# Statistica
## Regressione e Correlazione Multipla

Andrea Giommi

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze

Scuola di Psicologia
Corso di Studio in Scienze e Tecniche Psicologiche

# Introduzione

- Il modello di regressione lineare più semplice (modello binario), che mette in relazione lineare una variabile risposta con una variabile esplicativa (predittore o regressore), può essere scritto come segue:

$$E(y) = \alpha + \beta x$$

- Nella pratica è spesso ragionevole assumere che altre variabili abbiano una qualche influenza sulla variabile risposta ed è quindi necessario inseririrle nel modello
- il modello binario può essere generalizzato nel seguente **modello di regressione multipla**:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

se sono $k$ le variabili che si ritengono collegate linearmente con la variabile risposta

# Interpretazione dei parametri del modello

- Nel modello di regressione multipla:
- $\alpha = E(y)$ quando $x_1 = x_2 = \cdots = x_k = 0$; e
- $\beta_1, \beta_2, \cdots, \beta_k$ sono detti **coefficienti di regressione parziali**
- Facendo riferimento alla prima variabile esplicativa, $x_1$, il coefficiente parziale $\beta_1$ ci dice di quanto varia la $y$ a seguito di un incremento unitario della $x_1$, per una qualsiasi combinazione di **valori costanti** delle altre variabili, o, in altri termini, controllando per le altre variabili

# Interpretazione dei parametri del modello

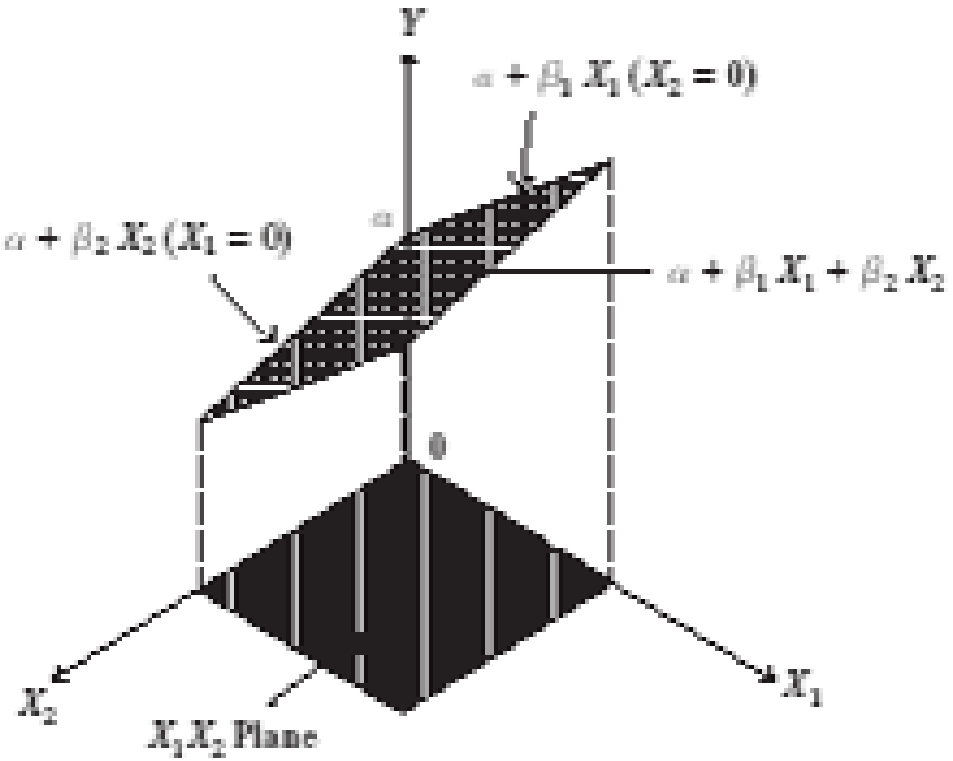
- Per esempio con due variabili esplicative ($k = 2$),

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

- Se $x_1$ aumenta di una unità e $x_2$ resta costante, la variazione di $E(y)$ è pari a:

$$[\alpha + \beta_1(x_1 + 1) + \beta_2 x_2] - [\alpha + \beta_1 x_1 + \beta_2 x_2] = \beta_1$$

# Espressione grafica del modello con due variabili esplicative



# Equazione di previsione

- I parametri del modello vengono stimati dai dati campionari utilizzando il criterio dei Minimi Quadrati.
- In modo del tutto analogo a quello visto per la regressione binaria, il criterio porta ad individuare i valori dei parametri che minimizzano la somma del quadrato dei residui:

$$SSE = \sum (y \text{ osservati} - y \text{ previsti})^2 = \sum (y - \hat{y})^2$$

- Stimati i parametri, possiamo scrivere l'equazione di previsione dei minimi quadrati:

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

# Esempio: studio sul disagio mentale

- $y$ = disagio mentale (in breve "disagio", espresso da un punteggio che sintetizza e riassume la presenza e l'importanza di sintomi psichiatrici come: stati ansionsi, depressione, ecc., ricavato da domande del questionario della "Healt opinion survey" con risposte di tipo ordinale quali: quasi mai, qualche volta, spesso, ecc.)

  $y$ varia da 17 a 41 nel campione, con media pari a 27 e $s = 5$

- $x_1$ = "Eventi di vita" (in breve "Eventi", punteggio che combina numero e importanza di eventi, fatti, accadimenti, degli ultimi tre anni)

  $x_1$ varia da 0 a 100 con media campionaria pari a 44 e $s = 23$

- $x_2$ = stato socio economico (in breve "SES", ancora un punteggio che combina stato occupazionale, reddito e livello di istruzione)

  $x_2$ varia da 0 a 100 com media campionaria 57 e $s = 25$

# Esempio sul disagio mentale (segue)

- La dimensione del campione per l'esempio è $n = 40$. I dati sono a pag.331 del libro di testo.
- Lo studio comprende altre variabili esplicative che non vengono utilizzate nell'esempio, quali: età, stato civile, genere, razza.
- Se facciamo un'analisi di tipo bivariato otteniamo le seguenti due espressioni di previsione:

$$\hat{y} = 23,3 + 0,090x_1$$
$$\hat{y} = 32,2 - 0,086x_2$$

- E matrice di correlazione:

| | Disagio | Eventi | SES |
|---|---|---|---|
| Disagio | 1,000 | $0,372^*$ | $-0,399^*$ |
| Eventi | $0,372^*$ | 1,000 | 0,123 |
| SES | $-0,399^*$ | 0,123 | 1,000 |

# Esempio sul disagio mentale (segue)

- L'equazione di previsione nella regressione multipla è:

$$\hat{y} = 28,23 + 0,103x_1 - 0,097x_2$$

| | Coeff. non standard. | | Coeff. standard. | | | Intervallo al 95% | |
|---|---|---|---|---|---|---|---|
| Modello | $b$ | stand.err. | $b^*$ | $t$ | sig. | $L_i$ | $L_s$ |
| Costante | 28,23 | 2,174 | | 12,984 | 0,000 | 23,824 | 32,635 |
| Eventi | 0,103 | 0,032 | 0,428 | 3,177 | 0,003 | 0,037 | 0,169 |
| SES | -0,97 | 0,029 | -0,451 | -3,351 | 0,002 | -0,156 | -0,039 |

# Esempio sul disagio mentale (segue)

Il disagio mentale previsto:

- aumenta di 0,103 all'aumento unitario del punteggio degli eventi di vita, mantenendo sotto controllo (a valori costanti) la variabile stato economico-sociale
- decresce di 0,097 all'aumento unitario dello stato economico-sociale, controllando per la variabile eventi.
- Questo significa, ad esempio, che il disagio mentale si riduce di 9,7 punti se la variabile *SES* passa da 0 a 100, mantenendo costante il valore degli eventi.

# Esempio sul disagio mentale (segue)

- Non possiamo confrotare i coefficienti di regressione parziali per stabilire quale variabile esplicativa abbia maggiore impatto sulla variabile risposta poiché tali coefficienti sono "non standardizzati" e quindi dipendono da unità di misura e media.
- Il software per la stima dei coefficienti calcola anche coefficienti standardizzati che esprimono la variazione della $y$ al variare di una deviazione standard della variabile esplicativa $x_i$, controllando per le altre variabili esplicative
- Nella regressione binaria il coefficiente di regressione standardizzato è il coefficiente di correlazione $r$. Nella regressione multipla, il coefficiente di regressione standardizzato è legato algebricamente al coefficiente di correlazione parziale.

# Valori previsti e residui

- Nel file di dati (a pag. 331) uno dei soggeti ha i seguenti valori:

  $Y = 33$; $x_1 = 45$ (vicino alla media); $x_2 = 55$ (vicino alla media)

- questo soggetto ha un disagio mentale previsto pari a:

  $\hat{y} = 28,23 + 0,103(45) - 0,097(55) = 27,5$ (vicino alla media)

- Il residuo (errore di previsione) per questo individuo è: $33 - 27,5 = 5,5$; cioè questa persona ha un disagio mentale di 5,5 punti maggiore rispetto a quello previsto sulla base delle variabili: *Eventi* e *SES*
- La somma dei residui al quadrato $SSE = 768,2$ per l'intero campione è inferiore a quella che deriva dalle due regressioni binarie e da qualsiasi relazione lineare con regressori $x_1$ e $x_2$ diversa da quella dei minimi quadrati.

# Commenti sul modello di regressione multipla

- Nella regressione multipla gli effetti dei vari coefficienti possono essere definiti "netti" in quanto non risentono degli effetti delle altre variabili nel modello, diversamente dalla regressione binaria, nella quale gli effetti di altre possibili variabili rispetto a quella del modello sono completamente ignorati

- L'effetto parziale di $x_1$ (tenendo sotto controllo $x_2$) è uguale all'effetto della stessa variabile nella regressione binaria solo se la correlazione tra $x_1$ e $x_2$ è uguale a 0. Questo, peraltro è auspicabile anche se difficilmente realizzabile nelle indagini osservazionali.

# Commenti sulla regressione multipla (segue)

- L'effetto parziale di ciascuna variabile esplicativa in questo modello di regressione è lo stesso per qualsiasi valore fisso dell'altra variabile:

**Esempio:**

$$\hat{y} = 28,23 + 0,103x_1 - 0,097x_2$$

Per $x_2 = 0$

$$\hat{y} = 28,23 + 0,103x_1 - 0,097(0) = \hat{y} = 28,23 + 0,103x_1$$

Per $x_2 = 100$

$$\hat{y} = 28,23 + 0,103x_1 - 0,097(100) = \hat{y} = 18,5 + 0,103x_1$$

# Commenti sulla regressione multipla (segue)

- Questo parallelismo delle due linee rette mostra l'assenza di effetto del valore di $x_2$ su la relazione tra $y$ e $x_1$. In questo caso parliamo di un modello con assenza di interazione tra le due variabili esplicative
- Se vi fosse un'interazione questo parallelismo non sarebbe presente. Vedremo successivamente come modificare il modello per poterne tener conto
- Il modello:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

può essere equivalentemente scritto nella forma:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

dove $\varepsilon = y - E(y)$ è un errore che ha $E(\varepsilon) = 0$

# Correlazione multipla e $R^2$

- Possiamo domandarci in che misura le variabili esplicative scelte per il modello riescano a prevedere i valori della variabile risposta
- La correlazione multipla, $R$, e il coefficiente di determinazione multiplo, $R^2$, rispondono entrambi a questa domanda.
- La correlazione multipla è la correlazione tra i valori $y$ osservati e i corrispondenti valori ricavati dall'equazione lineare di previsione:

$$\hat{y} = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

cioè la correlazione ordinaria tra le coppie di valori osservati della variabile risposta, $y$, e quelli teorici, $\hat{y}$, provenienti dal modello lineare di previsione per ciascuna delle $n$ unità del campione osservato

# Esempio: disagio mentale previsto sulla base di eventi di vita e stato economico-sociale

- La correlazione multipla è la correlazione tra le $n = 40$ coppie di valori $y$ osservati e previsti dal modello

| Unità | $y$ osserv. | $\hat{y} = 28,23 + 0,103x_1 - 0,097x_2$ |
|---|---|---|
| 1 | 17 | $24,8 = 28,23 + 0,103(46) - 0,097(84)$ |
| 2 | 19 | $22,8 = 28,23 + 0,103(39) - 0,097(97)$ |
| 3 | 20 | $28,7 = 28,23 + 0,103(27) - 0,097(24)$ |
| ... | | |

- Per l'intero campione, $R = 0,58$
- Le correlazioni bivariate con la $y$ sono 0,37 per $x_1$ e -0,40 per $x_2$

# Esempio sul disagio mentale (segue)

- **Il Coefficiente di Determinazione Multiplo** è la riduzione relativa (quota di riduzione) nell'errore totale che si ottiene stimando (prevedendo) la $y$ mediante il modello lineare anziché utilizzando soltanto la sua media campionaria

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{\sum(y - \bar{y})^2 - \sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

**Esempio:**

| *Var.esplic.* | *TSS* | *SSE* | $R^2$ |
|---|---|---|---|
| $x_1$ | 1162,4 | 1001,4 | 0,14 |
| $x_2$ | 1162,4 | 977,7 | 0,16 |
| $x_1, x_2$ | 1162,4 | 768,2 | 0,34 |

- Per il modello di regressione multipla:

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{1162,4 - 768,2}{1162,4} = 0,339$$

# Esempio sul disagio mentale (segue)

- I software per la regressione multipla forniscono la tavola dell'ANOVA con le somme dei quadrati utilizzate per l'$R^2$ e la tavola riassuntiva del modello (Model Summary) con i valori di $R$ e $R^2$

Tavola dell'ANOVA

| Modello | Somma dei quadrati | gdl | Media quad. | F | Sign. |
|---|---|---|---|---|---|
| Regressione | 394,238 | 2 | 197,119 | 9,495 | 0,000 |
| Residuo | 768,162 | 37 | 20,761 | | |
| Totale | 1162,400 | 39 | | | |

Model Summary

| $R$ | $R^2$ | $R^2$ corretto | Errore standard delle stime |
|---|---|---|---|
| 0,582 | 0,339 | 0,303 | 4,556 |

# Esempio sul disagio mentale (segue)

- $R^2 = 0,34$ significa che si riduce del 34% l'errore che si commette nel prevedere il disagio mentale sulla base del modello lineare con regressori 'eventi di vita' e 'stato economico-sociale' rispetto ad una previsione basata soltanto sulla media campionaria $\bar{y}$
- Un modo alternativo per esprimere questo, è: la variabilità complessiva del disagio mentale è spiegata per il 34% dalle variabili eventi di vita e stato economico-sociale
- La correlazione multipla $\sqrt{R^2} = \sqrt{0,34} = 0,58$ è la correlazione tra le 40 coppie di valori $y$ osservati e $\hat{y}$ previsti mediante il modello

# Proprietà di $R$ e $R^2$

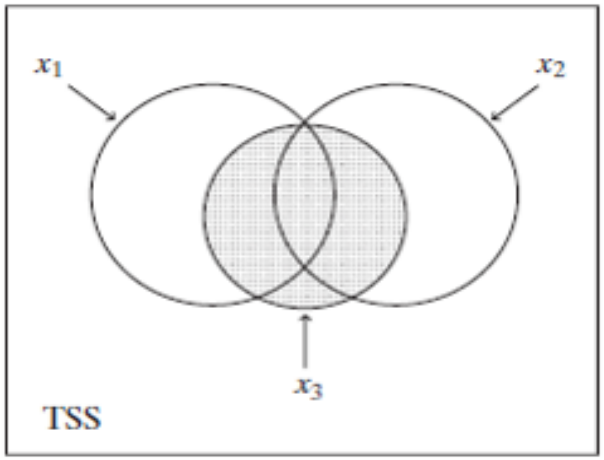
- $0 \leq R^2 \leq 1$
- $R = +\sqrt{R^2}$ e pertanto $0 \leq R \leq 1$ (cioè non può essere negativo)
- Maggiore è il loro valore e migliore è la capacità previsiva delle variabili esplicative del modello
- $R^2 = 1$ quando tutti i valori $y$ osservati sono uguali a quelli previsti. In questo caso abbiamo ovviamente $SSE = 0$
- $R^2 = 0$ quando tutti i valori previsti, $\hat{y}$ sono uguali alla media, $\bar{y}$ e, quindi, $TSS = SSE$. Quando accade questo, $b_1 = b_2 = \cdots = b_k = 0$ e, ovviamente, anche il coefficiente di correlazione $r$ tra la $y$ e ciascun regressore $x$ è uguale a 0
- $R^2$ **non può diminuire** se aggiungiamo una variabile esplicativa al modello.

# Proprietà di $R$ e $R^2$ (segue)

- Il numeratore di $R^2$, $TSS - SSE$ è chiamato somma dei quadrati di regressione ($RSS$), e rappresenta la variabilità dei valori $y$ previsti o 'spiegati' dal modello
- $R^2$ è additivo: cioè è uguale alla somma degli $r^2$ bivariati quando le coppie di variabili esplicative sono tra loro incorrelate. Ciò sarebbe sempre auspicabile, ma difficilmente si realizza nelle indagini osservazionali
- Lo stimatore campionario di $R^2$ è distorto e tende a sovrastimare il valore di $R^2$ nella popolazione. Il software riporta il calcolo dell'$R^2$ "corretto". Si tratta in realtà di un coefficiente che è affetto da una distorsione minore e che ha un valore di norma inferiore a quello dell'$R^2$ classico.

# Multicollinearità

- $R^2$ non può diminuire se aggiungiamo una variabile al modello
- Tuttavia se la variabile che si aggiunge è fortemente correlata con una o più variabili già presenti nel modello, l'$R^2$ potrebbe rimanere invariato o quasi invariato
- Questa situazione, che è qui illustrata graficamente, è detta di **multicollinearità**



# Inferenza sui parametri del modello di regressione multipla

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- L'inferenza sui parametri del modello si basa su una serie di assunzioni:
  - ✓ Il modello deve essere (approssimativamente) adeguato
  - ✓ La distribuzione nella popolazione della $y$ condizionatamente a ciascuno dei regressori $x_1, \cdots, x_k$ è normale
  - ✓ La deviazione standard della distribuzione condizionata della $y$ è costante per ogni combinazione di valori $x_1, \cdots, x_k$
  - ✓ Il campione è selezionato in modo casuale
- Queste assunzioni raramente sono completamente soddisfatte
- Tuttavia, se la prima è rispettata, l'inferenza di tipo bidirezionale è *robusta* rispetto a violazioni non pesanti di normalità e di variabilità costante della distribuzione condizionata della $y$

# Test sull'influenza complessiva delle variabili esplicative

- Poiché i coefficienti di regressione esprimono l'influenza delle variabili esplicative sulla variabile risposta, l'inferenza può riguardare la loro influenza complessiva o quella di ciascuna variabile
- per valutare l'influenza complessiva, si formulano le seguenti ipotesi:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

cioè che $y$ sia linearmente indipendente da ciascuna delle variabili esplicative;

$$H_a : \text{almeno un } \beta_i \neq 0; \ \ (i = 1, ..., k)$$

contro l'alternativa che almeno una di queste abbia influenza sulla $y$

- É del tutto equivalente sottoporre a test le ipotesi:

$$H_0 : R^2 = 0; \quad H_a : R^2 > 0$$

# Il test F

- La statistica test per la verifica delle ipotesi appena viste è:

$$F = \frac{R^2/k}{(1-R^2)/\left[n-(k+1)\right]}$$

- Quando l'ipotesi nulla è vera la statistica segue la distribuzione $F$ (R. A. Fisher)

- Maggiori valori di $R$ producono maggiori valori di $F$, cioè maggiore evidenza contro l'ipotesi nulla

# Proprietà della distribuzione $F$

- La $F$ può assumere solo valori non negativi
- La forma della distribuzione è asimmetrica positiva
- La sua media è approssimativamente uguale a 1 (l'approssimazione migliora al crescere di $n$)
- La distribuzione dipende da una coppia di gradi di libertà:

  $gdl_1 = k$ numero di variabili esplicative

  $gdl_2 = [n - (k + 1)]$ dimensione campione − n. parametri nel modello
- la tavola della $F$ riporta, per alcune probabilità della coda della distribuzione, quali, 0,05, 0,01, 0,001 ecc., i valori di $F$ ($F$-score) relativi a diverse combinazioni di $gdl_1$ e $gdl_2$

# Esempio: il disagio mentale è linearmente indipendente da Eventi e SES?

- $H_0 : \beta_1 = \beta_2 = 0$ (indipendenza lineare)
- $H_a : \beta_1 \neq 0$ oppure $\beta_2 \neq 0$ o entrambi diversi da 0
- Test statistico:

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{0,339/2}{(1-0,339)/[40-(2+1)]} = 9,5$$

- per $\alpha = 0,001$, $gdl_1 = 2$ e $gdl_2 = 37$, abbiamo $F_{2;37} = 8,37$
- Di conseguenza: $P(F > F_{2;37}) < 0,001$

# Esempio sul disagio mentale (segue)

- $F = 9,5$ ci dice che c'è una forte evidenza che almeno una delle due variabili esplicative sia associata con il disagio mentale
- il calcolo della statistica test $F$, in alternativa può essere effettuato utilizzando la tavola dell'ANOVA

$$F = \frac{197,119}{20,761} = 9,5$$

Tavola dell'ANOVA

| Modello | Somma dei quadrati | gdl | Media quad. | F | Sign. |
|---|---|---|---|---|---|
| Regressione | 394,238 | 2 | 197,119 | 9,495 | 0,000 |
| Residuo | 768,162 | 37 | 20,761 | | |
| Totale | 1162,400 | 39 | | | |

# Inferenza per singoli coefficienti di regressione

- Le variabili esplicative del modello sono tutte necessarie?
- Per valutare l'effetto parziale di $x_i$, controllando per le altre variabili del modello, si sottopone a test l'ipotesi nulla $H_0 : \beta_i = 0$ utilizzando la statistica test

$$t = (b_i - 0)/se; \text{ con } gdl = n - (k + 1),$$

gli stessi $gdl_2$ del test $F$, che possiamo anche trovare nella tavola della ANOVA, in corrispondenza della riga: *Residuo* del modello

- L'$IC$ per $\beta_i$ assume la forma $b_i \pm t(se)$ con il t-score in corrispondenza di $gdl = n - (k + 1)$
- I principali software per la regressione forniscono: stime dei coefficienti, errore standard, $t$-test e $p$-valore, usualmente riferito al test bidirezionale

# Esempio: effetto di *SES* sul disagio mentale, controllando per gli *Eventi*

- Sottoponiamo a test: $H_0 : \beta_2 = 0, \quad H_a : \beta_2 \neq 0$
- Il risultato del test $t$ è disponibile dalla tavola fornita dal software SPSS (ma anche da altri software ) che riassume le caratteristiche del modello di regressione stimato
- Possiamo concludere, osservando i dati della tavola, che c'è un'evidenza molto forte che la variabile *SES* eserciti un effetto riduttivo del disagio mentale ($p - valore < 0,002$). E, allo stesso modo, che la variabile *Eventi* ne abbia uno positivo ($p - valore < 0,003$)

| | Coeff. non standard. | | Coeff. standard. | | | Intervallo al 95% | |
|---|---|---|---|---|---|---|---|
| Modello | $b$ | stand.err. | $b^*$ | $t$ | sig. | $L_i$ | $L_s$ |
| Costante | 28,23 | 2,174 | | 12,984 | 0,000 | 23,824 | 32,635 |
| Eventi | 0,103 | 0,032 | 0,428 | 3,177 | 0,003 | 0,037 | 0,169 |
| SES | -0,97 | 0,029 | -0,451 | -3,351 | 0,002 | -0,156 | -0,039 |

# Attenzione alla multicollinearità

- Perché preoccuparsi di fare il test $F$? non potremmo passare direttamente al $t$-test sui singoli coefficienti?
- No; è possibile che dal test $F$ si ottenga un piccolo $P - valore$ mentre nessuno dei test $t$ produca un analogo risultato.
- É ugualmente possibile che si ottenga un piccolo $P - valore$ nella regressione binaria per una variabile esplicativa e non altrettanto se questa variabile viene controllata per altre variabili
- Ciò accade in presenza di multicollinearità. In questo caso la variabilità parziale spiegata da una singola variabile è piccola (il valore di una esplicativa può essere facilmente previsto dalle altre esplicative)
- Ovviamente assurdo, ma chiarificatore è il modello: $y=$ statura; $x_1=$ lunghezza gamba destra; $x_2=$ lunghezza gamba sinistra

# In presenza di multicollinearità

- L' errore standard di ciascun coefficiente di regressione può essere abbastanza elevato e, conseguentemente, non significativo il relativo $t$-test
- $R^2$ può mantenersi elevato anche se vengono eliminate una o più variabili esplicative
- É consigliabile semplificare il modello eliminando variabili esplicative che non apportano sostanziali benefici al modello
- Esistono strumenti diagnostici (VIF - fattori di incremento della varianza) per valutare la presenza e l'entità della multicollinearità

# Interazione tra i regressori del modello

- Il modello

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

  assume che ciascun coefficiente di regressione parziale $\beta_i$ resti costante per qualunque combinazione dei rimanenti regressori $x_j; (j \neq i)$.
- Ciò equivale a assumere che nel modello non vi sia *interazione* tra le variabili esplicative (come nell'esempio su Disagio mentale, Eventi e SES)
- Se c'è interazione tra le variabili $x_1$ e $x_2$ allora l'effetto di $x_1$ sulla variabile dipendente, $y$, può variare al variare di $x_2$

# Il più semplice modello di interazione

- Esempio: con $k = 2$:

$$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2)$$

che può essere considerato come un caso speciale del modello:

$$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

con $x_3 = x_1 x_2$

- Se inseriamo il termine di interazione nel modello sul disagio mentale, $R^2$ passa da 0,339 (senza interazione) a 0,347 e otteniamo la seguente espressione:

$$\hat{y} = 26 + 0,156 x_1 - 0,060 x_2 - 0,00087 x_1 x_2$$

# Il più semplice modello di interazione (segue)

- Il controllo per $x_2$ produce i seguenti risultati:

| $x_2$ (costante) | Equazione di previsione per $y$ e $x_1$ |
|---|---|
| 0 | $26 + 0,156x_1 - 0,060(0) - 0,00087x_1(0)$ $=26+0,16x_1$ |
| 50 | $26 + 0,156x_1 - 0,060(50) - 0,00087x_1(50)$ $=23+0,11x_1$ |
| 100 | $26+0,156x_1-0,060(100)-0,00087x_1(100)$ $=20+0,07x_1$ |

- E' facile verificare che al crescere di SES (variabile $x_2$) decresce l'effetto degli 'Eventi di vita' (variabile $x_1$ ) sulla variabile risposta $y$, disagio mentale

# Test sul termine di interazione

- Per sottoporre a test l'ipotesi nulla: $H_0$: non c'è nessuna interazione nel modello $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$, si sottopone a test, $H_0 : \beta_3 = 0$, per mezzo del test

$$t = b_3/se$$

- Nell'esempio sul disagio mentale:
  $t = -0,00087/0,0013 = -0,67$
  $gdl = n - 4 = 36$, $P - valore = 0,51$ per $H_a : \beta \neq 0$
  evidenza insufficiente per concludere che ci sia interazione.
  Ma con il data set completo, di oltre 1000 oservazioni, si ottiene un valore significativo del test $t$.

- In un modello con $k > 2$ l'interazione può riguardare tutte le possibili coppie di variabili:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$$

# Confronto tra modelli

- Come possiamo valutare se un modello ci garantisce un 'fit' migliore rispetto ad un altro più seplice, con un numero di variabili esplicative inferiore?
- Confrontiamo, per fare un esempio, i seguenti modelli:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$$

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- E sottoponiamo a test l'ipotesi nulla di assenza di interazione formulandola come segue:

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$$

# Confronto tra modelli (segue)

- Il confronto tra modelli può essere effettuato con un test $F$ confrontando le $SSE$ dei due modelli o, equivalentemente, i loro $R^2$.
- Il modello più complesso, che chiamiamo **completo**, sarà migliore se il valore di $SSE$ sarà sufficientemente inferiore a quello del modello più semplice, che chiamiamo **ridotto** o, equivalentemente, se il suo $R^2$ sarà sufficientemente maggiore
- Indichiamo rispettivamente con $SSE_c$ e con $SSE_r$ le devianze di errore dei modelli completo e ridotto e utilizziamo in modo analogo la notazione $R_c^2$ e $R_r^2$ per i loro $R^2$

$$F = \frac{(SSE_r - SSE_c)/gdl_1}{SSE_c/gdl_2} = \frac{(R_c^2 - R_r^2)/gdl_1}{(1 - R_c^2)/gdl_2}$$

$gdl_1$ = **differenza** tra il numero dei parametri nei due modelli
$gdl_2 = n - (k + 1)$ con $k$ = numero di variabili esplicative del modello completo

# Studio sul disagio mentale ($n = 40$)

- Modello ridotto: $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$
- Modello completo: $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
  con la variabile $x_3 = x_1 \cdot x_2$
- Con un solo parametro aggiuntivo: $gdl_1 = 1$ e
  $gdl_2 = n - (k + 1) = 40 - (3 + 1) = 36$
- La somma degli errori al quadrato per il modello completo è $SSE_c = 758,8$ e per il modello ridotto, $SSE_r = 768,2$

$$F = \frac{(SSE_r - SSE_c)/gdl_1}{SSE_c/gdl_2} = \frac{9,4/1}{758,8/36} = 0,45$$

# Studio sul disagio mentale (segue)

- In modo del tutto equivalente possiamo ricavare la statistica $F$ dai valori: $R_c^2 = 0,347$ e $R_r^2 = 0,339$

$$F = \frac{(R_c^2 - R_r^2)/gdl_1}{(1 - R_c^2)/gdl_2} = \frac{(0,347 - 0,339)/1}{(1 - 0,347)/36} = 0,45$$

- Il valore di $F$ corrisponde ad un $P - valore$ pari a 0,51; c'è quindi un'evidenza troppo bedole che il modello completo sia migliore e, di conseguenza, sembra opportuno adottare il modello ridotto.
- Poiché $gdl_1 = 1$, il test $F$ è uguale al quadrato della $t$ con $n - (k + 1) = 36 gdl$. Pertanto, è ancora equivalente calcolare:

$$t = \frac{b_3}{se} = \frac{-0,00087}{0,0013} = -0,67$$

# Esempio: Studio sul disagio mentale (segue)

- Nel confronto tra modelli non è necessario che i regressori aggiuntivi siano termini di interazione
- Modello ridotto: $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$
- Modello completo: $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
  con la variabile $x_3 =$ numero di partecipazioni a funzioni religiose nell'ultimo anno
- $R_r^2 = 0,339$; $R_c^2 = 0,358$

$$F = \frac{(R_c^2 - R_r^2)/gdl_1}{(1 - R_c^2)/gdl_2} = \frac{(0,358 - 0,339)/1}{(1 - 0,358)/36} = 1,07$$

con $gdl_1 = 1; gdl_2 = 36$; $P - valore = 0,31$

- Non è possibile respingere $H_0$ al livello $\alpha = 0,05$. Di conseguenza consideriamo adeguato il modello più semplice (ridotto)