Metodi statistici per la ricerca sociale

Capitolo 13.
Combinare regressione e ANOVA: predittori categoriali e quantitativi

Alessandra Mattei

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

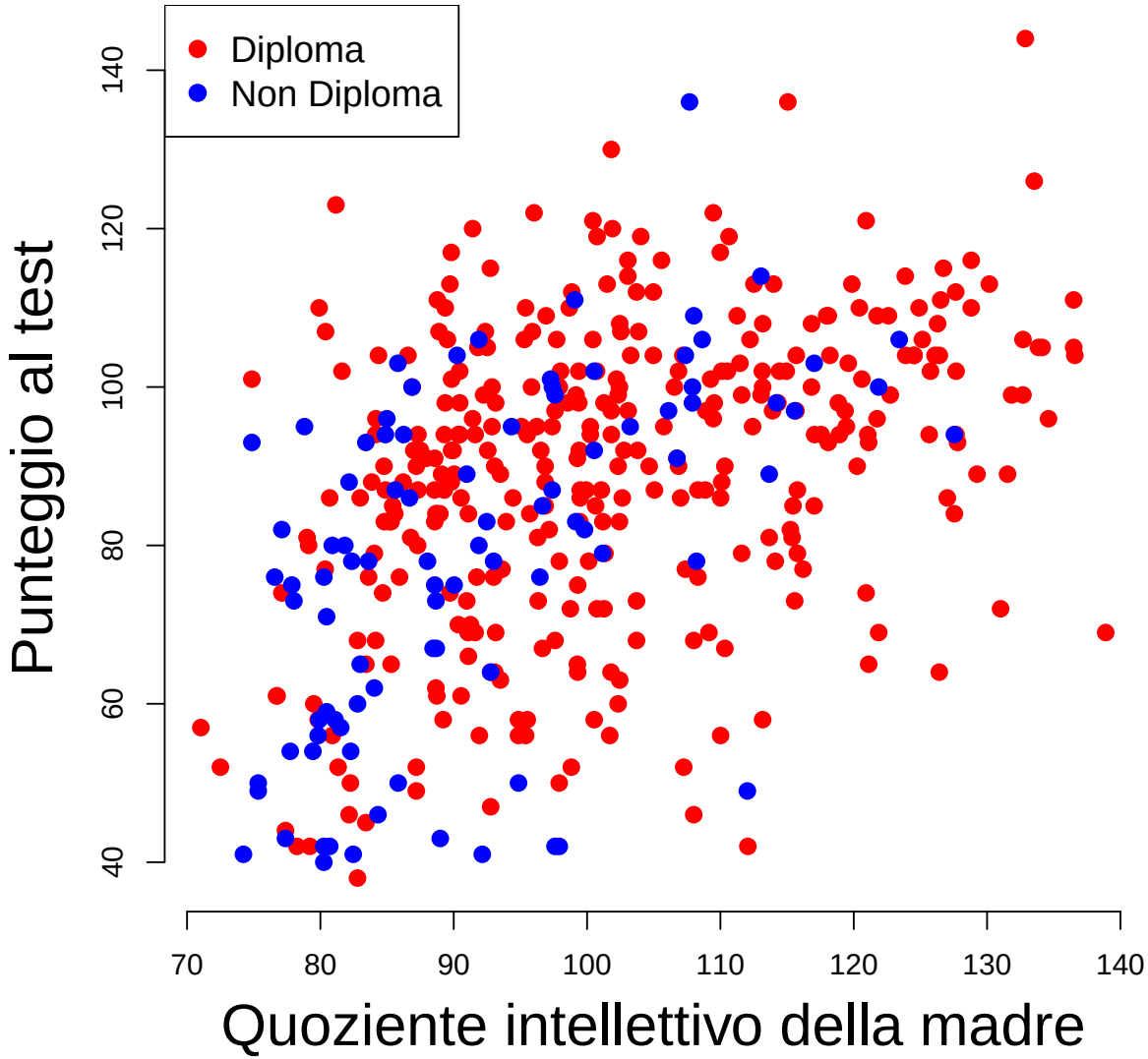
LM–88 SOCIOLOGIA E RICERCA SOCIALE

Analisi della covarianza (ANCOVA)

- Variabile risposta continua
- Variabili esplicative continue e categoriche
- Esempio: Variabile risposta continua che dipende da una variabile esplicativa continua X e un fattore A con g livelli
- Esempio:

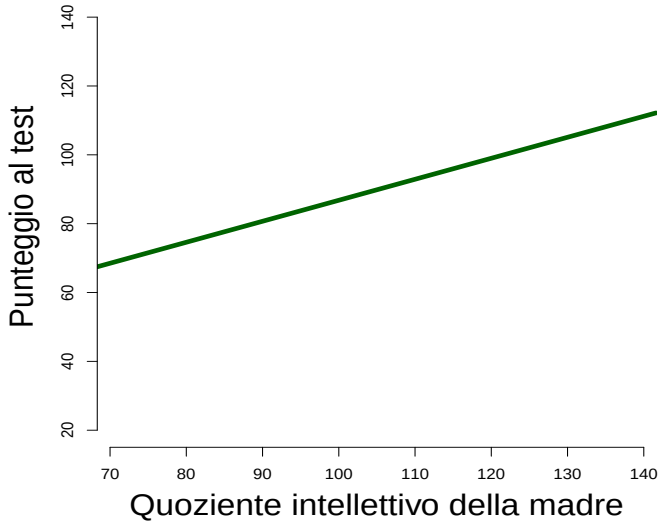
$$
\begin{array}{rcl}
Y & = & \text{Punteggio a un test finalizzato a valutare} \\
 & & \text{le capacità cognitive di bambini tra 3 e 4 anni} \\
X & = & \text{Quoziente intellettivo della madre} \\
A & = & \text{Livello di istruzione della madre} \\
 & & \text{(Inferiore al diploma = 0; Almeno il diploma = 1)}
\end{array}
$$

Esempio: Punteggio al test

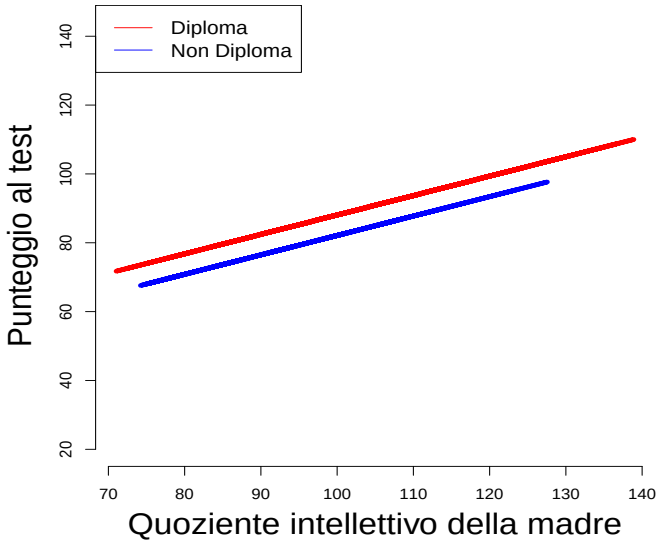


Esempio: Punteggio al test

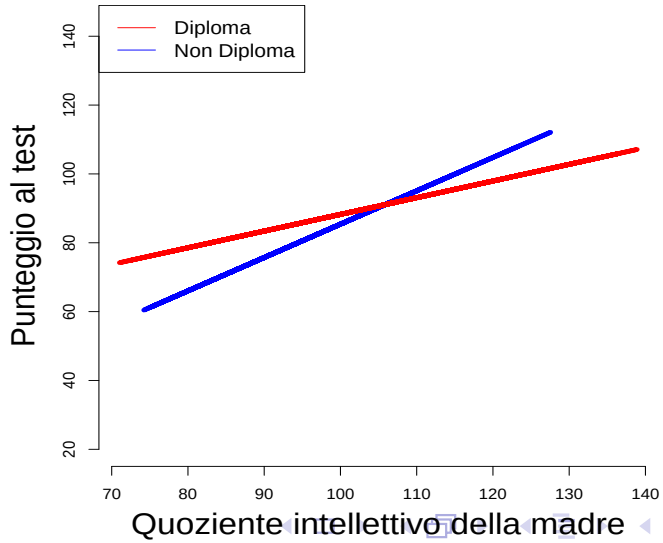
Ignorando il livello di instruzione della madre



Assenza di interazione



Con interazione



Il modello di analisi della covarianza senza interazione: Modello additivo

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \cdots + \beta_g^A \cdot A_{ig} + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

- β_0 = Valore atteso di Y per $x = 0$ nella categoria di riferimento
- β_1 = Per unità appartenenti allo stesso gruppo (ossia fissato il valore del fattore), un incremento unitario di X comporta una variazione attesa in Y di β_1
- β_ℓ^A = Differenza attesa nelle medie di Y per unità appartenenti al gruppo ℓ e unità appartenenti al gruppo di riferimento (gruppo 1) fissato il valore di X
- Valori attesi

$$\mathbb{E}(Y_i | X_i = x_i, A_{i\ell} = 1) = \begin{cases} \beta_0 + \beta_1 \cdot x_i & \text{se } A_{i1} = 1 \\ \beta_0 + \beta_1 \cdot x_i + \beta_2^A & \text{se } A_{i2} = 1 \\ \dots & \dots \\ \beta_0 + \beta_1 \cdot x_i + \beta_g^A & \text{se } A_{ig} = 1 \end{cases}$$

Modello additivo
Esempio: Punteggio al test

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \text{ indipendenti}$$

$$A_{i2} = \begin{cases} 1 & \text{se la madre del bambino } i \text{ ha almeno il diploma} \\ 0 & \text{Altrimeni} \end{cases}$$

- β_0 = Punteggio medio al test per bambini la cui madre non ha il diploma e ha un quoziente intellettivo nullo
- β_1 = Per bambini la cui madre ha almeno il diploma (non ha il diploma), un incremento unitario nel quoziente intellettivo della madre implica una variazione attesa nel punteggio al test pari a β_1
- β_2^A = A parità di quoziente intellettivo delle madri, la differenza tra la media del punteggio dei bambini con madri almeno diplomate e la media del punteggio dei bambini con madri che non hanno il diploma è pari a β_2^A
- Valori attesi

$$\begin{aligned} \mathbb{E}(Y_i | X_i = x_i, A_{i2} = 0) &= \beta_0 + \beta_1 \cdot x_i \\ \mathbb{E}(Y_i | X_i = x_i, A_{i2} = 1) &= \beta_0 + \beta_2^A + \beta_1 \cdot x_i \end{aligned}$$

Modello additivo: Stime dei coefficienti

$$\widehat{\beta}_0 = \overline{y}_1 - \widehat{\beta}_1 \cdot \overline{x}_1$$

$$\widehat{\beta}_\ell^A = (\overline{y}_\ell - \overline{y}_1) - \widehat{\beta}_1 \cdot (\overline{x}_\ell - \overline{x}_1) \quad \ell = 2, \ldots, g$$

$$\widehat{\beta}_1 = \frac{\sum_{\ell=1}^{g} \sum_{i=1}^{n_\ell} (y_{i\ell} - \overline{y}_\ell) \cdot (x_{i\ell} - \overline{x}_\ell)}{\sum_{\ell=1}^{g} \sum_{i=1}^{n_\ell} (x_{i\ell} - \overline{x}_\ell)^2}$$

Modello additivo: Stime dei coefficienti

Esempio: Punteggio al test

	Gruppo	
	Madre senza diploma	Madre con diploma
n_ℓ	93	341
$\overline{x}_\ell$	91.890	102.212
$\overline{y}_\ell$	77.548	89.32
$\sum_{i=1}^{n_\ell}(x_{i\ell} - \overline{x}_\ell)^2$	14676.71	74961.64
$\sum_{i=1}^{n_\ell}(y_{i\ell} - \overline{y}_\ell) \cdot (x_{i\ell} - \overline{x}_\ell)$	14220.11	36327.5

$$\widehat{\beta}_0 = \overline{y}_1 - \widehat{\beta}_1 \cdot \overline{x}_1 = 77.548 - 0.5639 \cdot 91.890 = 25.7315$$

$$\begin{aligned} \widehat{\beta}_2^A &= (\overline{y}_2 - \overline{y}_1) - \widehat{\beta}_1 \cdot (\overline{x}_2 - \overline{x}_1) \\ &= (89.32 - 77.548) - 0.5639 \cdot (102.212 - 91.890) = 5.9501 \end{aligned}$$

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{\ell=1}^{g}\sum_{i=1}^{n_\ell}(y_{i\ell} - \overline{y}_\ell) \cdot (x_{i\ell} - \overline{x}_\ell)}{\sum_{\ell=1}^{g}\sum_{i=1}^{n_\ell}(x_{i\ell} - \overline{x}_\ell)^2} = \frac{14220.11 + 36327.5}{14676.71 + 74961.64} \\ &= \frac{50547.61}{89638.35} = 0.5639 \end{aligned}$$

Modello additivo
Esempio: Punteggio al test

	Stima	Errore standard	$t - value$	$p - value$
Costante	25.7315	5.8752	4.38	0.0000
QI madre	0.5639	0.0606	9.31	0.0000
Istruzione madre	5.9501	2.2118	2.69	0.0074

- 25.73 è il punteggio medio al test per bambini la cui madre non ha il diploma e ha un quoziente intellettivo nullo
- Per bambini la cui madre ha almeno il diploma (non ha il diploma), un incremento unitario nel quoziente intellettivo della madre implica una variazione attesa nel punteggio al test pari a 0.5639 punti
- La media del punteggio dei bambini con madri almeno diplomate è maggiore della media del punteggio dei bambini con madri che non hanno il diploma di 5.9501 punti, controllando per il quoziente intellettivo delle madri (per ogni valore fissato quoziente intellettivo delle madri)

Il modello di analisi della covarianza con interazione

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \cdots + \beta_g^A \cdot A_{ig} + \beta_2^{AX} \cdot A_{i2} \cdot x_i + \cdots + \beta_g^{AX} \cdot A_{ig} \cdot x_i + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

- β_0 = Valore atteso di Y per $x = 0$ nel gruppo di riferimento
- β_1 = variazione attesa in Y per un incremento unitario di X per unità appartenenti al gruppo di riferimento (gruppo 1)
- β_ℓ^A = differenza tra intercette: differenza attesa nelle medie di Y per unità appartenenti al gruppo ℓ e unità appartenenti al gruppo di riferimento (gruppo 1) fissato il valore di X uguale a zero
- β_ℓ^{AX} = differenza nelle pendenze: differenza tra l'effetto di X nel gruppo ℓ e l'effetto di X nel gruppo di riferimento (gruppo 1)
- Valori attesi

$$\mathbb{E}(Y_i | X_i = x_i, A_{i\ell} = 1) = \begin{cases} \beta_0 + \beta_1 \cdot x_i & \text{se } A_{i1} = 1 \\ \beta_0 + \beta_2^A + (\beta_1 + \beta_2^{AX}) \cdot x_i & \text{se } A_{i2} = 1 \\ \cdots & \cdots \\ \beta_0 + \beta_g^A + (\beta_1 + \beta_g^{AX}) \cdot x_i & \text{se } A_{ig} = 1 \end{cases}$$

Modello con interazione – Esempio: Punteggio al test

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \beta_2^{AX} \cdot A_{i2} \cdot x_i + \epsilon_i$$

con $\epsilon \sim N(0, \sigma^2)$ indipendenti

	Stima	Errore standard	$t - value$	$p - value$
Costante	−11.4820	13.7580	−0.83	0.4044
QI madre	0.9689	0.1483	6.53	0.0000
Istruzione madre	51.2682	15.3376	3.34	0.0009
Interazione	−0.4843	0.1622	−2.99	0.0030

- −11.48 = punteggio medio al test per bambini la cui madre non ha il diploma e ha un QI nullo (non ha molto senso)
- Per bambini la cui madre **non ha il diploma**, un incremento unitario nel QI della madre implica una variazione attesa nel punteggio al test pari a 0.9689 punti
- Per bambini la cui madre ha un quoziente intellettivo pari a zero, la media del punteggio dei bambini con madri almeno diplomate è maggiore della media del punteggio dei bambini con madri che non hanno il diploma di 51.2682 punti
- Il coefficiente di interazione −0.4843 indica la differenza nelle pendenze tra bambini con madre diplomata e bambini con madre non diplomata $\Longrightarrow$ Per bambini la cui madre **ha almeno il diploma**, un incremento unitario nel QI della madre implica una variazione attesa nel punteggio al test pari a $0.9689 - 0.4843 = 0.4846$ punti

Modello con interazione – Esempio: Punteggio al test

$$Y_i = \beta_0 + \beta_1 \cdot (x_i - \overline{x}) + \beta_2^A \cdot A_{i2} + \beta_2^{AX} \cdot A_{i2} \cdot (x_i - \overline{x}) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \text{ indipendenti}$$

	Stima	Errore standard	$t - value$	$p - value$
Costante	85.4069	2.2182	38.50	0.0000
QI madre $- \overline{QI}$	0.9689	0.1483	6.53	0.0000
Istruzione madre	2.8408	2.4267	1.17	0.2424
Interazione	−0.4843	0.1622	−2.99	0.0030

- 85.4069 = punteggio medio al test per bambini la cui madre non ha il diploma e ha un QI uguale alla media ($\overline{QI} = 100$)
- Per bambini la cui madre **non ha il diploma**, un incremento unitario nel QI della madre implica una variazione attesa nel punteggio al test pari a 0.9689 punti
- Per bambini la cui madre ha un quoziente intellettivo pari alla media del QI nel campione, la media del punteggio dei bambini con madri almeno diplomate è maggiore della media del punteggio dei bambini con madri che non hanno il diploma di 2.8408 punti
- Il coefficiente di interazione −0.4843 indica la differenza nelle pendenze tra bambini con madre diplomata e bambini con madre non diplomata $\Longrightarrow$ Per bambini la cui madre **ha almeno il diploma**, un incremento unitario nel QI della madre implica una variazione attesa nel punteggio al test pari a $0.9689 - 0.4843 = 0.4846$ punti

Valutare la significatività dell'interazione

Modello esteso:

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \cdots + \beta_g^A \cdot A_{ig} + \beta_2^{AX} \cdot A_{i2} \cdot x_i + \cdots + \beta_g^{AX} \cdot A_{ig} \cdot x_i + \epsilon_i$$

versus

Modello ridotto:

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \cdots + \beta_g^A \cdot A_{ig} + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

$$\Updownarrow$$

$$H_0 : \beta_2^{AX} = \cdots = \beta_g^{AX} = 0$$

versus

H_a : Almeno un'uguaglianza in H_0 è falsa

Valutare la significatività dell'interazione

- Valori teorici

$$\widehat{y}_{ie} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_i + \widehat{\beta}_2^A \cdot A_{i2} + \cdots + \widehat{\beta}_g^A \cdot A_{ig} + \widehat{\beta}_2^{AX} \cdot A_{i2} \cdot x_i + \cdots + \widehat{\beta}_g^{AX} \cdot A_{ig} \cdot x_i$$

$$\widehat{y}_{ir} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_i + \widehat{\beta}_2^A \cdot A_{i2} + \cdots + \widehat{\beta}_g^A \cdot A_{ig}$$

- Somma dei quadrati degli errori

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

$gdl_e = n - 1 - 1 - (g-1) - (g-1) = n - g - g$

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

$gdl_r = n - 1 - 1 - (g-1) = n - 1 - g$

Valutare la significatività dell'interazione

- Statistica Test

$$F = \frac{(SSE_r - SSE_e)/(gdl_r - gdl_e)}{SSE_e/gdl_e} = \frac{(R_e^2 - R_r^2)/(gdl_r - gdl_e)}{(1 - R_e^2)/gdl_e}$$
$$= \frac{(SSE_r - SSE_e)/(g-1)}{SSE_e/(n - 2 \cdot g)} = \frac{(R_e^2 - R_r^2)/(g-1)}{(1 - R_e^2)/(n - 2 \cdot g)}$$

  dove $gdl_r - gdl_e = g - 1$ numero di termini aggiuntivi presenti nel modello esteso

- Sotto l'ipotesi nulla $F \sim F_{g-1,n-2\cdot g}$
- Regione di rifiuto al livello di significatività α:

$$RC_\alpha : \quad F \geq f_{g-1,n-2\cdot g}(\alpha)$$

- $p - value = Pr(F_{g-1,n-2\cdot g} \geq F^{oss})$

Valutare la significatività dell'interazione

Esempio: Punteggio al test

$$H_0 : \beta_2^{AX} = 0 \quad \text{versus} \quad H_a : \beta_2^{AX} \neq 0$$

- Tavola ANOVA modello esteso (modello con interazione)

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p - value$
Regressione	41507.51	3	13835.84	42.839	0.0000
Residua	138878.6	430	322.9736		
Totale	180386.2	433	416.5962	$R^2 = 0.2301$	

- Tavola ANOVA modello ridotto (modello senza interazione)

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p - value$
Regressione	38629.07	2	19314.53	58.724	0.0000
Residua	141757.1	431	328.9028		
Totale	180386.2	433	416.5962	$R^2 = 0.2141$	

Valutare la significatività dell'interazione
Esempio: Punteggio al test

- Statistica test

$$F^{oss} = \frac{(41507.51 - 38629.07)}{138878.6/430} = \frac{2878.442}{322.9736} = 8.91$$

$$= \frac{0.2301 - 0.2141}{(1 - 0.2301)/430} = \frac{0.015957}{0.769896/430} = 8.91$$

- Si noti che $F^{oss} = (T^{oss}(\beta_2^{AX}))^2 = (-2.99)^2$ dato che l'interazione è caratterizzata da un unico parametro (essendo la variabile categorica binaria)
- Regione critica di livello di significatività $\alpha = 0.05$

$$RC_{0.05}: \quad F \geq f_{1,430}(0.05) = 3.863$$

- p–value

$$p - value = Pr(F_{1,430} \geq 8.91) = 0.00299$$

- Evidenza contro l'ipotesi nulla

Modelli – Esempio: Punteggio al test

- Modello con interazione: $Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \beta_2^{AX} \cdot A_{i2} \cdot x_i + \epsilon_i$

	SQ	GdL	MQ	$F - value$	$p - value$
Regressione	41507.51	3	13835.84	42.839	0.0000
Residua	138878.6	430	322.9736	$R^2 = 0.2301$	

- Modello additivo: $Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \epsilon_i$

	SQ	GdL	MQ	$F - value$	$p - value$
Regressione	38629.07	2	19314.53	58.724	0.0000
Residua	141757.1	431	328.9028	$R^2 = 0.2141$	

- Modello con solo il predittore continuo: $Y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$

	SQ	GdL	MQ	$F - value$	$p - value$
Regressione	36248.82	1	36248.82	108.643	0.0000
Residua	144137.3	432	333.6512	$R^2 = 0.2010$	

- Modello con solo il predittore categorico: $Y_i = \beta_0 + \beta_2^A \cdot A_{i2} + \epsilon_i$

	SQ	GdL	MQ	$F - value$	$p - value$
Regressione	10124.97	1	10124.97	25.69	0.0000
Residua	170261.2	432	394.1231	$R^2 = 0.0561$	

- Modello nullo: $Y_i = \beta_0 + \epsilon_i$ ($SSE = TSS = 180386.2$ con 433 gdl ($TMS = 416.6$))

Modelli a confronto - Esempio: Punteggio al test

Modello	SSE	GdL	$F - value$	$p - value$
Modello con interazione (Modello esteso)	138878.6	430		
Modello additivo: $H_0 : \beta_2^{AX} = 0$	141757.1	431	8.912	0.00299
Solo X = QI madre: $H_0 : \beta_2^{A} = 0$ e $\beta_2^{AX} = 0$	144137.3	432	8.141	0.00034
Solo A = Istruzione madre: $H_0 : \beta_1 = 0$ e $\beta_2^{AX} = 0$	170261.2	432	48.584	0.00000
Modello nullo: $H_0 : \beta_1 = \beta_2^{A} = \beta_2^{AX} = 0$	180386.2	433	42.839	0.00000

Modelli a confronto - Esempio: Punteggio al test

Modello	SSE	GdL	$F - value$	$p - value$
Modello senza interazione (Modello esteso)	141757.1	431		
Solo X = QI madre: $H_0 : \beta_2^A = 0$	144137.3	432	7.237	0.00742
Solo A = Istruzione madre: $H_0 : \beta_1 = 0$	170261.2	432	86.664	0.00000
Modello nullo: $H_0 : \beta_1 = \beta_2^A = 0$	180386.2	433	58.724	0.00000