $\widehat{B}_0, \widehat{B}_1, \ldots, \widehat{B}_k$ gli stimatori dei minimi quadrati
- Il valori assunti da $\widehat{B}_0, \widehat{B}_1, \ldots, \widehat{B}_k$ in corrispondenza di un particolare campione sono detto stime dei minimi quadrati dei coefficienti di regressione $\beta_0, \beta_1, \ldots, \beta_k$: $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ = stime dei minimi quadrati dei coefficienti di regressione $\beta_0, \beta_1, \ldots, \beta_k$
- Gli stimatori dei minimi quadrati hanno buone proprietà: Sono stimatori non distorti con varianza minima nella classe degli stimatori corretti di un certo tipo

Inferenza per i parametri del modello di regressione

$$\widehat{B}_j \sim N\left(\beta_j, var(\widehat{B}_j)\right) \quad \Longrightarrow \quad \frac{\widehat{B}_j - \beta_j}{e.s.(\widehat{B}_j)} \sim N(0,1) \quad j = 0, 1, \ldots, k$$

dove $e.s.(\widehat{B}_j) = \sqrt{var(\widehat{B}_j)} = \sqrt{\sigma^2 \cdot c_{jj}}$ errore standard dello stimatore $\widehat{B}_j$.

$$\Downarrow$$

$$\frac{\widehat{B}_j - \beta_j}{\widehat{e.s.}(\widehat{B}_j)} \sim t_{n-k-1} \qquad \widehat{e.s.}(\widehat{B}_j) = \sqrt{s^2 \cdot c_{jj}} \quad j = 0, 1, \ldots, k$$

dove

$$s^2 = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^n \widehat{e}_i^2}{n-k-1} = \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{n-k-1}$$

Intervalli di confidenza per i coefficienti di regressione

Fissato il livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\beta_j) = \left[\widehat{\beta}_j - t_{n-k-1}(\alpha/2) \cdot \widehat{e.s.}(\widehat{B}_j); \widehat{\beta}_j + t_{n-k-1}(\alpha/2) \cdot \widehat{e.s.}(\widehat{B}_j)\right]$$

dove $t_{n-k-1}(\alpha/2)$ è tale che

$$Pr\left(-t_{\alpha/2,n-k-1} < t_{n-k-1} < t_{\alpha/2,n-k-1}\right) = 1 - \alpha$$

Intervalli di confidenza per i coefficienti di regressione
Esempio: Performance delle scuole elementari

Variabile	Stima	Errore Standard
Constante	75.21	11.60
Dimensione di classe	−2.31	0.22
Percentuale insegnanti di ruolo	0.58	0.12
Errore standard di regressione $s = 3.44$ con 21 gdl		

Fissato il livello di confidenza $1 - \alpha = 0.95$, $t_{21}(0.025) = 2.0796$

$$
\begin{aligned}
IC_{0.95}(\beta_0) &= [75.21 - 2.0796 \cdot 11.60; 75.21 + 2.0796 \cdot 11.60] \\
&= [51.08; 99.34] \\
IC_{0.95}(\beta_1) &= [-2.31 - 2.0796 \cdot 0.22; -2.31 + 2.0796 \cdot 0.22] \\
&= [-2.77; -1.85] \\
IC_{0.95}(\beta_2) &= [0.58 - 2.0796 \cdot 0.12; 0.58 + 2.0796 \cdot 0.12] \\
&= [0.33; 0.83]
\end{aligned}
$$

Verifica di ipotesi su un singolo parametro

Per $j = 0, 1, \ldots, k$

$$H_0 : \beta_j = 0 \qquad \text{versus} \qquad H_a : \beta_j \neq 0$$

- Statistica Test
$$T = \frac{\widehat{B}_j}{\widehat{e.s.}(\widehat{B}_j)} \sim t_{n-k-1} \text{ Sotto } H_0$$
- Regione di rifiuto al livello di significatività α
$$RC_\alpha : T \leq -t_{(n-k-1),\alpha/2} \text{ o } T \geq t_{(n-k-1),\alpha/2}$$
- Utilizzo del $p - value = 2 \cdot [1 - Pr\,(t_{n-k-1} \leq |T^{oss}|; H_0)]$
- Se NON è rifiutata l'ipotesi nulla $H_0 : \beta_j = 0$, allora si conclude che il coefficiente di regressione β_j non è statisticamente diverso da zero al livello di significatività α: i dati suggeriscono che non esiste una associazione (parziale) tra Y e X_j

Verifica di ipotesi – Esempio: Performance delle scuole elementari

Variabile	Stima	Errore Standard
Constante	75.21	11.60
Dimensione di classe	−2.31	0.22
Percentuale insegnanti di ruolo	0.58	0.12
Errore standard di regressione $s = 3.44$ con 21 gdl		

Fissato il livello di significatività $\alpha = 0.05$, $t_{21}(0.025) = 2.0796$

$$T_{oss}^{\beta_0} = \frac{\widehat{\beta}_0}{\widehat{e.s.}(\widehat{B}_0)} = \frac{75.21}{11.60} = 6.48$$

$$T_{oss}^{\beta_1} = \frac{\widehat{\beta}_1}{\widehat{e.s.}(\widehat{B}_1)} = \frac{-2.31}{0.22} = -10.50 \qquad T_{oss}^{\beta_2} = \frac{\widehat{\beta}_2}{\widehat{e.s.}(\widehat{B}_2)} = \frac{0.58}{0.12} = 4.83$$

- Si ha evidenza al livello del 5% che le performance delle scuole elementari siano associate con la dimensione di classe, controllando per la percentuale di insegnanti di ruolo
- Si ha evidenza al livello del 5% che le performance delle scuole elementari siano associate con la percentuale di insegnanti di ruolo, controllando per la dimensione di classe

Intervalli di confidenza e verifica di ipotesi

- Ipotesi. Per $j = 0, 1, \ldots, k$

$$H_0 : \beta_j = 0 \qquad \text{versus} \qquad H_a : \beta_j \neq 0$$

- Regione critica al livello di significatività α

$$RC_\alpha = t \leq -t_{n-k-1}(\alpha/2) \quad \text{oppure} \quad t \geq t_{n-k-1}(\alpha/2)$$

Calcolato il valore osservato della statistica test $T_{oss}^{\beta_j} = \frac{\widehat{\beta}_j}{\widehat{e.s.}(\widehat{B}_j)}$ si rifiuta $H_0 : \beta_j = 0$ se $T_{oss}^{\beta_j} \in RC_\alpha$

- Fissato il livello di confidenza $1 - \alpha$

$$\widehat{\beta}_j \pm t_{n-k-1}(\alpha/2) \cdot \widehat{e.s.}(\widehat{B}_j)$$

- L'intervallo di confidenza include lo zero?
  - ✓ No $\Rightarrow$ il parametro è significativamente diverso da zero: il valore osservato della statistica test appartiene alla regione critica
  - ✓ Si $\Rightarrow$ il parametro NON è significativamente diverso da zero: il valore osservato della statistica test NON appartiene alla regione critica

Stimatore della risposta media: $\mathbb{E}[\, Y_i \mid X_{i1} = x_1^*, \ldots, X_{ik} = x_k^* \,]$

$$\widehat{Y}_{x^*} = \widehat{B}_0 + \widehat{B}_1\, x_1^* + \cdots + + \widehat{B}_k\, x_k^*$$

- Correttezza $\mathbb{E}\left(\widehat{Y}_{x^*}\right) = \beta_0 + \beta_1\, x^* + \cdots + \beta_k\, x_k^*$
- Varianza e errore Standard di $\widehat{Y}_{x_1^*,\ldots,x_k^*}$

$$Var\left(\widehat{Y}_{x_1^*,\ldots,x_k^*}\right) \qquad e.s.\left(\widehat{Y}_{x_1^*,\ldots,x_k^*}\right) = \sqrt{Var\left(\widehat{Y}_{x_1^*,\ldots,x_k^*}\right)}$$

  ✓ La varianza e l'errore standard dello stimatore della risposta media, $\widehat{Y}_{x_1^*,\ldots,x_k^*}$, dipende da σ^2 (la varianza degli errori del modello di regressione)

- Stimatore della varianza e dell'errore standard dello stimatore di $\widehat{Y}_{x_1^*,\ldots,x_k^*}$

$$\widehat{Var}\left(\widehat{Y}_{x_1^*,\ldots,x_k^*}\right) \qquad \widehat{e.s.}\left(\widehat{Y}_{x_1^*,\ldots,x_k^*}\right) = \sqrt{\widehat{Var}\left(\widehat{Y}_{x_1^*,\ldots,x_k^*}\right)}$$

- Intervallo di confidenza al livello di confidenza del $(1-\alpha)$

$$IC_{1-\alpha}(\mathbb{E}[\, Y_i \mid X_{i1} = x_1^*, \ldots, X_{ik} = x_k^* \,]) = \widehat{Y}_{x_1^*,\ldots,x_k^*} \pm t_{n-k-1,\alpha/2}\, \widehat{e.s.}\left(\widehat{Y}_{x_1^*,\ldots,x_k^*}\right)$$

Previsione

- Previsione

$$\widehat{Y}^P_{x_1^*,\ldots,x_k^*} = \widehat{B}_0 + \widehat{B}_1\, x_1^* + \cdots + + \widehat{B}_k\, x_k^*$$

- Varianza e errore Standard di $\widehat{Y}_{x_1^*,\ldots,x_k^*}$

$$Var\left(\widehat{Y}^P_{x_1^*,\ldots,x_k^*}\right) \qquad e.s.\left(\widehat{Y}^P_{x_1^*,\ldots,x_k^*}\right) = \sqrt{Var\left(\widehat{Y}^P_{x_1^*,\ldots,x_k^*}\right)}$$

  - ✓ La varianza e l'errore standard dello stimatore $\widehat{Y}^P_{x_1^*,\ldots,x_k^*}$, dipende da σ^2 (la varianza degli errori del modello di regressione)

- Stimatore della varianza e dell'errore standard dello stimatore di $\widehat{Y}^P_{x_1^*,\ldots,x_k^*}$

$$\widehat{Var}\left(\widehat{Y}^P_{x_1^*,\ldots,x_k^*}\right) \qquad \widehat{e.s.}\left(\widehat{Y}^P_{x_1^*,\ldots,x_k^*}\right) = \sqrt{\widehat{Var}\left(\widehat{Y}^P_{x_1^*,\ldots,x_k^*}\right)}$$

- Intervallo di confidenza al livello di confidenza del $(1-\alpha)$

$$IC_{1-\alpha}(Y^P_{x_1^*,\ldots,x_k^*}) = \widehat{Y}^P_{x_1^*,\ldots,x_k^*} \pm t_{n-k-1,\alpha/2}\widehat{e.s.}\left(\widehat{Y}^P_{x_1^*,\ldots,x_k^*}\right)$$

Esempio: Performance delle scuole elementari

$$\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} = 75.21 - 2.31 x_{i1} + 0.58 x_{i2}$$

- $x_1^* = 20$ e $x_2^* = 80$
- Stima della risposta media e previsione:

$$\widehat{Y}_{x_1^*, x_2^*} = \widehat{Y}^P_{x_1^*, x_2^*} = 75.21 - 2.31 \cdot 20 + 0.58 \cdot 80 = 75.235$$

- Stima dell'errore standard dello stimatore della risposta media e della previsione

$$\widehat{e.s.}\left(\widehat{Y}_{x_1^*, x_2^*}\right) = 0.712 \qquad \widehat{e.s.}\left(\widehat{Y}^P_{x_1^*, x_2^*}\right) = 3.509$$

- Intervallo di confidenza al livello di confidenza del $(1 - \alpha) = 0.95$

$$IC_{1-\alpha}(\mathbb{E}[Y_i \mid X_{i1} = 20, X_{2i} = 80]) = 75.235 \pm 2.0796 \cdot 0.712 = (73.754; 76.715)$$

- Intervallo di previsione al livello di confidenza del $(1 - \alpha) = 0.95$

$$IC_{1-\alpha}(Y^P_{x_1^*, x_2^*}) = 75.235 \pm 2.0796 \cdot 3.509 = (67.938; 82.532)$$

Il test F

- Il test F è una procedura per verificare ipotesi riguardanti uno o più coefficienti di regressione
- Obiettivo: verificare l'ipotesi che due o più parametri siano congiuntamente pari a zero
- Tale ipotesi sottintende che le variabili esplicative corrispondenti ai parametri supposti nulli non sono utili a spiegare la relazione lineare con la variabile dipendente Y

Il test F per valutare l'influenza complessiva delle variabili esplicative

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- Ipotesi

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \qquad \textit{versus} \qquad H_a : \text{Almeno un } \beta_j \neq 0 \quad j = 1, \ldots, k$$

Oppure, equivalentemente

$$H_0 : \text{Correlazione multipla nella popolazione } = 0$$

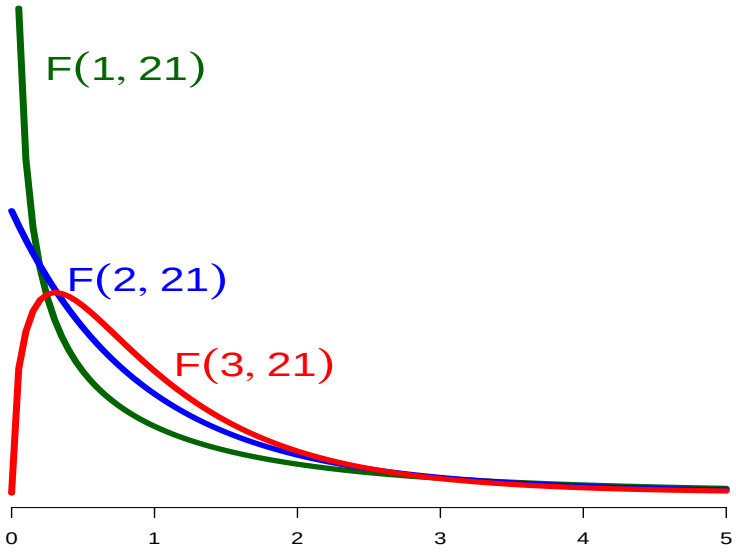
$$\textit{versus}$$

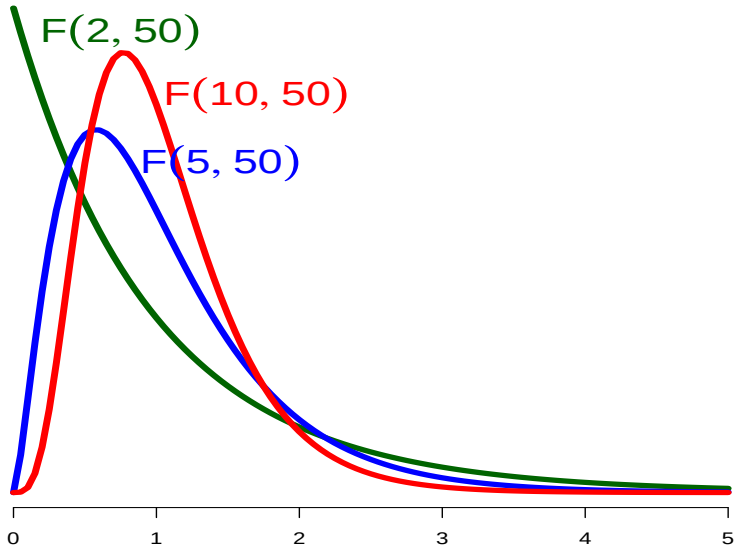
$$H_a : \text{Correlazione multipla nella popolazione } > 0$$

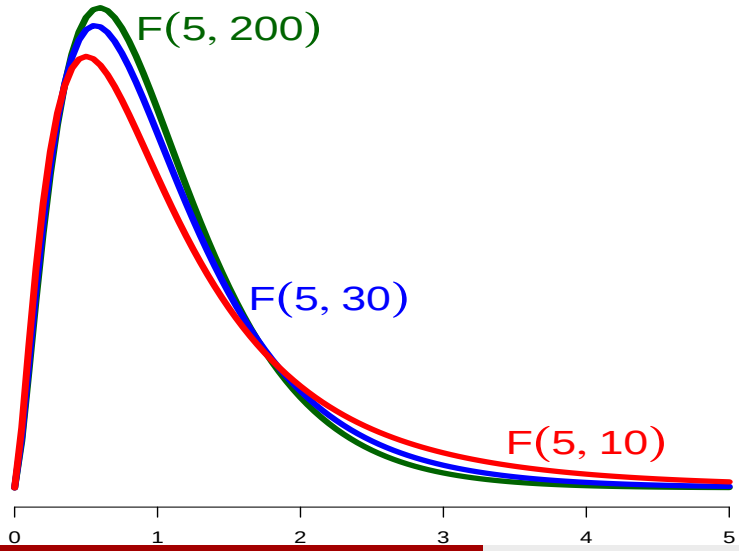
- Statistica test

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \sim F_{k,(n-k-1)} \quad \text{Sotto } H_0$$

La distribuzione F– Fisher







- $F_{gdl_1,gdl_2} \geq 0$
- $\mathbb{E}[F_{gdl_1,gdl_2}] = gdl_2/(gdl_2 - 2)$

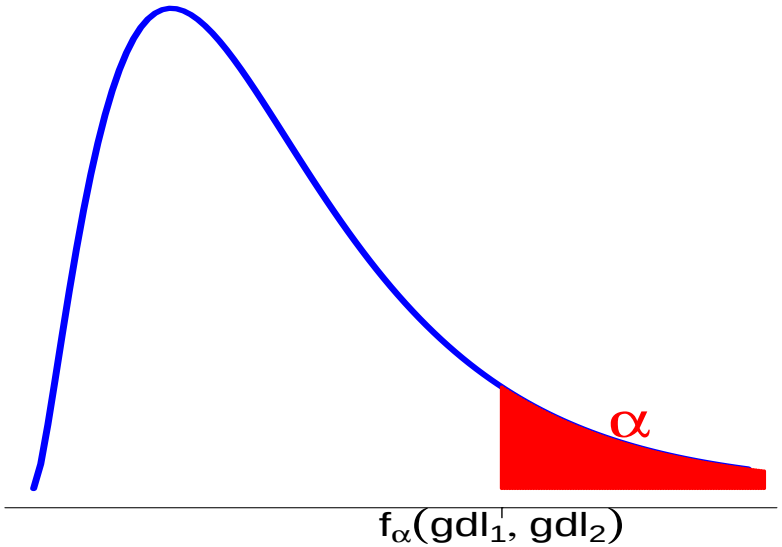
Il test F

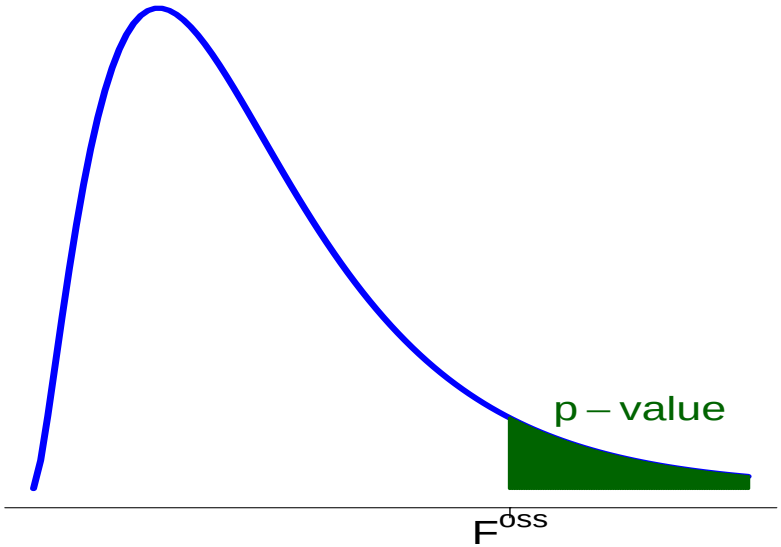
- α = Livello di significatività del test
- Regione critica di livello α

$$RC_\alpha : \quad F \geq f_{k,(n-k-1)}(\alpha)$$

dove $f_{k,(n-k-1)}(\alpha) : Pr(F_{gdl_1,gdl_2} \geq f_{k,(n-k-1)}(\alpha)) = \alpha$

- $p - value = Pr(F_{gdl_1,gdl_2} \geq F_{oss}; H_0)$





Il test F – Esempio: Performance delle scuole elementari

- Ipotesi: $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ o $\beta_2 \neq 0$
- Scomposizione della somma dei quadrati

Fonte di Variabilità	Simbolo	SQ	GdL
Regressione	RSS	2093.90	$k = 2$
Residua	SSE	247.91	$n - k - 1 = 24 - 2 - 1 = 21$
Totale	TSS	2341.81	$n - 1 = 24 - 1 = 23$

- Valore osservato della statistica test

$$R^2 = 0.89414 \Longrightarrow F_{oss} = \frac{0.89414/2}{(1 - 0.89414)/21} = 88.69$$

- $\alpha = 0.05 \Longrightarrow f_{2,21}(0.05) = 3.467$
- Regione critica di livello $\alpha = 0.05$: $RC_{0.05} = F \geq 3.467$
- $F_{oss} = 88.69 > 3.467 = f_{2,21}(0.05) \Longrightarrow$ I dati mostrano evidenza contro H_0 al livello di significatività del 5%
- $p - value = Pr(F_{2,21} \geq 88.69; H_0) = 0.000$

La statistica test F come rapporto tra medie dei quadrati

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

Si dimostra che

$$F = \frac{RSS/k}{SSE/(n-k-1)} = \frac{RMS}{MSE} = \frac{\text{Media dei quadrati di regressione}}{\text{Media dei quadrati dei residui}}$$

Tavola di Analisi della Varianza (Tavola ANOVA)

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p-value$
Regressione	RSS	k	RMS	RMS/MSE	$p-value$
Residua	SSE	$n-k-1$	MSE		
Totale	TSS	$n-1$	TMS		

Tavola di Analisi della Varianza (Tavola ANOVA)
Esempio: Performance delle scuole elementari

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F_{oss}	$p - value$
Regressione	2093.90	2	1046.948	88.69	0.000
Residua	247.91	21	11.805		
Totale	2341.81	23	101.818		

$$F_{oss} = \frac{1046.948}{11.805} = 88.69$$

Modelli di regressione a confronto

Modello esteso :

$$Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_h \cdot x_{ih} + \cdots + \beta_k \cdot x_{ik} + \epsilon_i$$

versus

Modell ridotto : $Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_h \cdot x_{ih} + \epsilon_i$

Oppure, equivalentemente,

$$H_0 : \beta_{h+1} = \beta_{h+2} = \cdots = \beta_k = 0$$

versus

H_a : Almeno un'uguaglianza in H_0 è falsa

Modelli di regressione a confronto
Esempio: Performance delle scuole elementari

Y = Indice di performance accademiche della scuola
X_1 = Dimensione media di classe
X_2 = Percentuale di insegnanti di ruolo
X_3 = Percentuale di studenti le cui famiglie hanno un reddito al di sotto della soglia di povertà

Quindi

$$M_e : Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + \epsilon_i$$

versus

$$M_r : Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \epsilon_i$$

Ossia

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{versus} \quad H_a : \beta_2 \neq 0 \text{ oppure } \beta_3 \neq 0$$

Modelli di regressione a confronto

- Valori teorici

$$\widehat{y}_{ie} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_{i1} + \cdots + \widehat{\beta}_h \cdot x_{ih} + \cdots + \widehat{\beta}_k \cdot x_{ik} \quad \text{e} \quad \widehat{y}_{ir} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_{i1} + \cdots + \widehat{\beta}_h \cdot x_{ih}$$

- Somma dei quadrati degli errori

$$\text{Modello Esteso} \quad SSE_e = \sum_{i=1}^{n} (y_i - \widehat{y}_{ie})^2 \quad \text{con } gdl_e = n - k - 1$$

$$\text{Modello Ridotto} \quad SSE_r = \sum_{i=1}^{n} (y_i - \widehat{y}_{ir})^2 \quad \text{con } gdl_r = n - h - 1$$

- Indice di determinazione multipla

$$R^2_e = 1 - \frac{SSE_e}{TSS} \qquad \text{e} \qquad R^2_r = 1 - \frac{SSE_r}{TSS}$$

- $SSE_r \geq SSE_e$ e $R^2_r \leq R^2_e$ e $gdl_r = n - h - 1 > gdl_e = n - k - 1$

Modelli di regressione a confronto

- Statistica Test

$$F = \frac{(SSE_r - SSE_e)/(gdl_r - gdl_e)}{SSE_e/gdl_e} = \frac{(SSE_r - SSE_e)/(k-h)}{SSE_e/(n-k-1)}$$

$$\Longleftrightarrow$$

$$F = \frac{(R_e^2 - R_r^2)/(gdl_r - gdl_e)}{(1-R_e^2)/gdl_e} = \frac{(R_e^2 - R_r^2)/(k-h)}{(1-R_e^2)/(n-k-1)}$$

  dove $gdl_r - gdl_e = k - h =$ numero di termini aggiuntivi presenti nel modello esteso (numero di parametri posti a zero sotto H_0)

- Sotto l'ipotesi nulla $F \sim F_{(gdl_r - gdl_e), gdl_e} = F_{k-h, n-k-1}$
- Regione di rifiuto al livello di significatività α:

$$RC_\alpha : \quad F \geq f_{k-h,(n-k-1)}(\alpha)$$

- $p - value = Pr(F_{k-h,n-k-1} \geq F_{oss}; H_0)$

Modelli di regressione a confronto
Esempio: Performance delle scuole elementari

- Ipotesi

$$M_e : Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + \epsilon_i$$

versus

$$M_r : Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \epsilon_i$$

$$\Longleftrightarrow$$

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{versus} \quad H_a : \beta_2 \neq 0 \text{ oppure } \beta_3 \neq 0$$

- Statistica test F

Modello	SSE	GdL	R^2
Ridotto	523.48	22	0.7765
Esteso	215.12	20	0.9081
Totale	2341.81	23	

$$F_{oss} = \frac{(523.48 - 215.12)/(22 - 20)}{215.12/20} = \frac{308.36/2}{215.12/20} = 14.3$$

$$= \frac{(0.9081 - 0.7765)/(22 - 20)}{(1 - 0.9081)/20} = \frac{0.1316/2}{0.0919/20} = 14.3$$

Modelli di regressione a confronto
Esempio: Performance delle scuole elementari

- Regione critica al livello di significatività $\alpha = 0.05$:

$$RC_{0.05}: \quad F \geq f_{2,20}(0.05) = 3.49$$

- $F_{oss} = 14.3 > 3.49 \Longrightarrow$ I dati mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 5%: Al livello di significatività del 5% il modello esteso è preferibile al modello ridotto
- $p - value$

$$p - value = Pr(F_{2,20} \geq 14.3; H_0) = 0.0001373$$

Modelli di regressione a confronto
Esempio: Performance delle scuole elementari

- Ipotesi

$$M_e : Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + \epsilon_i \text{ versus } M_r : Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \epsilon_i$$

$$\Longleftrightarrow$$

$$H_0 : \beta_3 = 0 \quad \text{versus} \quad H_a : \beta_3 \neq 0$$

- Statistica test F

Modello	SSE	GdL	R^2
Ridotto	247.91	21	0.8941
Esteso	215.12	20	0.9081
Totale	2341.81	23	

$$F_{oss} = \frac{(247.91 - 215.12)/(21 - 20)}{215.12/20} = \frac{32.79}{215.12/20} = 3.05$$

- Regione critica al livello di significatività $\alpha = 0.05$: $RC_{0.05}$: $F \geq f_{1,20}(0.05) = 4.35$
- $F_{oss} = 3.05 < 4.35 \Longrightarrow$ I dati non mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 5%: Al livello di significatività del 5% il modello ridotto è preferibile al modello esteso
- $p - value = Pr(F_{1,20} \geq 3.05; H_0) = 0.09615$

Relazione tra la statistica test F e la statistica test T

$$H_0 : \beta_j = 0 \qquad \text{versus} \qquad H_a : \beta_j \neq 0 \qquad j = 0, 1, \ldots, k$$

- Statistica Test

$$F = \left(\frac{\widehat{B}_j}{\widehat{e.s.}(\widehat{B}_j)} \right)^2 = (T)^2 \sim F_{1,n-k-1} \text{ Sotto } H_0$$

- $t^2_{n-k-1} \equiv F_{1,(n-k-1)}$
- Regione critica di livello α

$$RC_\alpha : \quad F \geq f_{1,(n-k-1)}(\alpha) = t^2_{(n-k-1)}(\alpha/2)$$

- $p - value$

$$Pr(F_{1,n-k-1} \geq F_{oss}; H_0) = Pr\left(t_{n-k-1} \leq -|T_{oss}| \quad o \quad t_{n-k-1} \geq |T_{oss}|; H_0\right)$$

Test F e Test T – Esempio: Performance delle scuole elementari

$$H_0 : \beta_3 = 0 \qquad \text{versus} \qquad H_a : \beta_3 \neq 0$$

Variabile	Stima	Errore Standard	t	$p - value$
Constante	84.40	12.26	6.88	0.0000
Dimensione media di classe	−2.08	0.25	−8.37	0.0000
Percentuale insegnanti di ruolo	0.49	0.13	3.90	0.0001
Percentuale studenti poveri	−0.35	0.20	−1.75	0.0962

Errore standard di regressione $s = 3.28$ con 20 gdl

Fissato il livello di significatività $\alpha = 0.05$,

$$t_{20}(0.025) = 2.086 \qquad F_{1,20}(0.05) = 4.35 = (2.086)^2 = t^2_{20}(0.025)$$

$$F^{\beta_3}_{oss} = (T^{\beta_3}_{oss})^2 = \left(\frac{-0.35}{0.20}\right)^2 = (-1.75)^2 = 3.05 \qquad p - value = Pr(F_{1,20} \geq 3.05; H_0) = 0.0962$$

- Un procedimento analogo può essere applicato per gli altri coefficienti

Test F e test T nel modello di regressione lineare semplice

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_{i1}$$

Variabile	Stima	Errore Standard	t	$p - value$
Constante	127.5566	5.8938	21.64	0.0000
Dimensione media di classe	−2.6154	0.2992	−8.742	0.0000

Errore standard di regressione $s = 4.878$ con 22 gdl

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p - value$
Regressione	1818.329	1	1818.329	76.42	0.0000
Residua	523.476	22	23.794		
Totale	2341.805	23	101.8176		

$$F_{oss} = \frac{1818.329}{23.794} = 76.42 = (-8.742)^2 = T^2$$

Modelli di regressione a confronto: Modello Esteso versus Modello Nullo

- Ipotesi

$$M_e : Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_k \cdot x_{ik} + \epsilon_i \qquad \text{versus} \qquad M_r : Y_i = \beta_0 + \epsilon_i$$

$$\Longleftrightarrow$$

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \text{versus} \quad H_a : \text{Almeno un'uguaglianza in } H_0 \text{ è falsa}$$

- Statistica test

Modello	SSE	GdL
Ridotto	TSS	$n-1$
Esteso	SSE	$n-k-1$

$$F = \frac{(SSE_r - SSE_e)/(k-h)}{SSE_e/(n-k-1)} = \frac{RSS/k}{SSE/(n-k-1)} = \frac{RMS}{MSE} \sim F_{k,n-k-1} \text{ Sotto } H_0$$

- Regione critica al livello di significatività α

$$RC_\alpha : \quad F \geq f_{k,n-k-1}(\alpha)$$

- $p - value = Pr(F_{k,n-k-1} \geq F_{oss}; H_0)$

Modello Esteso versus Modello Nullo
Esempio: Performance delle scuole elementari

- Sistema di ipotesi

$$M_e : Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \epsilon_i \qquad \text{versus} \qquad M_r : Y_i = \beta_0 + \epsilon_i$$

$$\Longleftrightarrow$$

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0 \text{ oppure } \beta_2 \neq 0$$

- Statistica test

Modello	*SSE*	GdL
Ridotto	2341.81	23
Esteso	247.91	21
Totale	2341.81	23

$$F_{oss} = \frac{(SSE_r - SSE_e)/(k-h)}{SSE_e/(n-k-1)} = \frac{(2341.81 - 247.91)/(2-0)}{247.91/21} = 88.69$$

oppure ...

Modello Esteso versus Modello Nullo
Esempio: Performance delle scuole elementari

Tavola ANOVA

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F_{oss}	$p-value$
Regressione	2093.90	2	1046.948	88.69	0.000
Residua	247.91	21	11.805		
Totale	2341.81	23	101.818		

$$F_{oss} = \frac{RSS/k}{SSE_e/(n-k-1)} = \frac{(2341.81 - 247.91)/(2-0)}{247.91/21} = \frac{1046.948}{11.805} = 88.69$$

- Regione critica al livello di significatività $\alpha = 0.05$

$$RC_{0.05}: \quad F \geq f_{2,21}(0.05) = 3.467$$

- $F_{oss} > 3.467$: I dati mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 5%.
- $p-value$

$$p-value = Pr(F_{2,21} \geq 88.69; H_0) = 0.00000$$

Interazione tra variabili (Effetti di secondo ordine)

- Esiste interazione tra due variabili esplicative quantitative se l'associazione tra la variabile risposta Y e ciascuna delle due variabili cambia al variare del valore dell'altra variabile esplicativa
- Termine di interazione = prodotto tra le due variabili (prodotto incrociato)
- Esempio

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

$$\Updownarrow$$

$$Y_i = (\beta_0 + \beta_2 x_{i2}) + (\beta_1 + \beta_3 x_{i2}) x_{i1} + \epsilon_i$$

$$\Longleftrightarrow$$

$$Y_i = (\beta_0 + \beta_1 x_{i1}) + (\beta_2 + \beta_3 x_{i1}) x_{i2} + \epsilon_i$$

La variazione attesa in Y per un incremento unitario di X_1 dipende dal valore di X_2. Analogamente la variazione attesa in Y per un incremento unitario di X_2 dipende dal valore di X_1

Interpretazione dei coefficienti di regressione
Esempio: Performance delle scuole elementari

$$\begin{aligned} \hat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i1} x_{i2} \\ &= 10.32 + 1.15 x_{i1} + 1.37 x_{i2} - 0.04 x_{i1} x_{i2} \end{aligned}$$

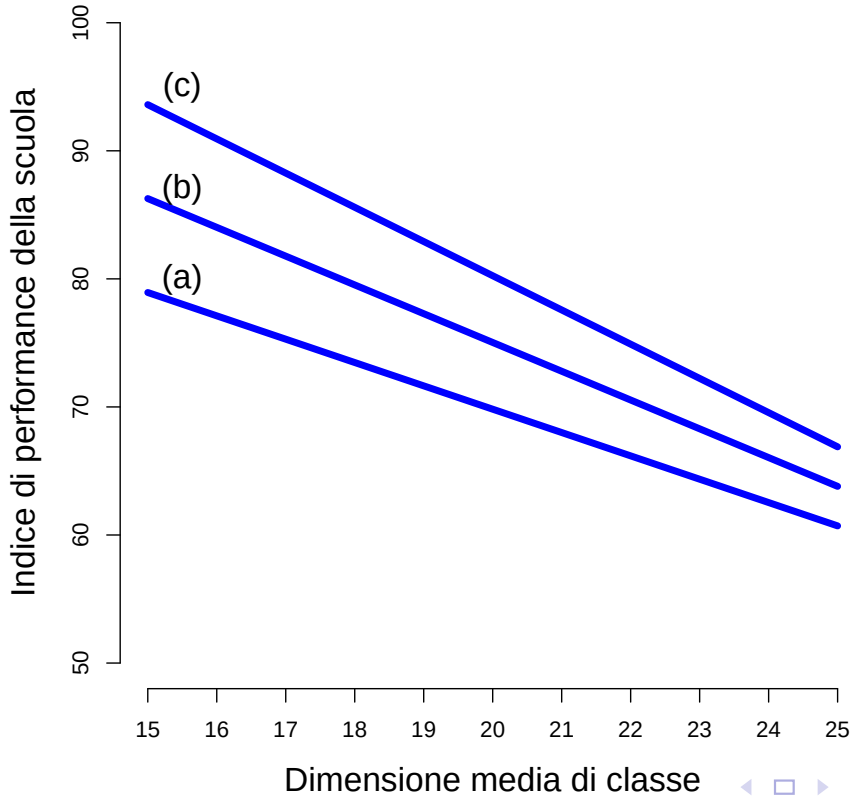
- $\widehat{\beta}_0 = 10.32$ = performance attese in scuole con dimensione media di classe zero e percentuale di insegnanti di ruolo zero
- $\widehat{\beta}_1 = 1.15$: Per le scuole in cui la percentuale di insegnanti di ruolo è zero, un incremento unitario nella dimensione media di classe implica un incremento atteso nelle performance medie di 1.15 punti
- $\widehat{\beta}_2 = 1.37$: Per le scuole in cui la dimensione media di classe è zero, un incremento unitario della percentuale di insegnanti di ruolo implica un incremento atteso nelle performance medie di 1.37 punti
- $\widehat{\beta}_3 = -0.04$: l'associazione parziale tra le performance delle scuole e la dimensione media di classe varia di -0.04 punti per ogni incremento unitario della percentuale di insegnanti di ruolo *ovvero* l'associazione parziale tra le performance delle scuole e la percentuale di insegnanti di ruolo varia di -0.04 punti per ogni incremento unitario della dimensione media di classe

Interpretazione dei coefficienti di regressione
Esempio: Performance delle scuole elementari

$$\hat{Y} = 10.32 + 1.15\, x_{i1} + 1.37\, x_{i2}^* - 0.04\, x_{i1}\, x_{i2}^*$$

Quindi

$$106.27 - 1.82\, x_{i1} \quad \text{se } x_2^* = 70 \quad (a)$$
$$119.98 - 2.25\, x_{i1} \quad \text{se } x_2^* = 80 \quad (b)$$
$$133.69 - 2.67\, x_{i1} \quad \text{se } x_2^* = 90 \quad (c)$$

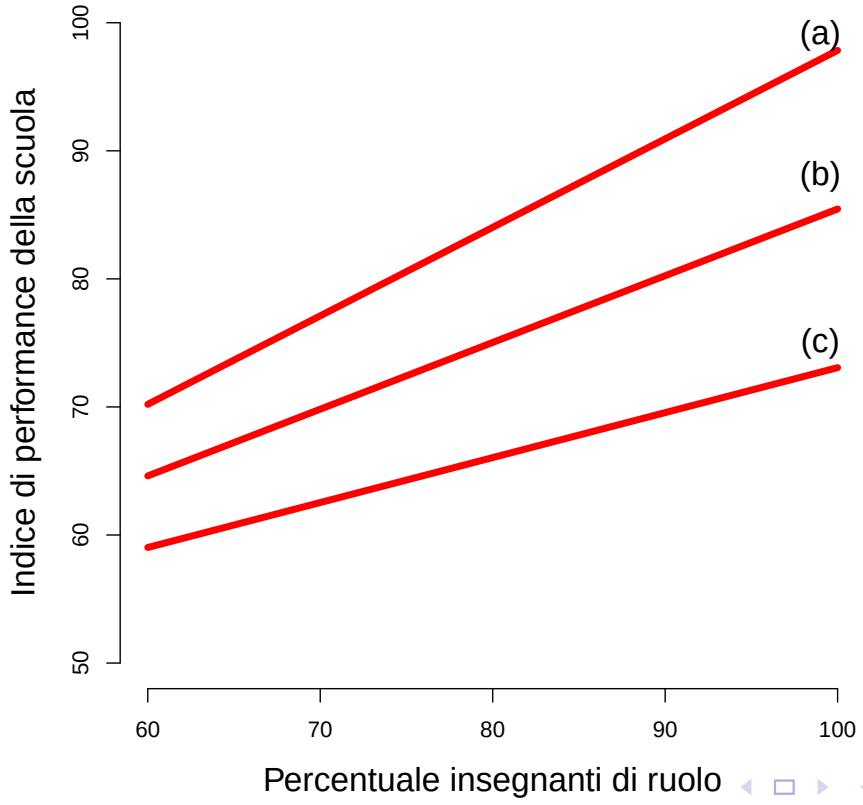


Interpretazione dei coefficienti di regressione
Esempio: Performance delle scuole elementari

$$\hat{Y} = 10.32 + 1.15\, x_{i1} + 1.37\, x^*_{i2} - 0.04\, x_{i1}\, x^*_{i2}$$

Quindi

$$28.74 + 0.69\, x_{i2} \quad \text{se } x^*_1 = 16 \quad (a)$$
$$33.35 + 0.52\, x_{i2} \quad \text{se } x^*_1 = 20 \quad (b)$$
$$37.95 + 0.35\, x_{i2} \quad \text{se } x^*_1 = 24 \quad (c)$$



Verifica di ipotesi per il termine di interazione

- Per la verifica di ipotesi su un termine di interazione si procede in modo analogo alla verifica di ipotesi su un singolo coefficiente del modello di regressione
- Esempio: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$
  - ✓ Ipotesi $H_0 : \beta_3 = 0$ versus $H_a : \beta_3 \neq 0$
  - ✓ Statistica Test

$$T = \frac{\widehat{B}_3}{\widehat{e.s.}(\widehat{B}_3)} \sim t_{n-k-1} \quad k = 3 \text{ Sotto } H_0$$

  - ✓ Regione critica al livello α: $RC_\alpha : T_{oss} \leq -t_{n-3-1,\alpha/2}$ o $T_{oss} \geq t_{n-3-1,\alpha/2}$
- Se l'evidenza contro l'ipotesi nulla (ossia a favore della presenza di in termine di interazione tra due variabili esplicative) è debole in genere si preferisce eliminare il termine di interazione e considerare un modello con solo gli effetti principali delle due variabili
- Se l'evidenza contro l'ipotesi nulla (ossia a favore della presenza di in termine di interazione tra due variabili esplicative) è forte, allora è opportuno mantenere sia il termine di interazione sia gli effetti principali delle due variabili

Esempio: Performance delle scuole elementari

Variabile	Stima	Errore Standard	t	$p-value$
Constante	10.32	54.77	0.19	0.8525
Dimensione media di classe	1.15	2.87	0.40	0.6920
Percentuale insegnanti di ruolo	1.37	0.66	2.06	0.0525
Interazione	−0.04	0.04	−1.21	0.2397

Errore standard di regressione $s = 3.398$ con 20 gdl

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p-value$
Regressione	2110.85	3	703.61	60.92	0.000
Residua	230.95	20	11.55		
Totale	2341.81	23	101.82		

- Il termine di interazione non è significativamente diverso da zero
- Notare la perdita di significatività degli effetti principali delle due variabili rispetto al modello senza interazione a causa di un'elevata correlazione tra le variabili esplicative e la loro interazione: $r_{X_1,X_1\cdot X_2} = 0.906$, $r_{X_2,X_1\cdot X_2} = 0.138$

Variabili esplicative centrate rispetto alla media
Esempio: Performance delle scuole elementari

$$Y_i = \beta_0 + \beta_1 (x_{i1} - \overline{x}_1) + \beta_2 (x_{i2} - \overline{x}_2) + \beta_3 (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) + \epsilon_i$$

$$\overline{x}_1 = 19.42 \qquad e \qquad \overline{x}_2 = 80.33$$

Quindi

Variabile	Stima	Errore Standard	t	$p - value$
Constante	76.53	0.72	105.80	0.0000
Dimensione media di classe centrata rispetto alla sua media	−2.26	0.22	−10.22	0.0000
Percentuale insegnanti di ruolo centrata rispetto alla sua media	0.55	0.12	4.50	0.0002
Interazione	−0.04	0.04	−1.21	0.2397
Errore standard di regressione $s = 3.398$ con 20 gdl				

Variabili esplicative centrate rispetto alla media
Esempio: Performance delle scuole elementari

- $\widehat{\beta}_0 = 76.53$ = performance attese in scuole con dimensione media di classe e percentuale di insegnanti di ruolo zero pari alle loro medie ($\overline{x}_1 = 19.42$ e $\overline{x}_2 = 80.33$, rispettivamente)
- $\widehat{\beta}_1 = -2.26$: Per le scuole in cui la percentuale di insegnanti di ruolo è pari alla media, $\overline{x}_2 = 80.33$, un incremento unitario nella dimensione media di classe (centrata) determina una riduzione attesa nelle performance medie di 2.26 punti
- $\widehat{\beta}_2 = 0.55$: Per le scuole in cui la dimensione media di classe è pari alla media, $\overline{x}_1 = 19.42$, un incremento unitario della percentuale di insegnanti di ruolo (centrata), implica un incremento atteso nelle performance medie di 0.55 punti
- $\widehat{\beta}_3 = -0.04$: uguale al modello con variabili non centrate
- Gli errori standard dei coefficienti di regressione parziali associati alle variabili esplicative (centrate) sono più simili a quelli del modello senza interazione
- Le stime degli effetti principali sono più significativi