IL MODELLO DI REGRESSIONE LINEARE MULTIPLA

19

19.1 Introduzione

Nel capitolo 16 abbiamo illustrato il modello di regressione lineare semplice, dove una variabile risposta Y è supposta dipendere soltanto da una variabile esplicativa X. Nella maggior parte dei casi, tuttavia, questa assunzione non è adeguata poiché Y potrebbe essere influenzata da più di una variabile, diciamo $X_1, X_2, \cdots, X_k$, con $k > 1$.

I concetti di relazione funzionale e relazione statistica devono essere estesi a questa situazione. Diremo che esiste una relazione funzionale tra Y e $X_1, X_2, \cdots, X_k$ se e solo se un valore di Y corrisponde a una data combinazione dei valori delle variabili esplicative. Per esempio, se X_1 e X_2 indicano la lunghezza dei due lati adiacenti di un rettangolo e Y è il suo perimetro, allora vale la seguente relazione funzionale lineare. $Y = 2X_1 + 2X_2$.

In generale una relazione funzionale di tipo lineare può essere espressa come:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

dove $\beta_0, \beta_1, \cdots, \beta_k$ sono i coefficienti.

In particolare, β_0 è detto intercetta, cioè il valore di Y quando $X_1 = X_2 = \cdots = X_k = 0$; mentre β_j (per $j = 1, \cdots, k$) esprime l'incremento di Y corrispondente a un incremento unitario di X_j, avendo fissato le restanti variabili.

Come abbiamo già visto per la regressione lineare semplice, negli studi empirici la relazione che può essere osservata tra le variabili Y e $X_1, X_2, \cdots, X_k$ non è mai una relazione matematica esatta: infatti ad una determinata combinazione di valori delle variabili esplicative possono corrispondere più valori di Y. In tali circostanze, diremo che tra la variabile risposta e le variabili esplicative sussiste una relazione statistica. Ad esempio:

- se Y è la *spesa annuale per consumi*, X_1 è il *prodotto interno lordo* (PIL), X_2 è la *Popolazione* e X_3 è *tasso di disoccupazione*, può accadere che per anni in cui si osservino gli stessi valori del PIL, della Popolazione e del Tasso di disoccupazione possano corrispondere valori diversi di spesa Y;
- se Y è il *salario d'impiego*, X_1 sono gli *anni di impiego nel lavoro attuale*, X_2 è

l'*ultimo titolo di studio conseguito*, X_3 è il *numero di componenti della famiglia*, a impiegati con gli stessi valori delle variabili esplicative possono corrispondere diversi valori del salario.

Come abbiamo già visto nel cap.16, per descrivere e analizzare i fenomeni empirici è opportuno introdurre una relazione più complessa di quella funzionale che prende il nome di **relazione statistica**.

Una **relazione statistica** può essere descritta dalla seguente equazione:

$$Y = f(X_1, X_2, \cdots, X_k) + \varepsilon \qquad (19.1.1)$$

in cui $f(X_1, X_2, \cdots, X_k)$ è la **funzione di regressione**, che esprime il contributo delle variabili esplicative al valore della variabile risposta Y mentre ε rappresenta il contributo di tutti gli altri fattori, non osservati, in grado di influenzare la risposta ed è quindi una variabile casuale.

In analogia con quanto già detto per la regressione lineare semplice, nel prossimo paragrafo introdurremo alcune assunzioni.

Per mostrare graficamente la relazione che intercorre tra due o più variabili esplicative e la variabile risposta può essere utilizzata una **matrice di diagrammi a dispersione**. Questa consiste in una serie di diagrammi di dispersione che descrivono tutte le possibili coppie di variabili selezionate da $(Y, X_1, X_2, \cdots, X_k)$. In tal modo è possibile visualizzare la relazione tra la Y e ogni singola variabile esplicativa e se, ad esempio, è di tipo lineare o non lineare. Tuttavia, l'uso di tale grafico non è idoneo a investigare se e come due o più variabili esplicative possono influire congiuntamente sulla variabile risposta.

ESEMPIO 19.1.1 – *Matrice di diagrammi di dispersione*

La seguente figura riporta il grafico a matrice di diagrammi di dispersione corrispondente alle variabili *Spesa annuale per consumi* (Y), *Prodotto interno lordo* (X_1), *Popolazione* (X_2) e *Tasso di disoccupazione* (X_3) osservati negli Stati Uniti tra il 1959 e il 1999 (dataset contenuto nel file *PIL*).

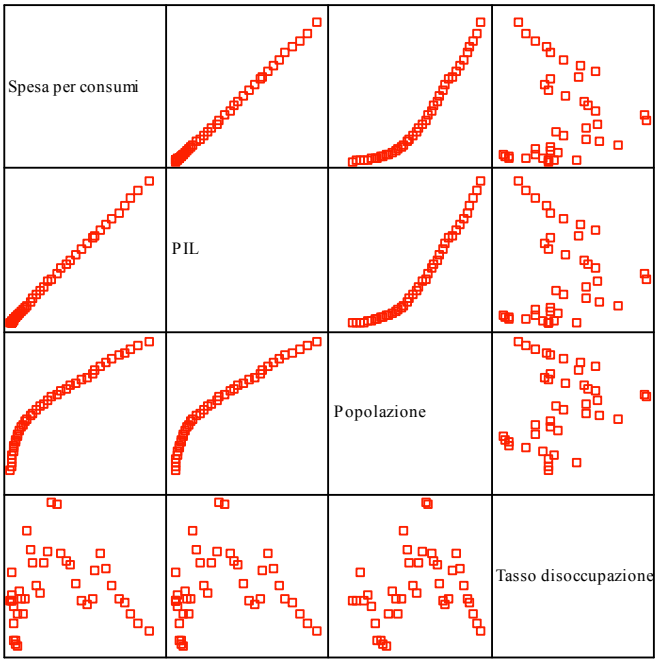

Figura 19.1.1 – *Matrice di diagrammi di dispersione*

In ogni quadrante è presente un diagramma di dispersione tra la variabile indicata nella corrispondente colonna con quella indicata nella corrispondente riga. Dal grafico si può notare che tra X_1 e X_2 sussiste un forte legame. Infatti, all'aumentare della Popolazione aumenta il PIL. Da ciò si potrebbe ritenere che le due variabili contengano la stessa informazione circa la Y e che quindi una delle due potrebbe essere ridondante nello spiegare la variabile dipendente.

Nei successivi paragrafi saranno introdotti i concetti principali alla base del modello di regressione lineare multipla, in particolare, il metodo di stima, la bontà d'adattamento, l'inferenza sui parametri e la multicollinearità. Rimandiamo il lettore a testi più specialistici per quanto riguarda l'approfondimento sulla generalizzazione del modello a variabili esplicative non solo quantitative e alla verifica delle assunzioni del modello. Dal terzo paragrafo in poi è richiesta la conoscenza degli elementi di base dell'algebra matriciale.

19.2 il modello di regressione lineare multipla

Tra tutti i possibili modelli di regressione multipla il più semplice è quello noto come **modello di regressione lineare multipla**. Le assunzioni del modello di regressione lineare multipla si riferiscono al processo che genera le n osservazioni disponibili composte ognuna da $k+1$ valori, $(x_{11}, x_{12}, \cdots, x_{1k}, y_1), (x_{21}, x_{22}, \cdots, x_{2k}, y_2), \ldots, (x_{n1}, x_{n2}, \cdots, x_{nk}, y_n)$ e sono le seguenti:

Assunzione 1. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$, per ogni osservazione $i = 1, \ldots, n$;

Assunzione 2. le ε_i sono v.c. indipendenti con valore atteso $E(\varepsilon_i) = 0$ e varianza costante $V(\varepsilon_i) = \sigma^2$ per ogni $i = 1, \ldots, n$ indipendentemente dai valori delle X_j (con $j = 1, .., k$);

Assunzione 3. i valori x_{ij} (per $i = 1, \ldots, n$) delle variabili esplicative X_j (per $j = 1, .., k$) sono noti senza errore.

La prima assunzione implica che tra le possibili funzioni di regressione $f(X_1, X_2, \cdots, X_k)$ che possono descrivere il legame tra la variabile dipendente e le variabili esplicative, si è scelta la funzione lineare. Ogni ε_i è una variabile casuale poiché rappresenta gli scostamenti di Y_i dal suo valore atteso. Si assume allora che le v.c. ε_i siano tra loro statisticamente indipendenti, con valore atteso uguale a zero e varianza costante pari a σ^2, tenendo fissate le variabili esplicative. La condizione di varianza costante del termine di errore ε_i corrisponde appunto all'ipotesi di **omoschedasticità**.

NOTA Per il modello di regressione è sufficiente assumere che le ε_i siano tra loro incorrelate, ossia che la covarianza tra ε_i e ε_j sia nulla per ogni $i \neq j$; tuttavia, per semplicità di esposizione, qui assumiamo l'indipendenza tra gli errori.

Poiché ε_i è una variabile casuale, anche la variabile dipendente Y_i, somma di una combinazione lineare di componenti deterministiche e di una stocastica, è una variabile casuale. Ora, per ogni $x_{i1}, x_{i2}, \ldots, x_{ik}$, $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ è una costante e $E(\varepsilon_i) = 0$ (assunzioni 2 e 3); da ciò discende che il valore atteso di Y_i condizionato al valore $X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_k = x_{ik}$ è:

$$E(Y_i | X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_k = x_{ik}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \quad (19.2.1)$$

La 19.2.1 rappresenta l'equazione di un iperpiano in uno spazio a $k+1$ dimensioni. Tale iperpiano prende il nome di **superficie di regressione** o di risposta e la sua posizione nello spazio dipende dai valori assunti dai **coefficienti di regressione**. In particolare, β_0 è l'intercetta e rappresenta il valore della risposta media $E(Y_i)$ quando $x_{i1} = x_{i2} = \ldots = x_{ik} = 0$, mentre β_j (per $j = 1, \ldots, k$) indica l'incremento di $E(Y_i)$ corrispondente ad un incremento unitario di X_{ij}, tenendo fisso il valore delle altre variabili esplicative.

ESEMPIO 19.2.1 – *Superficie di regressione*

Si consideri un modello con due variabili esplicative ($k = 2$) e $\beta_0 = 5$, $\beta_1 = -2$ e $\beta_2 = 1$. In questo caso si ottiene come superficie di regressione un piano di equazione $E(Y) = 5 - 2X_1 + X_2$ collocato in uno spazio tridimensionale. Il piano è mostrato nella seguente figura.

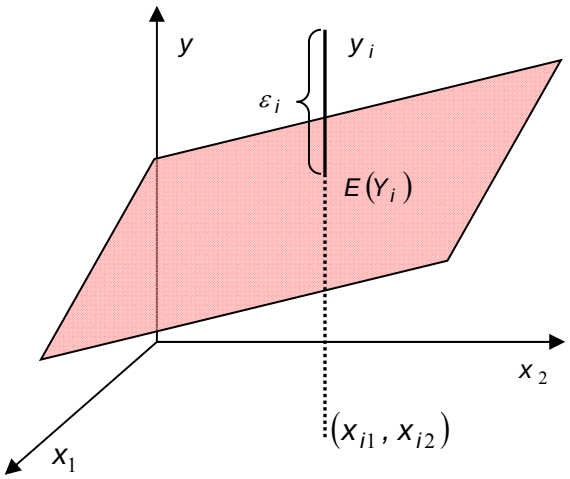

Figura 19.2.1 – *Superficie di regressione*

Quando $X_1 = X_2 = 0$, il valore di $E(Y)$ è pari a 5. Inoltre, $E(Y)$ decresce di 2 quando X_1 incrementa di 1 e X_2 è tenuta costante; similmente, $E(Y)$ incrementa di 1 in corrispondenza di un incremento unitario di X_2, tenuta costante X_1. Si può osservare che per l'i-esima osservazione la differenza tra il valore della variabile dipendente y_i e il suo valore atteso $E(Y_i) = 5 - 2x_{i1} + x_{i2}$ rappresenta il termine di errore ε_i.

Per quanto detto finora, le Y_i, essendo funzioni delle variabili casuali ε_i, hanno valore atteso pari a $E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ e varianza pari a $V(Y_i) = V(\varepsilon_i) = \sigma^2$. Pertanto, un modo equivalente di formulare il modello di regressione lineare multipla che tiene conto delle assunzioni 1-3, è dato da:

Le osservazioni y_i sono realizzazioni di variabili casuali indipendenti con valore atteso $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ e varianza comune σ^2.

Le procedure inferenziali sui parametri del modello di regressione lineare semplice richiedono, come abbiamo già visto nel capitolo 16, l'introduzione dell'assunzione di normalità delle variabili casuali ε_i, ossia che:

Assunzione 4. $\varepsilon_i \sim N(0, \sigma^2)$ per $i = 1, \ldots, n$.

Tenendo conto di questa ulteriore assunzione, il modello di regressione lineare multipla può essere definito nel modo seguente:

Le osservazioni y_i sono realizzazioni di variabili casuali Normali con valore atteso $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ e varianza σ^2, ossia

$$Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \sigma^2) \text{ per } i = 1, \ldots, n.$$

L'assunzione di Normalità del termine di errore non risulta generalmente troppo restrittiva nel senso che vale, almeno approssimativamente, in molti fenomeni reali.

NOTA Tra le assunzione fatte per il modello, la prima potrebbe sembrare troppo restrittiva per l'applicabilità del modello giacché presuppone che la superficie di risposta sia un iperpiano. Tuttavia, si deve notare che il modello impone solamente la **linearità dei suoi parametri** rendendolo molto più flessibile di quanto possa apparire. Si può infatti facilmente dimostrare che il **modello di regressione polinomiale** e il modello che presenta **uno o più termini di interazione** sono in sostanza ancora modelli di regressione lineare. Per esempio, consideriamo il modello:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i1} x_{i2} + \varepsilon_i$$

che include una variabile elevata al quadrato e il prodotto tra due variabili per tener conto di un possibile effetto d'interazione tra queste due. Questo modello è ancora un modello di regressione lineare multipla. Infatti, operando le sostituzioni $x_{i3} = x_{i1}^2$ e $x_{i4} = x_{i1} x_{i2}$, si può riscrivere nel seguente modo: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$.

Spesso anche quando la funzione di regressione non è lineare nei parametri si può ricondurla al caso lineare attraverso un'appropriata trasformazione.

Per esempio, $Y_i = \beta_0 x_{i1}^{\beta_1} \varepsilon_i$ si può ricondurre a un modello lineare nei parametri attraverso la trasformazione logaritmica, infatti: $log(Y_i) = log(\beta_0 x_{i1}^{\beta_1} \varepsilon_i) = log\, \beta_0 + \beta_1 log\, x_{i1} + log\, \varepsilon_i$ e ponendo $Y_i' = log(Y_i)$, $\beta_0' = log\, \beta_0$, $x_{i1}' = log\, x_{i1}$ e $\varepsilon_i' = log\, \varepsilon_i$ si ottiene il modello lineare $Y_i' = \beta_0' + \beta_1 x_{i1}' + \varepsilon_i'$.

19.3 Il modello di regressione lineare multipla in forma matriciale

Il modello di regressione lineare multipla può essere riformulato utilizzando l'algebra matriciale. L'uso dell'algebra matriciale permette di sintetizzare più facilmente i risultati. Indichiamo con $\mathbf{Y}$, $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$ i seguenti vettori colonna corrispondenti ai valori della variabile risposta, ai parametri del modello e ai termini di errore

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

e sia $\mathbf{X}$ la matrice:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

dove della seconda colonna in poi sono riportati i valori delle variabili esplicative osservati sulle *n* unità statistiche.
Le assunzioni del modello possono essere riscritte nel seguente modo:

Assunzione 1.	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$;
Assunzione 2.	$\boldsymbol{\varepsilon}$ è un vettore casuale di componenti indipendenti con valore atteso $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ e matrice di varianza-covarianza $V(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ (dove $\mathbf{I}$ indica la matrice identità di dimensione $n \times n$);
Assunzione 3.	$\mathbf{X}$ è una matrice di costanti e possiede rango pieno.

Poiché $\mathbf{X}$ è una matrice di dimensione $n \times (k+1)$ e $\boldsymbol{\beta}$ è un vettore colonna di dimensione $(k+1) \times 1$, allora $\mathbf{X}\boldsymbol{\beta}$ è un vettore colonna di dimensione $n \times 1$, dove l'*i*-esimo elemento è pari a $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ e dunque la prima assunzione corrisponde a quella già vista nel paragrafo precedente. Osserviamo che $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ è equivalente a scrivere $E(\varepsilon_i) = 0$ per ogni $i = 1,\ldots,n$ e che la matrice di varianza-covarianza $V(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ è equivalente a

$$V(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

ossia che $V(\varepsilon_i) = \sigma^2$ per ogni $i = 1,\ldots,n$ e che $Cov(\varepsilon_i, \varepsilon_j) = 0$ per ogni $i = 1,\ldots,n$, con $j \neq i$, e cioè che i termini di errore sono incorrelati. L'ultima assunzione è equivalente a quella del paragrafo precedente solo che è messa in forma matriciale. L'assunzione sul rango pieno della matrice $\mathbf{X}$ è necessaria per ottenere la stima del vettore dei parametri $\boldsymbol{\beta}$ che richiede il calcolo della matrice $(\mathbf{X}'\mathbf{X})^{-1}$ che può essere effettuato solo se la $\mathbf{X}$ possiede rango pieno.

In conseguenza a quanto illustrato, i risultati sulle variabili aleatorie Y_i possono essere espressi in forma matriciale:

Il vettore aleatorio $\mathbf{Y}$ possiede valore atteso e matrice di varianza-covarianza pari a:

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}\,;\; V(\mathbf{Y}) = V(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I} \qquad (19.3.1)$$

Alle tre precedenti assunzioni si può aggiungere l'assunzione sulla normalità degli errori

Assunzione 4. $\boldsymbol{\varepsilon}$ si distribuisce come una Normale multivariata con vettore dei valori attesi pari a $\boldsymbol{\mu} = \mathbf{0}$ e matrice di varianza-covarianza pari a $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$.

NOTA La distribuzione Normale multivariata è una generalizzazione della distribuzione Normale al caso di due o più variabili. Indichiamo con $\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ un vettore aleatorio $\mathbf{z}$ che si distribuisce come una Normale multivariata con $\boldsymbol{\mu}$, vettore dei valori attesi, e $\boldsymbol{\Sigma}$, matrice di varianza-covarianza, i cui elementi costituiscono i parametri della distribuzione. Dall'assunzione 4. segue che il vettore aleatorio $\mathbf{Y}$ si distribuisce come $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ e ciò implica l'indipendenza delle Y_i e che $Y_i \sim N(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2)$ per $i=1,\ldots,n$.

19.4 Stima puntuale dei parametri

Il metodo di **stima dei minimi quadrati** descritto nel paragrafo 16.4 può essere applicato anche in questo caso più generale.

Il metodo dei minimi quadrati richiede la minimizzazione rispetto al vettore dei parametri $\boldsymbol{\beta}$ della funzione di perdita

$$G(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \qquad (19.4.1)$$

La funzione $G(\boldsymbol{\beta})$ è sempre non negativa ed è uguale a zero solamente quando per ogni osservazione il valore stimato $\hat{y}_i$ uguaglia il valore osservato y_i. La funzione aumenta al crescere della distanza tra i valori osservati e i valori stimati e pertanto, nell'individuare il miglior modello, si minimizza tale distanza rispetto al vettore di parametri $\boldsymbol{\beta}$. Si dimostra che

la **stima dei minimi quadrati ordinari** di $\boldsymbol{\beta}$, se esiste $(\mathbf{X}'\mathbf{X})^{-1}$, è data da

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (19.4.2)$$

NOTA La stima dei minimi quadrati ordinari è spesso denotata con OLS, abbreviazione del termine inglese *Ordinary Least Squares*, per differenziarla dal metodo dei minimi quadrati ponderati (WLS – *Wheighted Least Squares*) utilizzato nel caso in cui non valga l'assunzione di omoschedasticità del termine di errore.

Per meglio comprendere la formula 19.4.2 consideriamo in dettaglio il caso più semplice in cui si considera una sola variabile esplicativa, ossia il caso di un modello di regressione lineare semplice. Consideriamo le seguenti matrici iniziali:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$ allora si ha

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \end{bmatrix} \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_{i1}^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_{i1} y_i \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1}^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1}^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_{i1} y_i \end{bmatrix} =$$

$$= \frac{1}{\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}^2} \begin{bmatrix} \bar{y}\sum_{i=1}^{n} x_{i1}^2 - \bar{x}\sum_{i=1}^{n} x_{i1} y_i \\ \sum_{i=1}^{n} x_{i1} y_i - n\bar{x}\bar{y} \end{bmatrix} = \begin{bmatrix} \bar{y} - \frac{\sigma_{XY}}{\sigma_X^2}\bar{x} \\ \frac{\sigma_{XY}}{\sigma_X^2} \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

ricordando che $n\bar{x} = \sum_{i=1}^{n} x_{i1}$, $\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}^2 = n\sigma_X^2$ e $\sum_{i=1}^{n} x_{i1} y_i - n\bar{x}\bar{y} = n\sigma_{XY}$. Pertanto si ottengono le stesse stime della 16.4.2.

ESEMPIO 19.4.1 – *Calcolo matriciale della stima dei parametri di un modello di regressione lineare semplice*

Riprendiamo i dati dell'esempio 16.2.3 e la stima dei parametri del modello di regressione lineare semplice riportati nell'esempio 16.4.2. Possiamo riformulare il problema in termini matriciali:

$$\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 47{,}8 & 27{,}9 & 36{,}6 & 54{,}2 & 41{,}9 & 44{,}4 & 54{,}3 & 42{,}3 & 48{,}2 & 41{,}5 & 43{,}2 & 56{,}3 & 63{,}3 & 46{,}8 & 45{,}2 & 38{,}7 & 36{,}3 & 39{,}5 & 30{,}9 & 52{,}6 \end{bmatrix}$$

$$\mathbf{y}' = [63{,}0 \quad 33{,}4 \quad 42{,}0 \quad 72{,}8 \quad 52{,}0 \quad 54{,}0 \quad 63{,}4 \quad 60{,}7 \quad 58{,}0 \quad 54{,}4 \quad 55{,}5 \quad 74{,}0 \quad 79{,}2 \quad 53{,}1 \quad 59{,}6 \quad 52{,}0 \quad 47{,}2 \quad 48{,}7 \quad 41{,}4 \quad 66{,}9]$$

e dunque

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 20 & 891{,}9 \\ 891{,}9 & 41223{,}99 \end{bmatrix} \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1{,}42180 & -0{,}03076 \\ -0{,}03076 & 0{,}00069 \end{bmatrix} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 1131{,}3 \\ 52269{,}82 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 1{,}42180 & -0{,}03076 \\ -0{,}03076 & 0{,}00069 \end{bmatrix} \begin{bmatrix} 1131{,}3 \\ 52269{,}82 \end{bmatrix} = \begin{bmatrix} 0{,}59486 \\ 1{,}25508 \end{bmatrix}$$

I calcoli matriciali possono essere eseguiti, ad esempio, attraverso le funzione del software Excel. In particolare, le funzioni MATR.PRODOTTO(.) e MATR.INVERSA(.) consentono rispettivamente di

calcolare il prodotto tra due matrici e l'inversa di una matrice. Come si può osservare, gli elementi del vettore dei parametri $\mathbf{b}$ corrispondono esattamente ai valori delle stime di β_0 e β_1 trovati nell'esempio 16.4.2.

I calcoli matriciali per la stima dei parametri del modello sono effettuati automaticamente dal software statistico utilizzato per analizzare i dati.

ESEMPIO 19.4.2 – *Stima dei parametri di un modello di regressione lineare multipla*

Consideriamo il file di dati *Impiegati* che contiene le determinazioni del *Sesso*, degli *Anni di istruzione*, della *Categoria lavorativa*, dello *Stipendio attuale*, e di quello *iniziale*, dei *Mesi trascorsi dall'assunzione* e del *Logaritmo dello stipendio attuale* relativamente a 474 dipendenti. Considerando il modello di regressione lineare multipla in cui si assume il *Logaritmo dello stipendio attuale* come variabile dipendente (Y) e gli *Anni di istruzione* (X_1) e i *Mesi trascorsi dall'assunzione* (X_2) come variabili esplicative, si ottengono le stime riportate nella seguente tabella:

Stima dei coefficienti

	Coefficiente	*Errore standard*	*t*	*p-value*
Costante	8,877	0,120	73,797	0,0000
X_1	0,096	0,005	21,043	0,0000
X_2	0,002	0,001	1,798	0,0728

Il modello di regressione stimato è: $\hat{y}_i = 8{,}877 + 0{,}096x_{i1} + 0{,}002x_{i2}$. Ciò significa, ad esempio, che il valore atteso del *Logaritmo dello stipendio attuale* aumenta di 0,096 aumentando di un anno gli *Anni di istruzione*, tenuto costante il numero di *Mesi trascorsi dall'assunzione*. Per individuare i valori di β_0 e β_1 che rendono minima la funzione di perdita $G(\beta_0, \beta_1)$, occorre calcolare le derivate parziali di tale funzione rispetto a β_0 e β_1 e porle uguali a zero. Dopo alcuni passaggi e semplificazioni si ottengono le stime dei coefficienti di regressione.

Lo stimatore dei minimi quadrati ordinari dei coefficienti di regressione, che per semplicità verrà indicato allo stesso modo del vettore delle stime, $\mathbf{b}$, è anche lo **stimatore di massima verosimiglianza**, sotto l'assunzione che il termine di errore è distribuito normalmente.
Come per il modello di regressione lineare semplice, dove gli stimatori dei coefficienti di regressione, B_0 e B_1, sono funzioni lineari delle Y_i, anche lo stimatore $\mathbf{b}$ è funzione lineare del vettore casuale $\mathbf{Y}$ giacché si può vedere come funzione lineare del tipo $\mathbf{b} = \mathbf{AY}$ con $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Possiamo pertanto enunciare le seguenti proprietà:

Proprietà degli stimatori dei minimi quadrati ordinari

1. $\mathbf{b}$ è uno stimatore corretto di $\boldsymbol{\beta}$, ossia $E(\mathbf{b}) = \boldsymbol{\beta}$.
2. La matrice di varianza-covarianza dello stimatore $\mathbf{b}$ è $V(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
3. Nella classe degli stimatori corretti di β_j (per $j = 0,1,\ldots,k$) che sono funzioni lineari delle Y_i, gli stimatori dei minimi quadrati ordinari sono i più efficienti, cioè sono quelli che hanno minima varianza per qualsiasi valore dei parametri (***Teorema di Gauss-Markov***).

NOTA Si denoti con c_{ij} l'elemento della matrice $(\mathbf{X}'\mathbf{X})^{-1}$ posto all'intersezione della riga $h+1$ con la colonna $k+1$. La proprietà 2. implica che la varianza di b_j, per $j = 0,1,\ldots,k$, è $V(b_j) = \sigma^2 c_{jj}$,

mentre la covarianza tra b_h e b_j, per $h, j = 0,1,\ldots,k$ con $h \neq j$, è $Cov(b_h, b_j) = \sigma^2 c_{hj}$. Il valore del parametro σ^2 in genere non è noto e dunque non è noto neppure il valore di $V(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ pertanto sarà necessario stimare il parametro σ^2 come verrà illustrato in seguito.

Considerando la riformulazione in forma matriciale del modello di regressione lineare semplice, si può vedere facilmente che gli elementi della matrice $V(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ portano agli stessi risultati dati nel paragrafo 16.6.
Consideriamo ora una stima del parametro σ^2 basata sui residui $\hat{e}_i = y_i - \hat{y}_i$.

Lo stimatore **corretto** della varianza dei residui è dato da:

$$s^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n - k - 1} \qquad (19.4.3)$$

Sostituendo σ^2 con s^2, possiamo stimare la varianza degli stimatori dei minimi quadrati ordinari. La radice quadrata della stima della varianza di b_j, $s(b_j)$ è detta **errore standard** di b_j. Questa quantità misura la dispersione media dello stimatore intorno al suo valore atteso ed è una quantità fondamentale per l'inferenza sui coefficienti di regressione.

ESEMPIO 19.4.3 – *Errori standard delle stime dei coefficienti di regressione*
Riprendendo il file di dati considerato nell'esempio 19.4.2, le seguenti tabelle riportano l'output ottenuto tramite il software Excel.

Statistica della regressione	
R multiplo	0,699
R al quadrato	0,489
R al quadrato corretto	0,487
Errore standard	0,285
Osservazioni	474

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	2	36,512	18,256	225,320	0,000
Residuo	471	38,162	0,081		
Totale	473	74,675			

Stima dei coefficienti

	Coefficiente	*Errore standard*	*t*	*p-value*
Costante	8,877	0,120	73,797	0,0000
X_1	0,096	0,005	21,043	0,0000
X_2	0,002	0,001	1,798	0,0728

Nella seconda tabella, in corrispondenza della cella con riquadro rosso, è riportata la stima di σ^2 pari a 0,081; La sua radice quadrata, pari a 0,285, è riportata nel riquadro rosso della prima tabella e rappresenta una stima della distanza media tra il valore osservato e la stima del suo valore atteso. I valori degli errori standard dei coefficienti di regressione sono riportati nell'ultima tabella.

19.5 La decomposizione della varianza totale e il coefficiente di determinazione multiplo

Come abbiamo già visto per il modello di regressione lineare semplice, anche in questo caso vale la proprietà della **decomposizione della varianza totale**, dove la **somma totale dei quadrati** (SQT) si può scomporre nella **somma dei quadrati della regressione** (SQR) e nella **somma dei quadrati degli errori** (SQE):

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{e}_i^2$$

A partire da tale relazione si può costruire, come già visto nel paragrafo 16.5, una misura di bontà di adattamento nota come **coefficiente di determinazione multiplo** e dato da:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} \qquad (19.5.1)$$

Questo indice è un'estensione del coefficiente di determinazione al caso di due o più variabili esplicative e varia tra 0 e 1, indicando la proporzione di variabilità di Y spiegata dalle variabili esplicative attraverso il modello di regressione. Assume il valore minimo quando $SQR = 0$, vale a dire in assenza di relazione statistica di tipo lineare tra le osservazioni; vale 1 quando $SQR = SQT$, ossia nel caso di perfetta dipendenza lineare.
Quando alle variabili esplicative del modello di regressione si aggiunge una nuova variabile, la somma dei quadrati degli errori (SQE) non aumenta e normalmente i valori stimati $\hat{y}_i$ risultano essere più vicini ai valori osservati y_i. Pertanto, il valore di R^2 con l'introduzione di una nuova variabile esplicativa nel modello non può mai decrescere. Ne consegue che un modello che differisce da un altro solamente per possedere una variabile in più rispetto all'altro mostrerà sempre un valore maggiore o uguale di R^2. Per poter confrontare la capacità d'adattamento di diversi modelli, neutralizzando l'effetto dovuto al diverso numero di variabili esplicative, possiamo prendere in considerazione il seguente indice.

Il **coefficiente di determinazione multiplo corretto** è dato da:

$$R_c^2 = 1 - \frac{SQE/(n-k-1)}{SQT/(n-1)} \qquad (19.5.2)$$

Questo indice tiene conto del numero di variabili esplicative incluse nel modello. Infatti, SQE è diviso per $n-k-1$ e quindi, aggiungendo una nuova variabile nel modello, k aumenta di un'unità e R_c^2 potrebbe aumentare o diminuire a seconda di quanto si è ridotto l'SQE.
Infine, il **coefficiente di correlazione multiplo** è dato dalla radice quadrata del coefficiente di determinazione multiplo, $R = \sqrt{R^2}$, e misura la correlazione lineare tra i valori osservati y_i e i corrispondenti valori stimati $\hat{y}_i$. Si noti che questo indice, a differenza del coefficiente di correlazione lineare, può assumere solo valori non negativi.

ESEMPIO 19.5.1 – *Bontà di adattamento del modello di regressione lineare multipla*
Riprendendo l'esempio 19.4.3, la seconda tabella riporta la decomposizione della varianza totale, dove $SQT = 74{,}675$, $SQR = 36{,}512$ e $SQE = 38{,}162$. La prima tabella invece riporta il valore di $R^2 = SQR/SQT = 36{,}512/74{,}675 = 0{,}489$ che risulta non troppo elevato denotando un adattamento moderato.
Il valore del coefficiente di determinazione multiplo corretto è

$$R_c^2 = 1 - (38{,}162/471)/(74{,}675/473) = 0{,}487 .$$

Il coefficiente di correlazione multiplo è $R = \sqrt{0{,}489} = 0{,}699$.

19.6 Inferenza sui parametri del modello di regressione

Come abbiamo già visto, sotto le prime tre assunzioni, gli stimatori dei minimi quadrati ordinari o di massima verosimiglianza, definiti dagli elementi di $\mathbf{b}$, sono corretti, ossia $E(\mathbf{b}) = \boldsymbol{\beta}$, e possiedono matrice di varianza-covarianza pari a $V(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Tuttavia, per poter fare inferenza sui coefficienti di regressione si deve conoscere la distribuzione di ognuno dei $k+1$ stimatori. Per derivare questa distribuzione è richiesta l'assunzione di normalità dei termini di errore (assunzione 4).

Sotto l'assunzione 4, lo stimatore dei minimi quadrati $\mathbf{b}$ di $\boldsymbol{\beta}$ ha distribuzione Normale multivariata con vettore dei valori attesi pari a $E(\mathbf{b}) = \boldsymbol{\beta}$ e matrice di varianza-covarianza $V(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, ossia

$$\mathbf{b} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right). \qquad (19.6.1)$$

Pertanto, ogni elemento di $\mathbf{b}$ ha distribuzione Normale, $b_j \sim N(\beta_j, \sigma^2 c_{jj})$ per $j = 0, 1, \ldots, k$. Quindi, la statistica

$$\frac{b_j - \beta_j}{\sigma\sqrt{c_{jj}}} \sim N(0,1).$$

Tuttavia, solo raramente si è a conoscenza del valore del parametro σ^2 e il più delle volte viene stimato attraverso lo stimatore s^2 dato dalla 19.4.3; quindi sostituendo tale valore si ottiene la statistica

$$\frac{b_j - \beta_j}{s(b_j)} \sim t_{n-k-1} \qquad (19.6.2)$$

che si distribuisce come una variabile casuale $t - Student$ con $n - k - 1$ gradi di libertà. Il denominatore della 19.6.2 è dato dall'errore standard dello stimatore b_j.
Dalla 19.6.2 possiamo ricavare gli intervalli di confidenza per i parametri del modello e le statistiche test da utilizzare nella verifica d'ipotesi.

Intervalli di confidenza

Gli intervalli di confidenza per i parametri di regressione $\boldsymbol{\beta}$ a un livello di confidenza $1 - \alpha$ sono dati da:

$$b_j \pm t_{\alpha/2} s(b_j) \qquad (19.6.3)$$

dove $t_{\alpha/2}$ indica quel valore per cui la probabilità di osservare valori della $t - Student$ con $n - k - 1$ gradi di libertà superiori o uguali a $t_{\alpha/2}$ è pari a $\alpha/2$.

ESEMPIO 19.6.1 – *Intervallo di confidenza per i parametri del modello*
Consideriamo il file di dati *Demograf* che contiene le determinazioni del *Numero di crimini*, della *Popolazione (in migliaia)*, della *% di popolazione con 65 anni e più*, della *% di popolazione con diploma superiore* e della *Forza lavoro (in migliaia)*, osservate su 141 aree metropolitane degli Stati Uniti nel 1977. Consideriamo il modello di regressione lineare che pone il *Numero di crimini* in funzione delle altre variabili. Le stime dei coefficienti di regressione, degli errori standard e dei limiti inferiori e superiori degli intervalli di confidenza al 95% di confidenza, sono riportati nella seguente tabella:

Stima dei coefficienti			*Intervallo di confidenza*	
	Coefficienti	*Errore standard*	*Inferiore 95%*	*Superiore 95%*
Costante	-29059,26	13902,84	-56552,96	-1565,57
Popolazione (migliaia)	76,44	17,61	41,62	111,26
% popolazione di 65 e più anni	168,18	653,44	-1124,03	1460,39
% con diploma superiore	388,49	208,30	-23,44	800,41
Forza lavoro (migliaia)	-22,20	38,40	-98,15	53,74

Verifica d'ipotesi

Dalla formula 19.6.2 possiamo anche ricavare le statistiche test per la verifica d'ipotesi sui coefficienti di regressione. Di norma si è interessati a verificare l'ipotesi nulla $H_0 : \beta_j = 0$ per qualche $j = 1, \ldots, k$ contro l'ipotesi alternativa $H_1 : \beta_j \neq 0$. Infatti, se H_0 non è rifiutata, si può affermare che X_j non è utile alla previsione della Y. Più in generale, il sistema d'ipotesi può essere formulato come $H_0 : \beta_j = \beta_{j0}$ contro $H_1 : \beta_j \neq \beta_{j0}$ dove β_{j0} è il valore dato al parametro sotto l'ipotesi nulla. Sotto l'ipotesi nulla si ha che la statistica test:

$$\frac{b_j - \beta_{j0}}{s(b_j)} \sim t_{n-k-1}$$

si distribuisce come una variabile casuale $t - Student$ con $n - k - 1$ gradi di libertà.
Se si vuole verificare il sistema d'ipotesi bidirezionale a un livello di significatività α, la regione di rifiuto sarà data dai valori della statistica test superiori, in valore assoluto, a $t_{\alpha/2}$. Perciò,

Regione di Rifiuto di H_0: $|t| \geq t_{\alpha/2}$

In corrispondenza del valore osservato della statistica test possiamo calcolare il *p-value* che, come sappiamo, è una misura del "grado di disaccordo" rispetto all'ipotesi nulla: quanto più è piccolo il *p-value*, tanto maggiore è l'evidenza contro l'ipotesi nulla.

NOTA Se per alcune variabili i coefficienti non risultano significativamente diversi da zero, non è corretto eliminare dal modello più di una variabile alla volta. Infatti, ad esempio, l'eliminazione

contemporanea di due variabili dal modello corrisponde ad accettare l'ipotesi nulla $H_0 : \beta_i = \beta_j = 0$ ma a un livello di α superiore a quello utilizzato per i due singoli test. Come si vedrà in seguito una tale verifica può essere svolta attraverso il test F.

ESEMPIO 19.6.2 – *Verifica d'ipotesi per i parametri del modello*
Riprendendo i dati dell'esempio 19.6.1, nelle seguenti tabelle sono riportati i risultati della regressione. In particolare, nell'ultima tabella sono mostrati i valori della statistica test t e dei corrispondenti *p-value* (l'ipotesi nulla pone il valore del coefficiente di regressione pari a 0):

Statistica della regressione	
R multiplo	0,978
R al quadrato	0,956
R al quadrato aggiustato	0,954
Errore standard	18691,463
Osservazioni	141

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	4	1,026e+12	2,56695e+11	734,7357	0,000
Residuo	136	47514426254	349370781,3		
Totale	140	1,0743e+12			

Stima dei coefficienti

	Coefficienti	*Errore standard*	*Stat t*	*p-value*
Costante	-29059,26	13902,84	-2.090	0,038
Popolazione (migliaia)	76,44	17,61	4,341	0,000
% popolazione di 65 anni e più	168,18	653,44	0,257	0,797
% con diploma superiore	388,49	208,30	1,865	0,064
Forza lavoro (migliaia)	-22,20	38,40	-0,578	0,564

Per un livello di significatività pari a $\alpha = 0{,}10$ solo la *Costante* e le variabili *Popolazione* e *% di popolazione con diploma superiore* risultano avere un valore del coefficiente significativamente diverso da zero.

19.7 Il test F per selezionare il modello di regressione

Il test F è una procedura per verificare ipotesi riguardanti uno o più coefficienti di regressione. Essa è impiegata soprattutto per 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 e che pertanto possono essere escluse dal modello di regressione.
In generale, si supponga di voler verificare l'ipotesi nulla:

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

per qualche $h \leq k$, contro l'ipotesi alternativa

$$H_1 : \text{almeno un' uguaglianza in } H_0 \text{ non è vera}$$

L'ipotesi nulla afferma che almeno $k - h$ variabili, $X_{h+1}, X_{h+2}, \ldots, X_k$, non sono utili per spiegare la relazione lineare con la variabile dipendente. Sotto l'ipotesi nulla vale

pertanto il seguente modello:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_h x_{ih} + \varepsilon_i$$

Un tale modello è detto **modello ridotto**, in contrapposizione al **modello completo** basato su tutte le variabili esplicative:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

Il modello ridotto è **annidato** all'interno del modello completo, ossia il primo è un caso particolare del secondo, quando è vera l'ipotesi nulla.
Ovviamente l'adattamento ai dati osservati del modello completo, giacché considera più variabili esplicative, è sempre migliore di quello ridotto. Utilizzando la notazione introdotta nel paragrafo 17.3 riguardante l'Analisi della Varianza, indichiamo con SQE la somma dei quadrati degli errori del modello completo e con SQE_0 quella del modello ridotto.
Quindi si ha che $SQE \leq SQE_0$, e

- se la differenza $SQE_0 - SQE$ è grande ⇒ H_0 dovrebbe essere rifiutata poiché gli ultimi $k - h$ parametri aumentano considerevolmente l'adattamento del modello;
- se la differenza $SQE_0 - SQE$ è piccola ⇒ H_0 non dovrebbe essere rifiutata poiché l'adattamento del modello non aumenta in modo considerevole.

Per stabilire se la differenza $SQE_0 - SQE$ è grande o piccola, si può utilizzare la seguente statistica test:

$$F = \frac{(SQE_0 - SQE)/(k - h)}{SQE/(n - k - 1)} \qquad (19.7.1)$$

Sotto l'ipotesi nulla questa statistica si distribuisce come una variabile casuale $F - Fisher$ con $k - h$ e $n - k - 1$ gradi di libertà (si veda paragrafo 9.8.5). Si noti che il denominatore della 19.7.1 è la stima del parametro σ^2 ottenuta dal modello completo. Quanto più il valore di F è grande tanto più l'ipotesi H_0 sarà rifiutata a favore dell'ipotesi H_1. Stabilito il livello di significatività α del test la regione di rifiuto è data da

Regione di Rifiuto di H_0: $F \geq F_\alpha$

dove F_α indica quel valore per cui la probabilità di osservare valori della $F - Fisher$ con $k - h$ e $n - k - 1$ gradi di libertà superiori o uguali a F_α è pari ad α.
Quando $F < F_\alpha$, l'ipotesi nulla non può essere rifiutata e quindi le $k - h$ variabili esplicative possono essere escluse dal modello di regressione.

NOTA Come già notato nel paragrafo 17.3, quando $h = k - 1$ il test F verifica l'ipotesi nulla $H_0 : \beta_j = 0$ e porta alle stesse conclusioni del test t descritto nel paragrafo 19.6. Quando $h = 0$, l'ipotesi nulla diventa $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, e corrisponde a verificare la validità del modello di

regressione nel suo complesso. In questo caso si può dimostrare che la statistica test F è una funzione del coefficiente di determinazione multiplo: $F = \dfrac{R^2/k}{(1-R^2)/(n-k-1)}$.

ESEMPIO 19.7.1 – *Test F per il modello di regressione*

Riprendendo l'esempio 19.6.2, si era visto che i coefficienti di regressione trovati per le variabili esplicative *% di popolazioni di 65 anni e più* e *Forza lavoro* non erano significativamente diversi da zero (i valori del *p-value* sono rispettivamente pari a 0,797 e 0,564). Per decidere se entrambe le variabili possono essere rimosse dal modello si deve utilizzare il test F. Bisogna, pertanto, comparare il modello completo, quello mostrato nella tabella dell'esempio 19.6.2, con il modello ridotto in cui sono state eliminate le suddette variabili esplicative. Di seguito sono presentate le tabelle riassuntive ottenute con Excel della stima del modello ridotto.

Statistica della regressione	
R multiplo	0,978
R al quadrato	0,956
R al quadrato aggiustato	0,955
Errore standard	18588,248
Osservazioni	141

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	2	1,027e+12	5,11307e+11	1485,5931	0,000
Residuo	138	47682166985	345522949,2		
Totale	140	1,0743e+12			

Stima dei coefficienti

	Coefficienti	*Errore standard*	*Stat t*	*p-value*
Costante	-25345,93	10955,14	-2,314	0,022
Popolazione (migliaia)	66,32	1,22	54,216	0,000
% con diploma superiore	349,80	199,20	1,756	0,081

La somma dei quadrati degli errori del modello completo è $SQE = 47514426254$ mentre quella del modello ridotto è $SQE_0 = 47682166985$. Quindi

$$F = \frac{(47682166985 - 47514426254)/2}{47514426254/136} = 0{,}24 .$$

Compariamo il valore della statistica test con il valore soglia per un $\alpha = 0{,}05$ ottenuto da una $F - Fisher$ con 2 e 136 gradi di libertà e pari a $F_{0,05} = 3{,}063$. Poiché 0,25<3,063 non può essere rifiutata l'ipotesi nulla. Ciò significa che per analizzare il *Numero di Crimini* il modello ridotto è preferibile al modello completo. Ciò è confermato anche dal fatto che i due modelli hanno praticamente lo stesso valore del coefficiente di determinazione multiplo e che il valore di quello corretto nel modello ridotto è, anche se di poco, superiore a quello del modello completo.
Dalla tabella dell'Analisi della Varianza si può verificare il caso di $h = 0$ attraverso il valore della statistica test F (pari a 1485,5931) e il corrispondente valore del *p-value* (pari a 0,000) che in questo caso conducono a rifiutare l'ipotesi nulla che il modello nel suo complesso non è utile a spiegare la variabile dipendente.

19.8 Inferenza per la risposta media e per la previsione

Considerando la **funzione di regressione stimata** $\hat{Y} = b_0 + b_1 x_1 + \cdots + b_k x_k$, per una data combinazione di valori delle variabili esplicative, $x_{i1}, x_{i2}, \ldots, x_{ik}$, la stima del valore atteso della variabile dipendente è $\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}$, che può essere riscritta in forma matriciale come $\hat{y}_i = \mathbf{x}_i' \mathbf{b}$, con $\mathbf{x}_i' = \begin{pmatrix} 1 & x_{i1} & \cdots & x_{ik} \end{pmatrix}$.
La stima di $E(Y_i | x_{i1}, \ldots, x_{ik})$, ossia $\hat{y}_i$, varierà a seconda del campione estratto, generando la variabile casuale *stimatore della risposta media* $\hat{Y}_i$.

Proprietà dello stimatore della risposta media

1. $\hat{Y}_i$ è uno stimatore corretto di $E(Y_i | x_{i1}, \ldots, x_{ik})$.
2. La varianza di $\hat{Y}_i$ è data da $V(\hat{Y}_i) = \sigma^2 \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}$

Assumendo l'ipotesi di normalità dei termini di errore, la $\hat{Y}_i$ si distribuisce normalmente e quindi:

$$\frac{\hat{Y}_i - E(Y_i | x_{i1}, \ldots, x_{ik})}{\sqrt{\sigma^2 \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}}} \sim N(0,1)$$

Sostituendo la varianza del termine di errore, σ^2, con lo stimatore s^2, si ottiene la variabile casuale

$$t = \frac{\hat{Y}_i - E(Y_i | x_{i1}, \ldots, x_{ik})}{\sqrt{s^2 \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}}} \sim t_{n-k-1}$$

che si distribuisce come una v.c. $t - Student$ con $n - k - 1$ gradi di libertà.
Indicando con $s(\hat{Y}_i) = \sqrt{s^2 \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}}$ l'**errore standard** di $\hat{Y}_i$, si ottiene l'espressione dell'intervallo di confidenza per la risposta media a un livello di confidenza $1 - \alpha$

$$\hat{Y}_i \pm t_{\alpha/2} s(\hat{Y}_i) \qquad (19.8.1)$$

Per la previsione del singolo valore della Y_i, in corrispondenza di una nuova osservazione che presenta valori delle variabili indipendenti uguali a quelli di $\mathbf{x_i}$, si utilizza lo stesso stimatore utilizzato per stimare la risposta media. Tuttavia, come abbiamo già visto nel paragrafo 17.4, lo stimatore per i singoli valori Y_i ha un errore standard più grande dato da $s(Y_i - \hat{Y}_i) = \sqrt{s^2\left(1 + \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}\right)}$.

Perciò, l'intervallo di confidenza per la previsione di un singolo valore Y_i a un livello di confidenza $1 - \alpha$ è dato da:

$$\hat{Y}_i \pm t_{\alpha/2} s\left(Y_i - \hat{Y}_i\right) \qquad (19.8.2)$$

ESEMPIO 19.8.1 – *Intervallo di confidenza per la risposta media e la previsione*
Riprendendo l'esempio 19.7.1, in cui il *Numero di crimini* viene messo in relazione con la *Popolazione* e la *% di persone con diploma superiore*, supponiamo di considerare un'area con una Popolazione di 2900 (migliaia) di individui di cui il 64% con diploma superiore. Sostituendo tali valori al modello stimato si ottiene la stima del *Numero di crimini* pari a 189356,6. Considerando l'intervallo di confidenza a livello del 95% si ottiene: per la risposta media l'intervallo [182770,4 ; 195942,7] mentre per il valore previsto l'intervallo più ampio [152016,5 ; 226696,6].

19.9 La multicollinearità

In molte situazioni, le variabili esplicative possono essere tra loro molto correlate e in questo caso diremo che ci troviamo in una situazione di **multicollinearità**. Ciò accade spesso nelle indagini di tipo socio-economico dove la maggior parte delle variabili considerate non possono essere tenute completamente sotto controllo. L'effetto principale dovuto alla multicollinearità è quello di aumentare considerevolmente la varianza degli stimatori dei minimi quadrati dei coefficienti di regressione. Consideriamo il caso di due variabili esplicative, X_1 e X_2, e per semplicità supponiamo che siano standardizzate (ossia con media nulla e varianza unitaria). In questo caso si può dimostrare che

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{1-\rho_{12}^2}\begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix},$$

dove ρ_{12} è il coefficiente di correlazione tra X_1 e X_2. Pertanto, la varianza degli stimatori dei minimi quadrati di β_1 e β_2 è $V(b_1) = V(b_2) = \sigma^2 \frac{1}{1-\rho_{12}^2}$ ed è chiaro che tanto maggiore è il valore di ρ_{12} tanto più grande è il valore della varianza degli stimatori. Quando $\rho_{12} = 0$, ossia le due variabili sono incorrelate, le varianze assumono il valore minimo. Più in generale, la varianza degli stimatori dei minimi quadrati b_j aumenta al crescere della dipendenza lineare della variabile X_j dalle altre variabili esplicative. L'aumento della varianza dovuto alla multicollinearità ha degli effetti negativi sull'inferenza dei coefficienti di regressione. In particolare, come si può vedere dalla 19.6.3, l'aumento di $s(b_j)$ porta all'aumento dell'ampiezza dell'intervallo di confidenza. Inoltre, dalla 19.6.2, l'aumento dell'errore standard fa diminuire il valore assoluto della statistica test portando più facilmente a non rifiutare l'ipotesi nulla anche se questa non è vera, ossia a commettere con maggiore probabilità un errore del secondo tipo.
Un indice utilizzato per misurare il livello di multicollinearità della variabile X_j con le altre variabili esplicative è il **Variance inflaction factor** dato da

$$VIF_j = \frac{1}{1-R_j^2} \qquad (19.9.1)$$

dove R_j^2 è il coefficiente di determinazione multiplo del modello di regressione nel quale la variabile X_j dipende dalle altre $k-1$ variabili esplicative. Il valore minimo del VIF_j è 1 e indica che la variabile X_j è incorrelata dalle altre. In genere, un valore superiore a 2 indica già un livello sufficientemente alto di multicollinearità.
La presenza di elevata multicollinearità comporta il cambiamento nei valori delle stime dei coefficienti di regressione in conseguenza a lievi modificazioni dei valori osservati, a eliminazione o aggiunta di qualche variabile esplicativa, all'aggiunta di nuove osservazioni.
Tuttavia, la multicollinearità non altera la bontà di adattamento del modello e la sua capacità previsiva rispetto alla variabile risposta.

ESEMPIO 19.9.1 – *Verifica delle multicollinearità*
Riprendendo il dataset *Demograf* dell'esempio 19.6.1, si ottengono in corrispondenza del modello stimato i seguenti valori del VIF :

Variabili	VIF
Popolazione (migliaia)	205,957
% con diploma superiore	1,087
% popolazione di 65 e più anni	1,089
Forza lavoro (migliaia)	206,150

Come si può vedere la variabile *Popolazione* e *Forza Lavoro* presentano un valore molto elevato del VIF indicando presenza di multicollinearità. Si può verificare ad esempio che $R^2_{ForzaLavoro} = 0{,}995$.
Nel modello ridotto riportato nell'esempio 17.7.1 non vi è multicollinearità come si può osservare dai valori riportati nella seguente tabella.

Variabili	VIF
Popolazione (migliaia)	1,005
% con diploma superiore	1,005

Appendice

A.19.1 Stima dei coefficienti di regressione

Il problema consiste nel minimizzare $G(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2$ rispetto a $\boldsymbol{\beta}$. In forma matriciale possiamo scrivere $G(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ ed espandendo l'espressione (ricordando che $(\mathbf{X}\boldsymbol{\beta})' = \boldsymbol{\beta}'\mathbf{X}'$) si trova $G(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$.

Notando che $\mathbf{y}'\mathbf{X}\boldsymbol{\beta}$ ha dimensione $(1 \times n)(n \times k)(k \times 1) = 1 \times 1$ darà lo stesso valore della sua trasposta $\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$. Pertanto si trova che $G(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$. Differenziando la funzione rispetto al vettore di parametri e uguagliandola a zero si ottiene:

$$\frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

da cui, ponendo $\mathbf{b}$ al posto di $\boldsymbol{\beta}$ e dividendo per 2, si ottiene il seguente **sistema normale**:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

Per ottenere la stima dei coefficienti di regressione dal sistema normale si deve premoltiplicare entrambi i membri dell'equazione per la matrice inversa di $\mathbf{X}'\mathbf{X}$ (assumendo che esista) $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ e, poiché $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$ e $\mathbf{I}\mathbf{b} = \mathbf{b}$, si ottiene

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

A.19.2 Correttezza degli stimatori dei minimi quadrati

Poichè $\mathbf{b}$ si può riscrivere come funzione lineare del tipo $\mathbf{b} = \mathbf{A}\mathbf{Y}$, con $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ matrice costante, possiamo facilmente dimostrare che $\mathbf{b}$ è uno stimatore corretto di $\boldsymbol{\beta}$.

Infatti, dato che $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$, si ha $E(\mathbf{b}) = E(\mathbf{A}\mathbf{Y}) = \mathbf{A}E(\mathbf{Y}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{I}\boldsymbol{\beta} = \boldsymbol{\beta}$.

A.19.3 Varianza e covarianza degli stimatori dei minimi quadrati

Dalle proprietà delle matrici sappiamo che dato un vettore $\mathbf{W} = \mathbf{A}\mathbf{Y}$, con $\mathbf{A}$ matrice costante, allora $V(\mathbf{W}) = V(\mathbf{A}\mathbf{Y}) = \mathbf{A}V(\mathbf{Y})\mathbf{A}'$. Pertanto $V(\mathbf{b}) = V(\mathbf{A}\mathbf{Y}) = \mathbf{A}V(\mathbf{Y})\mathbf{A}'$. Ricordando che $V(\mathbf{Y}) = \sigma^2\mathbf{I}$ e $\mathbf{A}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ si ha $V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{I} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

A.19.4 Proprietà dello stimatore della risposta media

Lo stimatore è corretto, infatti

$$E(\hat{Y}_i | x_{i1}, \ldots, x_{ik}) = E(\mathbf{x}_i'\mathbf{b}) = \mathbf{x}_i'E(\mathbf{b}) = \mathbf{x}_i'\boldsymbol{\beta} = E(Y_i | x_{i1}, \ldots, x_{ik})$$

La varianza dello stimatore è data da

$$V(\hat{Y}_i) = V(\mathbf{x}_i'\mathbf{b}) = \mathbf{x}_i'V(\mathbf{b})\mathbf{x}_i = \sigma^2\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i.$$

VERIFICA

Indicare se le seguenti affermazioni sono vere o false. *(Le risposte esatte sono in fondo a questo documento)*	V	F
19.1 - $Y_i = \beta_0 + \beta_1 log(x_{i1}) + \beta_2 x_{i2} + \varepsilon_i$ è un modello di regressione lineare multipla.	☐	☐
19.2 - Aggiungendo al modello una variabile esplicativa il coefficiente di determinazione multiplo non può mai ridursi.	☐	☐
19.3 - Il coefficiente di determinazione multiplo corretto cresce sempre all'aumentare delle variabili esplicative incluse nel modello.	☐	☐
19.4 - Il coefficiente di correlazione multiplo può essere negativo.	☐	☐
19.5 - Nel modello di regressione si assume che le osservazioni della variabile risposta siano incorrelate.	☐	☐
19.6 - Gli stimatori di massima verosimiglianza per i coefficienti di regressione sono gli stessi dei minimi quadrati ordinari.	☐	☐
19.7 - La funzione di regressione descrive l'andamento del valore atteso della variabile dipendente al variare del valore delle variabili esplicative.	☐	☐
19.8 - Se dal test t risulta che due variabili esplicative hanno coefficienti di regressione non significativamente diversi da zero, allora possiamo direttamente eliminarle dal modello.	☐	☐
19.9 - Il modello completo differisce dal modello ridotto perché include una sola variabile esplicativa in più.	☐	☐
19.10 - Il coefficiente di determinazione multiplo indica la proporzione di variabilità totale dovuta all'errore.	☐	☐
19.11 - La variabilità degli stimatori dei minimi quadrati per i coefficienti di regressione dipende dalla matrice $(\mathbf{X}'\mathbf{X})^{-1}$.	☐	☐
19.12 - Più le variabili esplicative sono correlate fra loro e maggiore è la variabilità dello stimatore $\mathbf{b}$.	☐	☐
19.13 - In presenza di multicollinearità il modello di regressione peggiora la sua capacità di previsione.	☐	☐
19.14 - Lo stimatore $\mathbf{b}$ di $\boldsymbol{\beta}$ ha minima variabilità quando le variabili esplicative sono fra loro massimamente correlate.	☐	☐
19.15 - Il coefficiente di determinazione corretto tiene conto del numero di variabili esplicative incluse nel modello.	☐	☐
19.16 - L'intervallo di confidenza per la risposta media non è mai più grande di quello per il valore previsto.	☐	☐
19.17 - Se $R^2 = R_c^2 = R = 1$ allora la funzione di regressione si adatta perfettamente ai dati osservati.	☐	☐
19.18 - Per fare inferenza sui parametri del modello di regressione è necessaria l'assunzione di normalità del termine di errore.	☐	☐
19.19 - La normalità del termine di errore non implica la normalità degli stimatori dei minimi quadrati dei coefficienti di regressione.	☐	☐
19.20 - Se $V(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ allora i termini di errore sono tra loro incorrelati.	☐	☐

ESERCIZI *(le soluzioni sono disponibili sul sito WEB)*

Gli esercizi proposti prevedono l'utilizzazione di file di dati in formato JMP, SPSS o EXCEL.

19.1 Costruire la matrice $\mathbf{X}$ e il vettore $\boldsymbol{\beta}$ per ognuno dei seguenti modelli di regressione multipla (si assume $i = 1, \ldots, 4$):

a. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

b. $\log Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

c. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i2}^2 + \varepsilon_i$

d. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$

19.2 Per ognuno dei seguenti modelli di regressione, dire se si tratta di un modello di regressione lineare multipla. Se non lo è, dire se, mediante un'opportuna trasformazione, è possibile esprimerlo nella forma dell'Assunzione 1.

a. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 \log x_{i2} + \beta_3 x_{i2}^2 + \varepsilon_i$

b. $Y_i = \varepsilon_i \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2\right)$

c. $Y_i = \beta_0 + \log\left(\beta_1 x_{i1}\right) + \beta_2 x_{i2} + \varepsilon_i$

d. $Y_i = \left[1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \varepsilon_i\right)\right]^{-1}$

19.3 Si consideri il file-dati "Cellulare" (.jmp, .sav o .xls) contenente la media mensile in minuti di utilizzo del cellulare (*Minuti*), il costo medio mensile delle telefonate (*Bolletta*), la percentuale per uso ufficio (*Lavoro*) e il reddito famigliare (*Reddito*) di 250 individui. Stimando il modello di regressione lineare multipla che fa dipendere la variabile *media mensile in minuti d'utilizzo del cellulare* dalle restanti variabili, si ottengono le seguenti tabelle di output:

Statistica della regressione	
R multiplo	0,540
R al quadrato	0,292
R al quadrato corretto	0,283
Errore standard	39,424
Osservazioni	250

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	3	157695,7	52565,2	33,821	2,44903E-18
Residuo	246	382340,71	1554,23		
Totale	249	540036,41			

	Coefficienti	*Errore standard*	*Stat t*	*p-value*	*Inferiore 95%*	*Superiore 95%*	VIF
Intercetta	29,625	15,503	1,911	0,057	-0,910	60,161	
BOLLETTA	0,885	0,147	6,016	0,000	0,595	1,175	1,360
LAVORO	0,536	0,323	1,662	0,098	-0,099	1,172	1,372
REDDITO	0,956	0,233	4,112	0,000	0,498	1,414	1,071

a. Aumentando di un euro il costo medio della bolletta di quanto aumenta la media mensile di utilizzo del cellulare (tenendo costante il valore delle altre variabili)?

b. Considerando un livello di significatività pari a $\alpha = 0{,}10$, indicare quali sono le variabili esplicative che presentano un coefficiente di regressione significativamente diverso da zero.
c. Considerando un livello di confidenza pari a $1 - \alpha = 0{,}95$, il coefficiente di regressione della variabile *Bolletta* può essere pari a 1,2?
d. La bontà di adattamento del modello di regressione lineare è molto elevata?
e. Si può rifiutare l'ipotesi nulla che i coefficienti di regressioni sono tutti uguali a zero per un $\alpha = 0{,}05$?
f. Possiamo dire che c'è multicollinearità tra le variabili esplicative?

19.4 Si consideri il file di dati "Auto" (.jmp, .sav o .xls) dell'esercizio 18.2. Stimando il modello di regressione lineare multipla che fa dipendere la variabile *Consumo dell'automobile* dalla *Cilindrata* (MOTORE), dai *Cavalli* (CV), dal *Peso* e dall'*Accelerazione* si ottengono le seguenti tabelle di output:

Statistica della regressione	
R multiplo	0,839
R al quadrato	0,704
R al quadrato corretto	0,701
Errore standard	4,254
Osservazioni	391

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	4	16592,1	4148,03	229,22	1,49E-100
Residuo	386	6985,3	18,10		
Totale	390	23577,4			

	Coefficienti	*Errore standard*	*Stat t*	*p-value*	*Inferiore 95%*	*Superiore 95%*	VIF
Intercetta	40,847	2,297	17,784	0,000	36,331	45,363	
MOTORE	-0,007	0,007	-1,040	0,299	-0,020	0,006	10,875
CV	-0,052	0,017	-3,065	0,002	-0,085	-0,018	9,139
PESO	-0,015	0,002	-6,071	0,000	-0,020	-0,010	10,181
ACCEL	-0,121	0,127	-0,953	0,341	-0,371	0,129	2,686

a. Considerando un livello di significatività pari a $\alpha = 0{,}05$, indicare quali sono le variabili esplicative che presentano un coefficiente di regressione significativamente diverso da zero.
b. Considerando un livello di confidenza pari a $1 - \alpha = 0{,}95$, il coefficiente di regressione della variabile *Peso* può essere di segno positivo?
c. La bontà di adattamento del modello di regressione lineare è sufficientemente elevata?
d. Si può accettare l'ipotesi nulla che i coefficienti di regressioni sono tutti uguali a zero per un $\alpha = 0{,}01$?
e. Possiamo dire che c'è multicollinearità tra le variabili esplicative?

19.5 Si consideri il file di dati "Auto" (.jmp, .sav o .xls). Stimando il modello di regressione lineare multipla che fa dipendere la variabile *Consumo dell'automobile* dai *Cavalli* (CV) e dal *Peso* si ottiene la seguente tabella di output:

Statistica della regressione	
R multiplo	0,838
R al quadrato	0,702
R al quadrato corretto	0,701
Errore standard	4,252
Osservazioni	391

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	2	16561,6	8280,8	457,96	7,50E-103
Residuo	388	7015,8	18,1		
Totale	390	23577,4			

a. Sulla base del valore del coefficiente di determinazione multiplo e di quello corretto si può affermare che il modello suddetto si adatta meno bene di quello dell'esercizio 19.4 che include tutte le variabili esplicative?
b. Sulla base del test F possiamo accettare il modello ridotto (il valore della $F_{0,05} = 3,019$)?

19.6 Si consideri il file di dati "PIL" (.jmp, .sav o .xls) dell'esempio 19.1.1. Stimando il modello di regressione lineare multipla che fa dipendere la variabile *Spesa annuale per consumi* dalla *Popolazione* e dal *Tasso di disoccupazione*, osservati negli Stati Uniti tra il 1959 e il 1999, si ottengono le seguenti tabelle di output:

Statistica della regressione	
R multiplo	0,968
R al quadrato	0,937
R al quadrato corretto	0,934
Errore standard	468,891
Osservazioni	41

	Coefficienti	*Errore standard*	*Stat t*	*p-value*	*Inferiore 95%*	*Superiore 95%*
Intercetta	-11478,921	638,102	-17,989	0,000	-12770,690	-10187,151
Popolazione (migliaia)	0,065	0,003	23,786	0,000	0,059	0,070
Tasso % di disoccupazione	-161,084	51,501	-3,128	0,003	-265,343	-56,825

a. Dalla stima del coefficiente di regressione si può dire che la *Spesa per consumi* è legata inversamente al *Tasso di disoccupazione*?
b. Le due variabili esplicative sono significative a un livello $\alpha = 0,01$?
c. L'adattamento del modello ai dati si può ritenere molto elevato?
d. Sapendo che nel 1990 la *Popolazione* è di 249973 migliaia e il *Tasso % di disoccupazione* è di 5,6, qual è il valore atteso della *Spesa annuale per consumi* prevista dal modello?
e. Volendo verificare l'ipotesi nulla $H_0 : \beta_1 = \beta_2 = 0$, quanto vale la statistica test F ?
f. Sapendo che il *p-value* della statistica test F calcolata nel precedente punto è pari a 0,000 , si può rifiutare l'ipotesi nulla? Cosa possiamo concludere?
g. Considerando i valori del VIF per la *Popolazione* e il *Tasso di disoccupazione*, pari rispettivamente a 1,038 e 1,038, cosa possiamo concludere circa la multicollinearità delle variabili?