Regressione Multipla e Regressione Logistica: concetti introduttivi ed esempi

I Edizione

– ottobre 2016 –

Vincenzo Paolo Senese
vincenzopaolo.senese@unina2.it

Indice

Note preliminari alla I edizione

Questo materiale è stato preparato per gli studenti del Corso di Laurea Magistrale in "Psicologia Clinica", del Corso di Laurea Specialistica in "Psicologia clinica e dello sviluppo" e del Corso di Laurea Magistrale in "Psicologia dei processi cognitivi" della Seconda Università degli studi di Napoli (SUN) nell'ambito dei corsi di Metodi e Tecniche della ricerca in Psicologia Clinica, Analisi dei dati, e di Metodi e tecniche di analisi dei dati. L'obiettivo del presente testo è fornire un supporto utile alla preparazione dell'esame da utilizzare come materiale integrativo al testo di base.

Napoli, ottobre 2016

Il presente materiale è scaricabile gratuitamente dal sito del Dipartimento di Psicologia della Seconda Università di Napoli (www.psicologia.unina2.it) alla pagina relativa all'insegnamento di Metodi e Tecniche della ricerca in Psicologia Clinica del Corso di Laurea Magistrale in Psicologia Clinica (http://psiclab.altervista.org/MetTecPsicClinica2017/2016_2017.html) [link].

Per informazioni
Senese Vincenzo Paolo
Dipartimento di Psicologia, SUN
Viale Ellittico, 31
81100 – Caserta – Italy
Stampato in Italia

1

Regressione semplice e multipla

Introduzione

Il modello di regressione lineare consente di analizzare la relazione causale (ipotesi) tra una variabile dipendente quantitativa (misurata su scala almeno a intervalli) e una o più variabili indipendenti quantitative.

Dal momento che non è possibile indagare nella popolazione la presunta relazione tra le variabili considerate, per la verifica delle ipotesi si procede estraendo un campione rappresentativo della popolazione e descrivendo su questo la relazione tra le variabili considerate. Successivamente, mediante la statistica inferenziale, si verifica se la relazione descritta al livello campionario può essere generalizzata alla popolazione di riferimento.

1. La Regressione lineare Semplice e Multipla

Nell'analisi della `regressione lineare semplice` è possibile verificare se due variabili sono legate da una relazione di tipo lineare e verificare la forza della relazione. In termini formali, la relazione lineare tra due variabili può essere descritta dall'equazione della retta:

$$Y = \alpha + \beta X \qquad [1]$$

dove: `Y` è la variabile che deve essere prevista (variabile dipendente, criterio, risposta o variabile endogena); `X` è la variabile i cui valori vengono utilizzati per prevedere `Y` (variabile indipendente, predittore o variabile esogena); mentre α e β sono i parametri della popolazione che indicizzano la relazione tra le variabili considerate e che devono essere stimati.

In particolare, il parametro α viene detto intercetta (*intercept*) o costante e rappresenta il valore previsto in `Y` quando la variabile `X` è `0`. Il parametro β chiamato coefficiente di regressione o *slope*, rappresenta il cambiamento in `Y` per ogni variazione unitaria della `X` o l'inclinazione della retta che rappresenta meglio la relazione tra `X` e `Y`. Per tale ragione il valore di tale parametro dipende dall'unità di misura delle variabili.

Nella `regressione lineare multipla` ci sono molte variabili esogene, molti predittori e una variabile criterio. Se k denota il numero di

variabili indipendenti allora l'equazione che descrive la relazione tra le variabili indipendenti e la variabile dipendente diventa:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \qquad [2];$$

dove β_1, β_2, ..., β_k sono i coefficienti di regressione parziali e riflettono il fatto che ognuno dei predittori X_1, X_2, …, X_k considerati fornisce una spiegazione parziale (o predizione) della variabile endogena Y. Ovvero, il modello assume che ciascun predittore agisca in modo indipendente e lineare sulla variabile dipendente, e che le variazioni della variabile dipendente siano il risultato della somma dei singoli effetti indipendenti. Per questa ragione la Y viene simboleggiata come $\hat{Y}$.

È importante sottolineare che per una maggiore correttezza, sia nella regressione semplice sia nella regressione multipla, l'equazione andrebbe scritta includendo il termine d'errore (ε) relativo alla previsione della variabile dipendente. Pertanto la formula corretta diventa:

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \qquad [3].$$

In questo modo, infatti, si riconduce il modello alla sua natura probabilistica. Questo implica che il valore di Y predetto dal modello sia considerato un valore medio di riferimento attorno al quale si distribuiscono i valori osservati in funzione della distribuzione teorica di riferimento (distribuzione dell'errore o casuale) associata alla Y. Nella regressione lineare multipla si assume che la distribuzione teorica di riferimento della Y sia la distribuzione normale.

1.1. La stima dei parametri

Nell'analisi della regressione multipla il primo passo prevede la valutazione o stima dei parametri. Come precedentemente sottolineato, nella pratica, i parametri della popolazione non sono sempre noti. In tali casi i valori sono stimati considerando un numero finito di osservazioni: le osservazioni campionarie. Alla base di questo passaggio vi è l'assunzione che il campione corrisponda a una sottoparte rappresentativa della popolazione. Ovvero che nel campione siano rappresentate tutte le caratteristiche della popolazione, e che i fenomeni al livello campionario agiscano in maniera omologa a quanto

avviene nella popolazione. Le tecniche di campionamento servono a garantire che i campioni siano rappresentativi (De Carlo e Robusto, 1996).

Per distinguere la regressione campionaria da quella della popolazione il modello di regressione viene scritto in questo modo:

$$\hat{Y}_j = a + b_1 X_{1j} + b_2 X_{2j} + b_3 X_{3j} + ... + b_k X_{kj} + e_j \qquad [4],$$

dove le lettere latine (*a*, *b*, *e*) indicano i parametri del modello stimati a partire dal campione (*N*), e *j* rappresenta il singolo valore predetto (*j* = 1, 2, …, *N*).

Per la stima dei parametri *a* e b_i (*i* = 1, 2, …, *k*) il metodo più frequentemente impiegato è il principio dei minimi quadrati (*Ordinary Least Square* [OLS]; si veda Agresti e Finlay, 1997; Bohrnstedt e Knoke, 1994).

Tale metodo si pone come obiettivo di stimare i parametri *a* e b_i in modo tale da ridurre al minimo l'errore di misura: la distanza al quadrato tra i valori predetti in base al modello ($\hat{Y}_j$) e i valori osservati (Y_j). In termini matematici, l'OLS tende a minimizzare la seguente sommatoria:

$$\sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 \qquad [5].$$

Ovvero la sommatoria degli scarti dalla media al quadrato (SQ). Per questa ragione questo metodo viene definito il metodo dei "minimi quadrati".

Nella regressione semplice le formule per il calcolo dei parametri (secondo il metodo OLS) sono le seguenti:

$$b_i = \frac{\sum_{j=1}^{n} (X_{ij} - \overline{X}_i)(Y_j - \overline{Y})}{\sum_{j=1}^{n} (X_{ij} - \overline{X}_i)^2} \qquad [6];$$

$$a = \overline{Y} - b_i \overline{X} \qquad [7].$$

Nella regressione multipla le formule per il calcolo dei parametri richiedono l'algebra matriciale.

Poiché si tratta di stime campionare dei parametri è necessario conoscere l'effetto dell'errore sulla stima. Per fare ciò è necessario calcolare l'errore standard (s_i) del coefficiente stimato:

$$s_{b_i} = \sqrt{\frac{\sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2}{\sum_{j=1}^{n} (X_{ij} - \overline{X}_i)^2 (1 - R_i^2)(N - k - 1)}} \qquad \text{[6a].}$$

Dove: *N* è l'ampiezza campionaria; *k* è il numero di variabili indipendenti del modello; R^2_i è la correlazione multipla al quadrato delle variabili indipendenti sulla variabile indipendente considerata (*i*).

Della formula [6a] è utile notare che l'errore di stima di b_i (s_b) si riduce se: al numeratore è minore l'errore di stima di Y_j; al denominatore è maggiore la varianza di X_i; è minore la correlazione di X_i con le altre variabili indipendenti; è maggiore il numero delle osservazioni *N*, mentre se il numero di predittori aumenta e si approssima all'ampiezza campionaria, s_b aumenta notevolmente.

1.2. Le assunzioni dell'OLS

Perché la stima dei parametri possa essere considerara robusta, l'OLS presuppone che alcune assunzioni siano verificate. Le assunzioni sono le seguenti:

- misure: tutte le variabili indipendenti sono misurate su scala ad intervalli, a rapporti o dicotomica, la variabile dipendente è continua e misurata su scala ad intervalli o a rapporti. Tutte le variabili sono misurate senza errore;
- specificazioni: tutti i predittori rilevanti per la variabile dipendente sono stati inseriti nell'analisi, nessun predittore irrilevante è stato inserito, e la forma della relazione tra variabili indipendenti e dipendenti è lineare;
- valore atteso dell'errore: il valore atteso dell'errore ε è 0;
- omoschedasticità: la varianza del termine d'errore ε è la stessa (o è costante) per tutti i valori delle variabili indipendenti;
- normalità degli errori: gli errori della Y sono distribuiti normalmente (distribuzione normale) per ogni gruppo di valori delle variabili indipendenti;

- assenza di autocorrelazioni: non ci devono essere correlazioni tra i termini dell'errore prodotti da ciascun predittore (matematicamente $E(\varepsilon_i, \varepsilon_j) = 0$);
- assenza di correlazione tra errori e predittori: i termini d'errore devono essere non correlati con le variabili indipendenti, matematicamente $E(\varepsilon_j, X_j) = 0$;
- assenza di perfetta multicollinearità: nessuna delle variabili indipendenti deve essere una combinazione lineare perfetta delle altre variabili indipendenti (matematicamente, per ogni variabile *i* il valore di R^2_i deve essere minore di 1, dove R^2_i è la varianza della variabile indipendente X_i spiegata da tutti gli altri predittori nel modello $X_1, X_2, \ldots, X_k$).

1.3. La valutazione del *fit* del modello

Un altro aspetto utile alla valutazione del modello di regressione è la valutazione della bontà di adattamento del modello (*goodness-of-fit*). Vale a dire la capacità del modello di migliorare la previsione della variabile Y considerando come valore di riferimento il valore stimato mediante il modello di regressione (ipotesi alternativa H_1) piuttosto che il valore medio di Y (ipotesi nulla H_0).

Le statistiche maggiormente impiegate a tal scopo sono l'errore standard della stima e l'R^2.

L'errore standard della stima corrisponde all'errore standard dei residui, $\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2$, e rappresenta un indice che esprime l'ampiezza dell'errore di misura del modello considerato. Tale statistica viene stimata mediante la seguente formula:

$$s_e = \sqrt{\frac{\sum_{j=1}^{n}\left(Y_j - \hat{Y}_j\right)^2}{N-2}} \qquad [8a].$$

L'altra statistica utile a valutare il *fit* del modello è R^2 ed esprime la parte di varianza della variabile dipendente spiegata attraverso il modello. L'R^2 viene stimata con le seguenti formule, tra loro equivalenti:

$$R^2 = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \overline{Y})^2}{\sum_{j=1}^{n} (Y_j - \overline{Y})^2} \qquad [8b];$$

$$R^2 = 1 - \frac{\sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2}{\sum_{j=1}^{n} (Y_j - \overline{Y})^2} \qquad [8c].$$

Nella prima formula [8b] si mette in evidenza come l'R^2 rappresenti il rapporto tra la devianza spiegata dal modello e la devianza totale (la devianza di `Y` osservata), mentre nella seconda formula [8c] si pone in luce che l'R^2 rappresenta l'inverso del rapporto tra la devianza d'errore (o non spiegata dal modello) e la devianza totale. Ovvero ci fa capire che l'R^2 rappresenta una stima di quanto si riduce l'errore di previsione della variabile dipendente considerando le variabili nel modello.

L'R^2 varia sempre tra `0` e `1`. Può essere interpretato come la percentuale di varianza (`%`) della variabile dipendente spiegata dalle variabili indipendenti considerate nel modello. Oppure, considerando la seconda formulazione [8c], come la `%` di riduzione dell'errore nella previsione della variabile dipendente. Per l'utilizzo e l'interpretazione dell'R^2 due aspetti devono essere sottolineati. Primo, l'R^2 è dipendente dal campione. Modelli di regressione con le stesse variabili se sono applicati su campioni diversi possono avere identici parametri *b* ma R^2 differenti; questo è determinato dalla diversa varianza di `Y` nei campioni considerati. Secondo, l'R^2 è influenzato dal numero di predittori. A parità di campione per confrontare due modelli è necessario calcolare un valore corretto (*adjusted* R^2) (Wonnacott e Wonnacott, 1979) stimabile con la seguente formula:

$$R_{adjusted}{}^2 = \left(R^2 - \frac{k}{N-1} \right) \left(\frac{N-1}{N-k-1} \right) \qquad [9].$$

Per sottoporre a verifica l'ipotesi che prevede che la previsione della variabile dipendente `Y` migliora significativamente mediante il modello di regressione si pone a confronto la varianza spiegata dal modello con la varianza non spiegata (o varianza residua). Per la verifica delle ipotesi si utilizza il test del rapporto tra le varianza che si distribuisce come la variabile casuale *F* di Fischer.

Per il calcolo delle varianza si utilizza il teorema della scomposizione della devianza. Secondo tale teorema la devianza totale è data dalla somma della devianza d'errore e della devianza dell'effetto:

$$SQ_{tot} = SQ_{reg} + SQ_{err} \qquad [10a].$$

Nella regressione si assume che la somma dei quadrati totale (SQ_{tot} o devianza) è data da una componente di errore (SQ_{err}) e da una componente spiegata dalla regressione (SQ_{reg}). In termini formali, possiamo riscrivere la [10a] nel seguente modo:

$$\sum \left(Y_i - \overline{Y}\right)^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 + \sum \left(\hat{Y}_i - \overline{Y}\right)^2 \qquad [10b].$$

Dove: $\sum \left(Y_i - \overline{Y}\right)^2$ (sommatoria degli scarti tra valori osservati in `Y` e il valore medio di `Y`) corrisponde alla devianza totale (devianza di `Y` o devianza osservata); $\sum \left(Y_i - \hat{Y}_i\right)^2$ (sommatoria degli scarti tra valori osservati in `Y` e valori stimati mediante la retta di regressione) corrisponde alla devianza non spiegata (o devianza residua); $\sum \left(\hat{Y}_i - \overline{Y}\right)^2$ (sommatoria degli scarti tra i valori stimati mediante la regressione e il valore medio di `Y`) corrisponde alla devianza dovuta all'effetto o devianza spiegata dalla regressione.

Una volta calcolate le devianze, le varianze rispettive si calcolano dividendo le devianze per i relativi gradi di libertà. Secondo il teorema della scomposizione della devianza, i gradi di libertà si calcolano secondo le seguenti formule:

$$\left(gdl_T\right) = \left(gdl_{errore}\right) + \left(gdl_{regression\ e}\right) \qquad [10c];$$

$$\left(N - 1\right) = \left(N - k - 1\right) + \left(k\right) \qquad [10d].$$

Dove: N è dato dal numero di osservazioni (numero di soggetti); k è dato dal numero di variabili indipendenti inserite nel modello.

Per confrontare la due varianze e verificare se quella spiegata dalla regressione è significativamente maggiore di quella residua, si calcola la statistica F di Fischer. La varianza spiegata dalla regressione va al numeratore, quella residua al denominatore:

$$F = \frac{Var_{reg}}{Var_{res}} = \frac{\frac{Dev_{reg}}{k}}{\frac{Dev_{res}}{N-k-1}} \qquad [11].$$

L'ipotesi che sottoponiamo a verifica (ipotesi nulla o H_0) è che la varianza spiegata è uguale alla varianza residua, vale a dire che il modello di regressione non riduce l'errore di previsione della variabile dipendente. In altri termini l'ipotesi nulla che si sottopone a verifica assume che tutti i parametri *b* siano uguali a 0:

$$H_0 \Rightarrow \beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0 \qquad [11a].$$

Qualora questa ipotesi venga rifiutata viene considerata come vera l'ipotesi alternativa che assume che <u>almeno uno</u> dei predittori abbia un valore di *b* diverso da 0:

$$H_1 \Rightarrow \beta_1 \, o \, \beta_2 \, o \, \beta_3 \, o ... o \, \beta_k \neq 0 \qquad [11b].$$

Per la verifica dell'ipotesi si utilizza la distribuzione teorica della variabile casuale *F* di Fischer con gradi di libertà uguali ai gradi di libertà del numeratore e del denominatore nel rapporto tra le varianze (vedi formula [11]).

1.4. Il contributo dei singoli predittori

Se la verifica dell'ipotesi relativa alla capacità predittiva del modello ha portato a scartare l'ipotesi nulla è possibile approfondire l'analisi indagando il contributo di ciascun predittore considerato singolarmente.

A tal scopo si formula per ciascun predittore una specifica ipotesi nulla e la si sottopone a verifica. L'ipotesi nulla che viene formulata assume che ciascun valore di *b* osservato al livello campionario corrisponda a un valore nella popolazione uguale a 0. In altri termini, l'ipotesi nulla assume che il valore di *b* osservato sia una variazione casuale del valore 0 della popolazione. Per la verifica delle ipotesi relative a ciascun predittore si utilizza la statistica *t* che pone a confronto il valore *b* osservato con il valore $\beta_{i_{H_0}}$ atteso in base all'ipotesi nulla (vale a dire $\beta_{i_{H_0}} = 0$):

$$t = \frac{b_i - \beta_{i_{H_0}}}{s_{b_i}} \Rightarrow \frac{b_i}{s_{b_i}} \qquad [12].$$

Dove b_i corrisponde al coefficiente di regressione calcolato sul campione e s_i corrisponde alla deviazione standard del coefficiente stimata con la formula [6a].

Per l'interpretazione del valore *t* si utilizza la distribuzione della variabile casuale *t* di Student calcolando i gradi di libertà secondo la seguente equazione:

$$\text{gdl}_t = N - k - 1 \qquad [12a].$$

Dove *N* corrisponde al numero di osservazioni (o ampiezza del campione totale) e *k* al numero di predittori considerati nel modello.

Una volta verificata la relazione tra ciascun predittore e la variabile dipendente possiamo procedere all'interpretazione della relazione. È importante sottolineare che possono essere considerati esclusivamente i predittori i cui valori *b* sono risultati espressioni di β significativamente diversi da `0`, ma che nell'interpretazione si deve fare riferimento a tutte le variabili nel modello dato che gli effetti sono sempre parziali.

Nella regressione multipla, il coefficiente di regressione viene detto `parziale` dal momento che esprime la relazione che una data variabile indipendente ha con la variabile dipendente al netto delle altre variabili considerate nel modello. In pratica, il coefficiente di regressione parziale esprime la relazione unica che una data variabile indipendente ha con la variabile dipendente mantenendo costanti i valori delle altre variabili.

In termini pratici, ciascun coefficiente di regressione viene interpretato come la variazione in unità del valore atteso della variabile dipendente per una variazione unitaria della variabile esplicativa, mantenendo costanti i valori delle altre variabili nel modello.

Da ciò deriva che il valore del coefficiente *b* dipende dall'unità di misura delle variabili considerate. Ad esempio, se il valore di $b_j = 1.2$ significa che per ogni variazione unitaria della variabile indipendente *j* il valore atteso della variabile dipendente aumenta di 1.2 unità. Per cui se ad $X_j = 10$ corrisponde $\hat{Y} = 123$, ad $X_j = 11$ corrisponderà $\hat{Y} = 124.2$ (dove 124.2 - 123 = 1.2).

A causa della dipendenza dall'unità di misura delle variabili considerate, il coefficiente di regressione viene interpretato esclusivamente in base al

segno. Quando il segno è positivo significa che la relazione tra le variabili è positiva: al crescere di X_j corrisponde un aumento nei valori di $\hat{Y}$ o, in modo del tutto equivalente, al decrescere della X_j la $\hat{Y}$ decresce. Al contrario, quando il segno del coefficiente *b* è negativo significa che le due variabili sono legate da una relazione inversa per cui se aumenta il valore della variabile X_j i valori attesi della variabile $\hat{Y}$ diminuiscono, e viceversa. Il valore, la grandezza del coefficiente non standardizzato non viene mai interpretato come indice di forza della relazione.

Per avere un indice che esprima la forza della relazione tra la variabile indipendente e la variabile dipendente o per confrontare i coefficienti di regressione parziale tra loro, è necessario calcolare i `coefficienti di regressione standardizzati`.

I coefficienti di regressione standardizzati (simboleggiati con la lettera greca *beta*) possono essere ottenuti in due modi: considerando nel modello di regressione le variabili standardizzate (variabili espresse in *`punti z`*) o trasformando i coefficienti di regressione attraverso la seguente formula:

$$\beta_j = b_j \cdot \frac{s_j}{s_Y} \qquad [13].$$

I coefficienti di regressione standardizzati esprimono la relazione tra variabile indipendente e variabile dipendente usando come metrica le deviazioni standard delle due variabili. Pertanto, un cambiamento in X pari a una deviazione standard produce un cambiamento in Y pari a β deviazioni standard. Anche i coefficienti di regressione standardizzati sono dei coefficienti parziali.

I coefficienti β vengono interpretati in funzione del segno e dell'entità. Il segno esprime il tipo di relazione esistente tra variabile indipendente e variabile dipendente. Al pari dei coefficienti non standardizzati, quando il coefficiente standardizzato è positivo indica una relazione lineare positiva, per cui le variabili tendono a covariare nella stessa direzione, quando è negativo indica che le variabili hanno una relazione inversa.

L'entità viene valutata in base alla grandezza del coefficiente. Infatti, al pari dei `coefficienti di correlazione`, i coefficienti di regressione standardizzati hanno valori compresi tra `+1` e `−1`. Un valore pari ad `+1`

indica una relazione perfetta positiva, un valore pari a `−1` una relazione perfetta negativa, mentre valori uguali a `0` indicano un'assenza di relazione tra le variabili.

Quando sono riferiti allo stesso campione, i coefficienti di regressione standardizzati consentono di confrontare il peso dei diversi predittori nella determinazione delle variabile dipendente.

1.5. Esempio Regressione Semplice

Immaginiamo che un ricercatore sia interessato a verificare se il numero di amici che manifestano comportamenti devianti (ad esempio tendenza a bere superalcolici, uso di droghe, messa in atto di comportamenti al limite della legalità, ecc.) incide sulla tendenza a manifestare comportamenti devianti (Modello 1), e se tale effetto si manifesta indipendentemente dal livello intellettivo (QI) (Modello 2).

A tal scopo registra per 12 soggetti le seguenti variabili:

- amici: il numero di amici dediti alla messa in atto di comportamenti devianti (almeno uno);
- QI: valuta il quoziente intellettivo di ciascun soggetto mediante la scala WISCH-R;
- devianza: valuta per ciascun partecipante il numero di comportamenti devianti messi in atto negli ultimi 5 anni.

I dati raccolti sono riportati nella tabella 1.

A questo punto il ricercatore, per verificare il modello 1, procede analizzando la relazione tra la variabili indipendenti (amici) e la variabile dipendente (devianza) attraverso l'analisi della regressione semplice lineare.

Tabella 1
Matrice dei dati soggetti (SS)× variabili (VV)

Cod	*amici*	*QI*	*devianza*
1	3	106	0
2	7	93	10
3	3	94	9
4	5	107	12
5	8	96	17
6	2	118	0
7	7	110	14
8	5	107	12
9	4	105	1
10	5	113	11
11	3	104	7
12	8	119	12

A tal scopo, prima calcola per ciascuna variabile media, devianza e varianza (vedi tabella 2).

Tabella 2
Calcolo delle medie, devianza e varianza per ciascuna variabile

Cod	X	Y	$X - \overline{X}$	$Y - \overline{Y}$	$(X - \overline{X})^2$	$(Y - \overline{Y})^2$
1	3	0	-2	-8.75	4	76.5625
2	7	10	2	1.25	4	1.5625
3	3	9	-2	0.25	4	0.0625
4	5	12	0	3.25	0	10.5625
5	8	17	3	8.25	9	68.0625
6	2	0	-3	-8.75	9	76.5625
7	7	14	2	5.25	4	27.5625
8	5	12	0	3.25	0	10.5625
9	4	1	-1	-7.75	1	60.0625
10	5	11	0	2.25	0	5.0625
11	3	7	-2	-1.75	4	3.0625
12	8	12	3	3.25	9	10.5625
$\sum$	60	105		**Devianza**	48	350.25
Media	5	8.75		**Varianza**	4	29.1875

Una volta stimate varianze e devianze, il passo successivo è il calcolo dei parametri mediante il metodo `OLS` (formule [6] e [7]; vedi tabella 3).

Tabella 3
Calcolo dei parametri della regressione

Cod	$X - \overline{X}$	$(X - \overline{X})^2$	$Y - \overline{Y}$	$(X - \overline{X})*(Y - \overline{Y})$
1	-2	4	-8.8	17.5
2	2	4	1.3	2.5
3	-2	4	0.3	-0.5
4	0	0	3.3	0.0
5	3	9	8.3	24.8
6	-3	9	-8.8	26.3
7	2	4	5.3	10.5
8	0	0	3.3	0.0
9	-1	1	-7.8	7.8
10	0	0	2.3	0.0
11	-2	4	-1.8	3.5
12	3	9	3.3	9.8
totale		48		102

$$b_i = \frac{\sum_{j=1}^{n} (X_{ij} - \overline{X}_i)(Y_j - \overline{Y})}{\sum_{j=1}^{n} (X_{ij} - \overline{X}_i)^2} = \frac{102}{48} = 2.125$$

$$a = \overline{Y} - b_i \overline{X} = 8.75 - 2.125 \cdot (5) = -1.875$$

Dunque, la retta di regressione che, in base al metodo `OLS`, descrive al meglio la relazione tra numero di amici e messa in atto di comportamenti devianti è la seguente:

$$\hat{Y}_{devianza} = -1.875 + 2.125 \cdot (amici) \qquad [M1]$$

Ad esempio, secondo tale formula, chi non ha amici che mettono in atto comportamenti devianti (ovvero che hanno *amici* = `0`) ha un numero di comportamenti previsto pari a `-1.9` (intercetta); mentre chi ha almeno due amici che mettono in atto un comportamento deviante (*amici* = `2`) ha un numero di comportamenti devianti previsto pari a circa `2.4`.

$$\hat{Y}_{devianza} = -1.875 + 2.125 \cdot (2) = 2.375 \approx 2.4$$

A questo punto bisogna verificare se il modello ottenuto consente di migliorare la previsione rispetto al modello dell'ipotesi nulla. Il modello 1 infatti afferma che il numero di comportamenti devianti (variabile dipendente) dipende dal numero di amici che mettono in atto comportamenti devianti (variabile indipendente), mentre nel modello dell'ipotesi nulla si ipotizza che le diverse manifestazioni della variabile dipendente

(comportamenti devianti) non dipendano da alcuna variabile indipendente ma soltanto dalla componente d'errore.

Per valutare il *fit* del modello 1 bisogna verificare se lo scostamento tra valori attesi e valori osservati si riduce utilizzando il modello di regressione [M1]. Pertanto dobbiamo calcolare per ciascuna osservazione il valore atteso in base al modello [M1] e confrontarlo con il valore osservato. Successivamente, si procede alla verifica dell'ipotesi confrontando l'errore di previsione dei valori osservati mediante il modello e senza il modello. Per il calcolo dei valori attesi si procede per ciascuna osservazione come nell'esempio precedente, vale a dire sostituendo i valori osservati nella variabile indipendente nel modello [M1].

$$\hat{Y}_i = -1.875 + 2.125 \cdot (X_i) =$$

$$\hat{Y}_1 = -1.875 + 2.125 \cdot (3) = 4.5$$
$$\hat{Y}_2 = -1.875 + 2.125 \cdot (7) = 13$$
$$\hat{Y}_3 = -1.875 + 2.125 \cdot (3) = 4.5$$
$$\hat{Y}_4 = -1.875 + 2.125 \cdot (5) = 8.75$$
$$\hat{Y}_5 = -1.875 + 2.125 \cdot (8) = 15.125$$
$$\hat{Y}_6 = -1.875 + 2.125 \cdot (2) = 2.375$$
$$\hat{Y}_7 = -1.875 + 2.125 \cdot (7) = 13$$
$$\hat{Y}_8 = -1.875 + 2.125 \cdot (5) = 8.75$$
$$\hat{Y}_9 = -1.875 + 2.125 \cdot (4) = 6.625$$
$$\hat{Y}_{10} = -1.875 + 2.125 \cdot (5) = 8.75$$
$$\hat{Y}_{11} = -1.875 + 2.125 \cdot (3) = 4.5$$
$$\hat{Y}_{12} = -1.875 + 2.125 \cdot (8) = 15.125$$

Una volta calcolati per ciascuna osservazione i valori attesi in base al modello è possibile verificare la capacità predittiva del modello attraverso il confronto tra la varianza spiegata dal modello e la varianza non spiegata. Per il calcolo della varianza non spiegata si procede calcolando prima la devianza (vedi tabella 4) e poi dividendo la devianza per gli opportuni gradi di libertà. Per il calcolo della varianza spiegata si procede sottraendo alla devianza totale la devianza non spiegata dal modello e dividendo il risultato per gli opportuni gradi di libertà.

Tabella 4
Calcolo della devianza non spiegata (devianza residua)

Cod	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$\left(Y_i - \hat{Y}_i\right)^2$
1	0	4.5	-4.5	20.3
2	10	13.0	-3.0	9.0
3	9	4.5	4.5	20.3
4	12	8.8	3.3	10.6
5	17	15.1	1.9	3.5
6	0	2.4	-2.4	5.6
7	14	13.0	1.0	1.0
8	12	8.8	3.3	10.6
9	1	6.6	-5.6	31.6
10	11	8.8	2.3	5.1
11	7	4.5	2.5	6.3
12	12	15.1	-3.1	9.8
totale				133.5

Dal momento che la devianza totale è pari a 350.25 possiamo calcolare la devianza spiegata: 350.25 – 133.25 = 216.75. In base alla formula [10d] possiamo calcolare i gradi di libertà, che risultano rispettivamente 11 (12 – 1), 10 (12 – 1 – 1) e 1, e verificare l'ipotesi mediante il rapporto tra le varianze (vedi formula [11]):

$$F = \frac{Var_{reg}}{Var_{res}} = \frac{\frac{216.75}{1}}{\frac{133.5}{10}} = 16.236$$

Una volta calcolato il valore della statistica *F* questo deve essere interpretato utilizzando la distribuzione teorica *F* di Fischer con gradi di libertà pari a 1 e 10.

$$P(F[1,10] = 16.236) = .002401$$

Il valore di probabilità ottenuto p = .002 essendo inferiore al valore di probabilità critico α = .05 ci porta a rifiutare l'ipotesi nulla (H_0 = la varianza spiegata e la varianza d'errore sono uguali) e ad accettare l'ipotesi alternativa (H_1 = la varianza spiegata è maggiore della varianza d'errore). Infatti, possiamo dire che la probabilità che sia vera l'ipotesi nulla (H_0) è inferiore a .05 (quindi molto bassa) e per questo rifiutiamo l'ipotesi nulla.
Mediante la formula [8c] calcoliamo la percentuale di varianza spiegata dal modello:

$$R^2 = 1 - \frac{133.5}{350.25} = .6188$$

Dal momento che in questo caso si trattava di una regressione semplice (un solo predittore) siamo certi che il valore del coefficiente *b* o β (il coefficiente standardizzato) sono statisticamente diversi da `0` anche nella popolazione. Tuttavia, a scopo didattico, procediamo con la verifica dell'ipotesi anche del predittore utilizzando le formule [6a], [12] e [12a].

$$s_{b_i} = \sqrt{\frac{\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2}{\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2(1 - R_i^2)(n - k - 1)}} = \sqrt{\frac{133.5}{(48)\cdot(10)}} = .5274$$

$$\mathrm{t} = \frac{b_i - \beta_{i_{H_0}}}{s_{b_i}} \Rightarrow \frac{b_i}{s_{b_i}} = \frac{2.125}{.5274} = 4.029$$

$$\mathrm{gdl}_{\mathrm{t}} = N - k - 1 = 10$$

Una volta calcolato il valore della statistica *t* questo deve essere interpretato utilizzando la distribuzione teorica *t* di Student con gradi di libertà pari a 10.

$$P(t[10] = 4.029) = .002401$$

Come ci aspettavamo, il valore di probabilità ottenuto p = .002 essendo inferiore al valore di probabilità critico α = .05 ci porta a rifiutare l'ipotesi nulla (H_0 = il valore di *b* nella popolazione è uguale a `0`) e ad accettare l'ipotesi alternativa (H_1 = il valore di *b* della popolazione è diverso da `0`).

Di seguito vengono riportate le statistiche del modello 1 calcolate attraverso il *software* SPSS® (Tabella 5) e mediante il *software* libero R (versione R 2.4.1; R Development Core Team, 2006) (Tabella 6). In questo secondo caso la funzione utilizzata per il calcolo dei parametri è la funzione che consente di stimare i modelli lineari (`lm`[1]).

[1] Per un approfondimento sulla funzione `lm` digitare il comando `help(lm)` nella finestra di lavoro dei software R.

Tabella 5

Stima dei parametri del modello 1 mediante il software SPSS®

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	.787[a]	.619	.581	3.654

a. Stimatori: (Costante), amici

ANOVA[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	216.750	1	216.750	16.236	.002[a]
	Residuo	133.500	10	13.350		
	Totale	350.250	11			

Coefficienti[a]

		Coefficienti non standardizzati		Coefficienti standardizzati		
Modello		B	Errore std.	Beta	t	Sig.
1	(Costante)	-1.875	2.840		-.660	.524
	amici	2.125	.527	.787	4.029	.002

Tabella 6

Stima dei parametri del modello 1 mediante il software R

```
#Sintassi per la creazione dei dati
amici    <- c(3, 7, 3, 5, 8, 2, 7, 5, 4, 5, 3, 8)
devianza <- c(0, 10, 9, 12, 17, 0, 14, 12, 1, 11, 7, 12)

#Sintassi per l'esecuzione dell'analisi
Mod1 <- lm(devianza ~ amici)

# Visualizza i parametri del Modello 1
summary(Mod1)

CALL:
lm(formula = devianza ~ amici)

RESIDUALS:
   Min     1Q Median     3Q    Max
-5.625 -3.031  1.438  2.687  4.500

COEFFICIENTS:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.8750     2.8400  -0.660   0.5240
amici         2.1250     0.5274   4.029   0.0024 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.654 on 10 degrees of freedom
Multiple R-Squared: 0.6188,     Adjusted R-squared: 0.5807
F-statistic: 16.24 on 1 and 10 DF,  p-value: 0.002402

#Calcolo del coefficiente di regressione standardizzato
#definito "beta"
coef(Mod1)[2]*sd(amici)/sd(devianza)

    amici
0.7866662
```

1.6. Esempio Regressione Multipla standard

Nel Modello 2 il ricercatore si chiedeva se l'eventuale effetto della variabile indipendente sulla variabile dipendente fosse legato a una terza variabile: il QI. In pratica il ricercatore si chiede se la relazione tra le due variabili esiste anche quando tutti gli individui sono pareggiati rispetto al livello intellettivo. Per raggiungere tale obiettivo il ricercatore ha due possibilità. La prima, molto faticosa e complessa da realizzare, prevede di raccogliere un nuovo campione di osservazioni campionando esclusivamente individui con lo stesso QI. La seconda, più facilmente praticabile, prevede di utilizzare un modello di analisi dei dati che consenta di analizzare la relazione tra la variabile indipendente e la variabile dipendente al netto della terza variabile. In questa prospettiva la terza variabile può essere considerata come una variabile di disturbo o come una variabile che modera o che media la relazione tra le altre due (per un approfondimento sugli effetti di `moderazione` e `mediazione` si veda Baron e Kenny, 1986). Nell'esempio considerato il modello che risponde allo scopo è un modello di `regressione lineare multipla standard`. Infatti, come anticipato, nella regressione multipla i coefficienti di regressione vengono detti parziali dal momento che esprimono la relazione specifica che una data variabile indipendente ha con la variabile dipendente al netto delle altre variabili indipendenti considerate nel modello.

Individuato il modello di analisi dei dati da utilizzare non resta che eseguire l'analisi mediante uno dei programmi a scelta. I risultati sono riportati nella tabelle 7 e 8.

Tabella 7
Stima dei parametri del Modello 2 mediante il software SPSS®

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	.800[a]	.640	.560	3.745

a. Stimatori: (Costante), QI, amici

ANOVA[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	224.037	2	112.018	7.988	.010[a]
	Residuo	126.213	9	14.024		
	Totale	350.250	11			

Coefficienti[a]

		Coefficienti non standardizzati		Coefficienti standardizzati		
Modello		B	Errore std.	Beta	t	Sig.
1	(Costante)	8.448	14.614		.578	.577
	amici	2.093	.542	.775	3.859	.004
	QI	-.096	.133	-.145	-.721	.489

Analizzando i risultati, possiamo osservare come il valore di probabilità associato alla statistica *F* è pari a $p = .010$; essendo inferiore al valore di probabilità critico $\alpha = .05$, questo risultato ci porta a rifiutare l'ipotesi nulla (H_0 = la varianza spiegata e la varianza d'errore sono uguali, oppure che tutti i $\beta = 0$) e ad accettare l'ipotesi alternativa (H_1 = la varianza spiegata dal modello 2 è maggiore della varianza d'errore, oppure che almeno un β è diverso da 0). La varianza spiegata dal modello che considera entrambi i predittori (amici e QI) corrisponde a circa il 64%. Analizzando i risultati relativi alla verifica delle ipotesi per ciascun predittore i dati evidenziano che per il predittore amici il valore di probabilità associato alla statistica *t* corrisponde a $p = .004$, mentre il valore associato al predittore QI corrsisponde a $p = .489$. Nel primo caso, quindi, possiamo rifiutare l'ipotesi nulla, mentre nel secondo caso non possiamo rifiutare l'ipotesi nulla.

Da un punto di vista più generale, e rispondendo agli obiettivi dell'analisi, il ricercatore può concludere che l'avere degli amici che mettono in atto comportamenti devianti influenza il numero di comportamenti devianti che vengono messi in atto (Modello 1) e tale effetto si osserva indipendentemente dal livello intellettivo (Modello 2).

Tabella 8

Stima dei parametri del modello 2 mediante il software R

```
#Sintassi per la creazione dei dati
amici    <-  c(3, 7, 3, 5, 8, 2, 7, 5, 4, 5, 3, 8)
devianza <-  c(0, 10, 9, 12, 17, 0, 14, 12, 1, 11, 7, 12)
QI       <-  c(106,93,94,107,96,118,110,107,105,113,104,119)

#Sintassi per l'esecuzione dell'analisi
Mod2 <- lm(devianza ~ amici + QI)

# Visualizza i parametri del Modello 2
summary(Mod2)

CALL:
lm(formula = devianza ~ amici + QI)

RESIDUALS:
   Min     1Q Median     3Q    Max
-5.753 -2.383  1.230  3.012  3.346

COEFFICIENTS:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.44796   14.61362   0.578  0.57738
amici        2.09304    0.54233   3.859  0.00385 **
QI          -0.09588    0.13301  -0.721  0.48931
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Residual standard error: 3.745 on 9 degrees of freedom
Multiple R-Squared: 0.6396,     Adjusted R-squared: 0.5596
F-statistic: 7.988 on 2 and 9 DF,  p-value: 0.01012

#Calcolo dei coefficienti di regressione standardizzati
#definiti "beta"
coef(Mod2)[2]*(sd(amici))/sd(devianza)
coef(Mod2)[3]*(sd(QI))/sd(devianza)

    amici
0.7748349

        QI
-0.1447223
```

2

Regressione logistica semplice e multipla

Introduzione

Il modello di regressione logistica viene utilizzato quando si è interessati a studiare o analizzare la relazione causale tra una variabile dipendente dicotomica e una o più variabili indipendenti quantitative o dicotomiche.

2. La Regressione logistica Semplice e Multipla

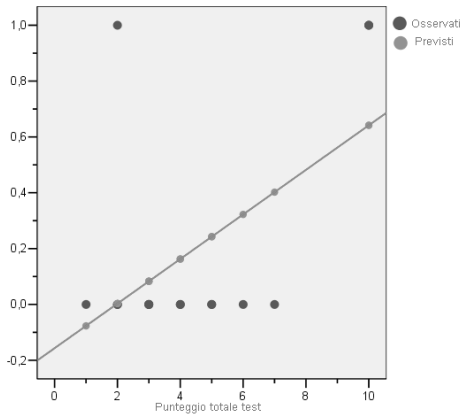
Come descritto nel capitolo precedente, quando la variabile dipendente (Y) è continua il valore stimato ($\hat{Y}$) può essere concepito come una stima della media condizionata di Y per ciascun valore della X (assumendo come vera la relazione tra X e Y)[2]. In questo caso, si assume che la variabile Y sia distribuita secondo la distribuzione normale.

Quando la variabile dipendente è dicotomica, e codificata come 0-1 (ad es. risposta giusta = 1, risposta sbagliata = 0), la distribuzione teorica di riferimento non dovrebbe essere più la normale ma la distribuzione binomiale. In questi casi, quindi, sebbene sia ugualmente possibile applicare il modello della regressione semplice, da un punto di vista matematico, un modello non lineare sarebbe più appropriato. Infatti, nella sua formulazione ($\hat{Y} = \alpha + \beta X$), il modello lineare implica che i valori della variabile dipendente ($\hat{Y}$) possano andare da $-\infty$ a $+\infty$. Se ad esempio si considera il grafico riportato in figura 1, in cui un modello dei regressione lineare è stato adattato con una variabile dipendente dicotomica, e si segue la linea di tendenza determinata dal modello lineare, possiamo notare che all'aumentare del punteggio totale nella variabile indipendente, sono accettabili valori previsti ($\hat{Y}$) maggiori di 1, e che al decrescere dei valori della X il modello prevede per la variabile dipendente ($\hat{Y}$) valori inferiori a 0.

[2] Se il predittore (X) è dicotomico (variabile *dummy*) questa relazione è ancora più evidente.

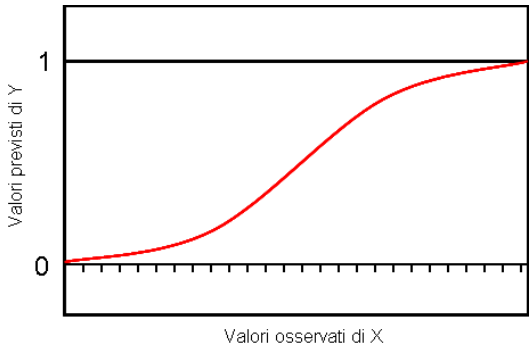
Figura 1

Modello di regressione lineare semplice con una variabile dipendente dicotomica



Tale previsione risulta non adeguata alla variabile dipendente che, come detto, può assumere esclusivamente valori `0 – 1`. In pratica, se la variabile dipendente è dicotomica, e se è influenzata dalla variabile `X`, allora si dovrebbe osservare che per valori molto alti di `X` (o molto bassi se la relazione è negativa) il valore in `Y` dovrebbe essere molto vicino ad `1` e non dovrebbe superare tale limite. Lo stesso dovrebbe avvenire in prossimità dello `0`. In pratica la curva che rappresenta la relazione tra `X` e `Y` dovrebbe essere di tipo `logistico` (vedi figura 2) e non lineare.

Figura 2



In questi casi, dunque, sarebbe più opportuno adattare un modello di regressione non-lineare.

Tra l'altro, la non linearità della relazione tra le variabili non consente di poter applicare il metodo OLS a meno che non si proceda ad opportune trasformazioni che rendano lineare la relazione. Si tratta di trasformazioni

che rendano lineare la relazione nei termini dei parametri (Berry e Feldman, 1985).

Una delle trasformazioni possibili è, ad esempio, la trasformazione logaritmica della variabile dipendente. Prima di illustrare i passaggi che portano alla linearizzazione della relazione è necessario aprire una breve parentesi per fare alcune considerazioni relative alla variabile dipendente considerata.

Nella regressione logistica la variabile dipendente definisce l'appartenenza a un gruppo (o all'altro). I valori che vengono assegnati ai livelli sono attribuiti in maniera arbitrara. Ciò che interessa, dunque, non è il valore atteso (o predetto), come nella regressione lineare, ma la probabilità che un dato soggetto appartenga a meno a uno dei due gruppi. Nonostante questo, è importante sottolineare che la scelta dei valori da assegnare influenza i risultati dell'analisi. Un modo per risolvere il dilemma dell'assegnazione dei valori ai livelli è quello di sostituire la probabilità (ad esempio di `Y = 1`) con *l'`odds`*: *odds* (`Y = 1`).

L'*odds* è un modo di esprimere la probabilità mediante un rapporto. Si calcola facendo il rapporto tra le frequenze osservate in un livello con le frequenze osservate nell'altro. Il valore dell'*odds* esprime il rapporto tra due categorie. Ad esempio, se ci sono 30 uomini e 12 donne (N = 42) possiamo dire che la probabilità di essere uomini è .714 (formula [1]), oppure che gli uomini sono il 71%. Se vogliamo esprimere la stessa informazione, mettendo in relazione le due categorie, possiamo ricorrere all'*odds*. Mediante l'*odds* (formula [2]) vediamo che la relazione tra uomini e donne è pari a 2.5, questo equivale a dire che per ogni donna ci sono 2.5 uomini.

$$P(M) = \frac{30}{42} = .714 \qquad [1]$$

$$odds\,(M) = \frac{30}{12} = 2.5 \qquad [2]$$

Per esprimere la relazione tra due categorie in funzione di un'altra variabile (valutare cioè l'associazione tra due variabili) è possibile utilizzare un altro indice chiamato *`odds ratio`* o rapporto tra gli *odds*. Tale indice si ottiene facendo un rapporto tra gli *odds* di una data variabile (ad esempio, la

variabile Y) ottenuti per ciascun livello della seconda variabile (ad esempio, la variabile X).

Ad esempio, se vogliamo valutare la relazione tra tipo di lavoro e sesso possiamo utilizzare una tabella di contingenza a doppia entrata e rappresentare la distribuzione di frequenze congiunte (Tabella 1). Quindi, la domanda che possiamo porci è la seguente: il rapporto (*odds*) tra uomini e donne è uguale nei differenti lavori? Se calcoliamo la percentuale di uomini nelle due tipologie di lavoro osserviamo che tra gli ingegneri il 90% è composto da uomini, mentre tra gli insegnanti la percentuale di uomini è del 55%.

Tabella 1

Tabella a doppia entrata che descrive la distribuzione congiunta della variabile sesso e della variabile lavoro

	Sesso		
Lavoro	*Uomini*	*Donne*	*Totale*
Ingegneri	18	2	20
Insegnanti	12	10	22
Totale	30	12	42

Per descrivere la diversa distribuzione delle categorie della variabile sesso nelle categorie della variabile lavoro mediante una statistica unica possiamo utilizzare l'*odds ratio*, che come detto precedentemente corrisponde al rapporto tra i rapporti tra le categorie. Da un punto di vista pratico, e facendo riferimento ai dati raccolti nella tabella 1, si ottiene che l'*odds ratio* ha un valore pari a 7.5 (formula [3]);

$$OR = \frac{18/2}{12/10} = \frac{18}{2} \cdot \frac{10}{12} = 7.5 \qquad [3].$$

Per l'interpretazione degli *odds ratio* si procede nel seguente modo: valori diversi da 1 indicano un'associazione tra le variabili. In questo caso, poiché il valore è diverso da 1, si può dire che esiste un'associazione tra il sesso e la professione. In particolare, si osserva che la proporzione degli uomini è 7.5 volte maggiore tra gli ingegneri rispetto agli insegnanti.

Nella tabella 2 viene riportata la distribuzione congiunta delle variabili X e Y. Dove la variabile X è una variabile discreta quantitativa mentre la

variabile Y è una variabile dicotomica. Come si vede nella tabella è possibile calcolare per ciascun livello della variabile X la distribuzione di frequenza della variabile Y.

Tabella 2

Confronto tra frequenze, frequenze relative (%), odds e logit per la distribuzione della variabile Y in funzione dei valori della variabile X

Punteggio (x_i)		**0**	**1**	**2**	**3**	**4**	**5**	**6**
Y = 1	*f*	2	3	5	2	5	6	7
successo	%	.20	.30	.50	.20	.50	.60	.70
Y = 0	*f*	8	7	5	8	5	4	3
fallimento	%	.80	.70	.50	.80	.50	.40	.30
odds (1/0)		0.25	0.43	1	0.25	1	1.5	2.3
log(*odds*(1/0))		-1.39	-0.85	0	-1.39	0	0.41	0.85

Una volta calcolate le frequenze, è possibile calcolare le frequenze relative o frequenze percentuali. Se confrontiamo i valori ottenuti notiamo subito che le frequenze relative dei due livelli della variabile `Y` sono del tutto speculari. Conoscendo le frequenze è poi possibile calcolare il rapporto ovvero l'*odds* tra le categorie della variabile `Y` per ciascun livello della variabile `X`. Infine, una volta calcolato l'*odds* è possibile calcolare il suo logaritmo naturale ovvero il *logit*. Se confrontiamo la distribuzione delle frequenze, delle frequenze relative, degli *odds* e dei *logit* possiamo notare come tutte queste statistiche forniscono la stessa informazione sebbene con valori matematicamente differenti. Quando le categorie successi (`Y = 1`) e fallimenti (`Y = 0`) sono equiprobabili le frequenze relative sono uguali a `.5` per entrambe le categorie di `Y`, gli *odds* sono uguali ad `1`, mentre i *logit* sono uguali a `0`. Quando il numero di successi è maggiore del numero dei fallimenti le frequenze relative assumono valori superiori a `.5` per la categoria `Y = 1` e minori per la categoria `Y = 0`, gli *odds* assumono valori superiori ad `1`, mentre i *logit* valori superiori allo `0`. Infine, quando il numero di successi è inferiore al numero dei fallimenti le frequenze relative assumono valori inferiori a `.5` per la categoria `Y = 1` e superiori per la categoria `Y = 0`, gli *odds* assumono valori inferiori ad `1`, mentre i *logit* valori negativi. In pratica, mentre le frequenze relative hanno un *range* di variabilità che va da `0` a `1`, gli *odds* hanno un range di variabilità che va da

0 a più infinito, mentre i *logit* possono variare da meno infinito a più infinito.

Chiusa la parentesi, possiamo riprendere il filo del discorso. Per esprimere la relazione tra la variabile indipendente e la variabile dipendente in termini lineari possiamo partire dalla seguente formulazione in cui il valore atteso della variabile dipendente è la probabilità ($\hat{Y} = \mu_Y = P_{(Y=1)}$), per cui la probabilità di Y = 1 come funzione lineare di X diventa:

$$P(Y = 1) = \alpha + \beta X \qquad [4a].$$

Come illustrato, questo modello non è adeguato, poiché i valori della probabilità sono compresi tra 0 e 1, mentre il termine $\alpha + \beta X$ può assumere valori che vanno da – ∞ a + ∞. Allora, per provare a risolvere il problema possiamo applicare la trasformazione esponenziale al termine di destra della funzione che diventa:

$$P(Y = 1) = e^{\alpha + \beta X} \qquad [4b].$$

Anche questa trasformazione, seppure consente di restringere i valori dell'equazione entro il range 0 + ∞, non risolve completamente il problema. A tal scopo possiamo applicare la trasformazione logistica che consente di controllare i valori e restringerli nel range della probabilità (0; 1):

$$P(Y = 1) = \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}} \qquad [4c].$$

Nel caso di variabili dicotomiche l'*odds* diventa:

$$odds\ (Y = 1) = \frac{P(Y = 1)}{1 - P(Y = 1)} \qquad [4d].$$

Dove: $P(Y = 0) = [1 - P(Y = 1)]$ serve a esprimere la probabilità della seconda categoria in funzione della prima.

Se definiamo in questo modo la probabilità di Y = 0 possiamo calcolare l'*odds* di Y = 1 che diventa:

$$odds_{Y=1} = \frac{\frac{e^{(\alpha + \beta X)}}{1 + e^{(\alpha + \beta X)}}}{\frac{1}{1 + e^{(\alpha + \beta X)}}} = e^{(\alpha + \beta X)} \qquad [5].$$

Infine, per le proprietà dei logaritmi ($\ln(e^x) = x$), se calcoliamo il logaritmo dell'*odds* osserviamo che il `logaritmo naturale` dell'*odds* di `Y = 1` è funzione lineare della variabile `X`:

$$\ln\left(odds_{Y=1}\right) = \alpha + \beta X \qquad [5a].$$

Applicando queste trasformazioni, l'equazione della relazione tra le variabili X_k e `Y` diviene:

$$P(Y=1) = \frac{e^{(\alpha + X_1\beta_1 + X_2\beta_2 + \ldots + X_i\beta_i + \varepsilon)}}{1 + e^{(\alpha + X_1\beta_1 + X_2\beta_2 + \ldots + X_i\beta_i + \varepsilon)}} \qquad [6].$$

È importante sottolineare che la probabilità, l'*odds* e il *logit* sono tre differenti modi di esprimere esattamente la stessa cosa. La trasformazione in *logit* serve solo a garantire la correttezza matematica dell'analisi.

2.1. La stima dei parametri

Come nella regressione lineare, nell'analisi della regressione logistica l'interpretazione della relazione tra variabili indipendenti e variabile dipendente avviene mediante la valutazione dei parametri del modello.

Nella pratica, i valori dei parametri della popolazione non sono noti, essi vengono stimati a partire da un numero finito di osservazioni: le osservazioni campionarie.

Per distinguere la regressione campionaria da quella della popolazione il modello di regressione logistica viene scritto utilizzando le lettere latine:

$$P(Y=1) = \frac{e^{(a + b_1X_1 + b_2X_2 + \ldots + e)}}{1 + e^{(a + b_1X_1 + b_2X_2 + \ldots + e)}} \qquad [7].$$

Nella stima dei parametri della regressione logistica il metodo OLS non può essere applicato (non sono verificati gli assunti), si utilizza l'algoritmo di `massima verosimiglianza` (*maximum likelihood* - ML) che stima i parametri del modello in modo da massimizzare la funzione (*log-likelihood function*) che indica quanto è probabile ottenere il valore atteso di `Y` dati i valori delle variabili indipendenti.

Nel metodo della massima verosimiglianza, la soluzione ottimale viene raggiunta partendo da dei valori di prova per i parametri (valori arbitrari) i quali successivamente vengono modificati per vedere se la funzione può

essere migliorata[3]. Il processo viene ripetuto (*iteration*) fino a quando la capacità di miglioramento della funzione è infinitesimale (*converge*).

2.2. La valutazione del *fit* del modello

Nell'interpretazione del modello della regressione logistica ci si avvale di statistiche del tutto simili alle statistiche che esprimono l'adeguatezza del modello nel riprodurre i dati osservati nella regressione lineare (F e R^2).

Similmente alla somma dei quadrati, nella regressione logistica si utilizza il *log likelihood* come criterio per la scelta dei parametri del modello. In particolare, per ragioni matematiche, si utilizza il valore del *log likelihood* moltiplicato per -2, e abbreviato come $-2LL$. Valori grandi e positivi indicano una bassa capacità di previsione del modello.

Nel modello con la sola intercetta il valore della statistica $-2LL$ rappresenta quello che nella regressione lineare corrisponde alla devianza (o somma dei quadrati totale, SQ o SST) e può essere indicata come D_0.

Il calcolo della D_0 viene ottenuto mediante la seguente equazione:

$$D_0 = -2\{n_{Y=1} \ln[P(Y=1)] + n_{Y=0} \ln[P(Y=0)]\} \qquad [8]$$

Dove con $n_{Y=1}$ intendiamo il numero di casi per i quali $Y = 1$, con $n_{Y=0}$ il numero di casi per i quali $Y = 0$, N il numero totale di casi e con $P(Y=1) = {}^{n_{Y=1}}\!/_{N}$ la probabilità che $Y = 1$.

In un modello in cui la variabile Y si distribuisce come riportato nella tabella 2 il calcolo diventa:

Tabella 3
Distribuzione di frequenze della variabile sesso

Y	*f*
Y = 1	17
Y = 0	13
totale	*N* = 30

[3] È importante sottolineare che quando le assunzioni dell'OLS sono verificate, le stime dei parametri ottenute mediante il metodo OLS e il metodo ML sono identiche (Eliason, 1993). In questo senso il metodo OLS può essere considerato un caso particolare della ML; quando i parametri sono stimabili direttamente, senza iterazioni.

$$D_0 = -2\left\{17 \cdot \ln \frac{17}{30} + 13 \cdot \ln \frac{13}{30}\right\} = \qquad [8a];$$

$$D_0 = 41.054 \qquad [8b].$$

Nel modello che contiene sia l'intercetta sia la/le variabile/i indipendente/i, il valore della statistica `−2LL` rappresenta la parte di variabilità dei dati che non viene spiegata dal modello (devianza d'errore) e viene indicata come D_M. Lo scarto tra D_0 e D_M rappresenta la parte di variabilità spiegata dalle variabili indipendenti o variabilità spiegata dal modello; e viene indicata come G_M:

$$D_0 - D_M = G_M \qquad [9].$$

G_M viene anche chiamato Chi-quadrato (χ^2) del modello e indica la quantità di riduzione dell'errore dovuta al modello; ma solo se i modelli sono `nidificati` (*nested*). Un modello A (M_A) [9a] si dice *nested* in un modello B (M_B) [9b] se il modello A è composto da alcuni dei termini contenuti nel modello B, e non ve ne sono di diversi, mentre nel modello B vi sono anche termini aggiuntivi:

$$M_A = a + b \qquad [9a];$$

$$M_B = a + b + c \qquad [9b].$$

La differenza tra i due `−2LL` (G_M), se calcolata su modelli *nested*, può essere interpretata come statistica del $\chi 2$ e utilizzata per la verifica dell'ipotesi nulla del modello:

$$H_0 \Rightarrow \beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0 \qquad [10].$$

Se il G_M risulta statisticamente significativo (cioè quando il valore ha una $p <$ .05) l'ipotesi H_0 può essere rifiutata; vale a dire che la previsione (`Y = 1`) può essere migliorata se consideriamo i predittori. Per la verifica dell'ipotesi i gradi di libertà sono definiti dal numero di predittori (*gdl* = *k*).

Se manteniamo la similitudine tra la statistica `−2LL` e la devianza della regressione, per ottenere una statistica simile all'R^2 (*pseudo* R^2) si può utilizzare il rapporto di verosimiglianza (*likelihood ratio*) secondo la seguente formula:

$$R_L^2 = \frac{G_M}{D_0} = \frac{G_M}{G_M + D_M} \qquad [11].$$

Nella letteratura, questa statistica è nota come indice di McFadden (1974).

Analogamente a quanto avviene nella regressione, R^2_L può essere considerato come la porzione di riduzione dell'errore ($-2LL$) dovuta al modello. Detto in altri termini, indica quanto considerare i predittori riduce la variazione nei dati (stimata a partire dal modello nullo).

2.3. La valutazione della capacità predittiva del modello

In aggiunta alle statistiche relative alla valutazione dell'adeguatezza del modello (*goodness of fit*), un ulteriore aspetto che viene preso in considerazione è la capacità predittiva del modello.

Nella maggior parte dei casi, infatti, oltre ad essere interessati a conoscere se il modello è in grado di prevedere adeguatamente $P(Y_i = 1)$ possiamo essere interessati anche a voler verificare se il modello è in grado di prevedere adeguatamente l'appartenenza dei casi ad un gruppo o ad un altro, quindi siamo molto più interessati alla *tabella delle classificazioni*.

Nella regressione logistica la tabella delle classificazioni è una tabella a due vie (2×2) nella quale, per ciascuna osservazione, si pongono a confronto i valori osservati con i valori previsti dal modello.

L'indice per la valutazione della capacità predittiva del modello maggiormente impiegato si basa sulla valutazione della riduzione dell'errore in percentuale (*proportional change in error*):

$$\text{Efficienza predittiva} = \frac{(\text{errori senza il modello}) - (\text{errori con il modello})}{(\text{errori senza il modello})} \qquad [12]$$

Gli errori con il modello sono dati dal numero di casi per cui il valore previsto è diverso dal valore osservato. Gli errori senza il modello si calcolano in modi diversi e dipendono dall'uso che si vuole fare del modello: *a*) predizione; *b*) classificazione; *c*) selezione.

Nei modelli predittivi, l'obiettivo è valutare se un dato caso soddisfa o meno un criterio (es. successo, presenza di un sintomo, ecc.). Non ci sono limiti posti a priori per cui tutti potrebbero appartenere a un'unica categoria.

Nei modelli di classificazione, l'obiettivo è simile ai modelli predittivi solo che si assume che il modello debba ricostruire le proporzioni tra le categorie

così come sono state osservate. Se il modello fallisce in questo compito viene valutato come non adeguato.

Nei modelli di selezione, l'obiettivo è quello di accettare o rifiutare casi stabilendo a priori il numero di elementi che possono entrare in una data categoria (es. stabilendo di volere selezionare il 10% dei candidati).

Per la valutazione dell'efficienza predittiva del modello sono state impiegate molte statistiche comunemente utilizzate per analizzare le tabelle di contingenza: ϕ; γ di Goodman e Kruskal; il κ; il coefficiente *r* di Pearson; e gli *odds ratio*.

Per i modelli di previsione, per il calcolo degli errori senza il modello si può utilizzare la moda come valore atteso di ogni caso. Questo metodo è lo stesso di quello impiegato per il calcolo della statistica λ (di Goodman e Kruskal). Tale statistica proposta da Ohlin e Duncan (1949) viene detta λ_p.

$$\lambda_p = \frac{\sum_{j=1}^{k} n_{Mj} - \max(R_i)}{N - \max(R_i)} \qquad [13];$$

dove:

N = ampiezza del campione
n_{Mj} = frequenza max nella colonna j-esima
R_i = il più grande totale di riga

Per i modelli di classificazione, un buon modo per calcolare gli errori senza il modello (esm) è basarsi sulla formula:

$$esm = \sum_{i=1}^{N} f_i \left[{}^{(N - f_i)} \big/ _{N} \right] \qquad [14];$$

dove:

N = ampiezza del campione
f_i = numero di casi nella categoria i

Quest'ultimo metodo è lo stesso usato per l'indice τ di Goodman e Kruskal. L'indice adattato alle tavole di previsione è stato proposto da Elecka (1980) e viene simboleggiato come τ_p. Questo indice corregge il numero atteso di errori in base alla differenza di partenza fra le categorie. Se il valore è negativo significa che il modello non migliora la previsione. Se è 1 la previsione è perfetta. Tale statistica si calcola nel modo seguente:

$$\tau_p = \frac{ad - bc}{\sqrt{[ad + bc + (ab + cd)][ad + bc + (ac + bd)]}} \qquad [15];$$

dove la lettera *a* corrisponde al numero di frequenze nella prima cella (in alto a sinistra) della tabella di classificazione (2×2); la lettera *b* alla seconda cella (in alto a destra); la lettera *c* alla terza cella (in basso a sinistra); e la lettera *d* alla quarta e ultima cella della tabella (in basso a sinistra).

Per i modelli di selezione, per il calcolo dell'efficienza del modello si può utilizzare una statistica che confronta per ciascuna cella lo scarto tra i valori attesi e i valori osservati. Se il valore è negativo significa che il modello non migliora la previsione. Tale statistica è detta ϕ_p e si calcola:

$$\phi_p = \frac{ad - bc}{.5[(a+b)(b+d)+(c+d)(a+c)]} \qquad [16].$$

2.4. Il contributo dei singoli predittori

Al pari della regressione lineare, anche nella regressione logistica siamo interessati a valutare il contributo specifico di ogni variabile indipendente sulla variabile dipendente, testandone la sua significatività. Come nella regressione lineare, la valutazione dei singoli contributi viene fatta solo quando il modello nel suo complesso si è rivelato utile a migliorare la previsione della variabile dipendente.

Per la valutazione del contributo di ciascuna variabile si considerano i coefficienti di regressione. A tal scopo possiamo considerare sia i coefficienti non standardizzati (se siamo interessati alle unità di misura) sia i coefficienti di regressione standardizzati (che esprimono la relazione tra le variabili nei termini delle deviazioni standard).

Il modo più utilizzato per valutare il contributo di ciascun predittore sulla variabile dipendente è mediante la statistica di Wald (W_k):

$$W_k^2 = \left(\frac{b_k}{s_{b_k}}\right)^2 \qquad [17]$$

Tale statistica segue la distribuzione della variabile casuale Chi-quadro con 1 grado di libertà.

Per porre a confronto variabili che hanno delle unità di misura differenti è necessario calcolare i coefficienti di regressione standardizzati.

Un coefficiente standardizzato è un coefficiente che è stato calcolato su variabili che hanno come unità di misura la deviazione standard. Tali coefficienti indicano di quante deviazioni standard varia la varabile

dipendente per ogni variazione unitaria (standard) della variabile indipendente.

Nel caso della regressione logistica i coefficienti standardizzati (b^*_{YX}) indicano di quante deviazioni standard si modifica il *logit* della Y_i per ogni variazione standard della variabile X_{ki}. La formula per il calcolo è la seguente:

$$b^*_{YX} = \frac{\left(b_{YX} \cdot s_X\right) \cdot R_{\hat{Y}Model}}{s_{\log it\left(\hat{Y}\right)}} \qquad [18];$$

dove:

b_{YX} = coeff. di regression e non stand.
s_X = dev. st. di X
$R_{\hat{Y}Model}$ = coeff. di regression e lineare
$s_{\log it(\hat{Y})}$ = dev. st. di log $it(\hat{Y})$ stimato

Un ulteriore parametro che può essere utilizzato per l'interpretazione della relazione tra le variabili è l'*odds ratio* che nell'output dei *software* viene riportato come *exp*(B).

Tale valore esprime la variazione della variabile dipendente in funzione di variazioni della variabile indipendente.

Se il valore è superiore a 1 significa che all'aumentare della variabile indipendente aumenta la probabilità di $Y = 1$. Al contrario, se il valore è inferiore a 1 significa che ad aumentare della variabile indipendente decresce la probabilità che $Y = 1$.

È importante sottolineare sia che l'*odds ratio* ha la stessa interpretazione del coefficiente di regressione, sia che per confrontare i differenti livelli di probabilità ($Y = 1$), nei diversi livelli delle variabili indipendenti, è necessario calcolare la probabilità e non basta rifarsi ai valori dell'*odds*.

2.5. Esempio Regressione Logistica Semplice

Immaginiamo che un ricercatore sia interessato a verificare se il livello di ansia incide sulla tendenza a manifestare attacchi di panico (Modello 1), e se l'eventuale effetto si manifesti indipendentemente dall'età (Modello 2).

A tal scopo registra per 14 soggetti le seguenti variabili (Tabella 4):

- `anx`: il punteggio riportato da ciascun individuo ad una scala che misura il livello di ansia STAI (misurazione ad intervalli nel range `1-10`);
- `age`: età espressa in anni;
- `panic`: valuta per ciascun partecipante la presenza/assenza di attacchi di panico negli ultimi 5 mesi (range `0-1`).

Tabella 4

Matrice dei dati soggetti (SS)× variabili (VV)

cod	*anx*	*age*	*panic*
1	23	28	0
2	23	28	0
3	23	40	0
4	23	18	0
5	23	40	0
6	23	28	1
7	27	41	0
8	27	18	1
9	28	20	0
10	28	41	1
11	28	16	1
12	28	19	1
13	28	20	1
14	28	18	1

Tabella 5

Costruzione della matrice (SS×VV) mediante il software R

```
#Sintassi per la creazione della matrice di dati: nevrosi
nevrosi<-data.frame (
       cod = c(1:14),
       anx =  c(23,23,23,23,23,23,27,28,27,28,28,28,28,28),
       age =    c(28,28,40,28,18,40,18,41,41,16,19,20,18,20),
       panic = c( 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0)  )
nevrosi

   cod  panic   anx age
1    1      0    23  28
2    2      0    23  28
3    3      0    23  40
4    4      1    23  28
5    5      0    23  18
6    6      0    23  40
7    7      1    27  18
8    8      1    28  41
9    9      0    27  41
10  10      1    28  16
11  11      1    28  19
12  12      1    28  20
13  13      1    28  18
14  14      0    28  20
```

Data l'ipotesi di causalità implicata nel Modello 1, il ricercatore procede analizzando la relazione tra la variabili indipendenti (ansia) e la variabile dipendente (panico) mediante una regressione. Dal momento che la variabile criterio è una variabile dicotomica, sceglie di eseguire una analisi

della regressione logistica semplice utilizzando come criterio la manifestazione di attacchi di panico (misurati come `0` = assenti, `1` = presenti) e come variabile indipendente il livello di ansia (misurato con la scala STAI).

Per verificare la sua ipotesi, calcola la devianza residua del modello dell'ipotesi nulla (Modello 0) e quella del modello dell'ipotesi alternativa (Modello 1). In entrambi i casi si tratta di utilizzare l'equazione [8]. Ottenute le due devianze, procede ad effettuare il loro confronto mediante la distribuzione teorica del Chi-quadrato (formula [9]). La sua ipotesi nulla è che il livello di ansia non influenzi la manifestazione di attacchi di panico o la probabilità di manifestare attacchi di panico. L'ipotesi alternativa, invece, assume che la probabilità di manifestare attacchi di panico dipenda dal livello di ansia. In termini parametrici, l'ipotesi nulla assume che il parametro β della variabile ansia (che esprime la relazione tra ansia e attacchi di panico nella popolazione) sia uguale a 0 e che quindi che il Modello 1 non sia in grado di ridurre significativamente la devianza residua del Modello 0. L'ipotesi alternativa, invece, assume che il parametro β della variabile ansia sia diverso da 0.

Attraverso il software R calcoliamo quindi il modello dell'ipotesi nulla (Modello 0) e il modello dell'ipotesi alternativa (Modello1). A tal fine applichiamo un modello generale linearizzato (utilizzando la funzione[4] `glm`) e definendo come funzione `link` la funzione `binomiale`.

Rispettivamente:

```
M0 <- (glm(panic ~ 1, family=binomial))
```
[19]

per il modello dell'ipotesi nulla (Modello 0) e

```
M1 <- (glm(panic ~ 1+anx, family=binomial))
```
[20]

per il modello dell'ipotesi alternativa (Modello 1).

Per confrontare le due devianze utilizziamo la funzione `anova` e specifichiamo come distribuzione teorica per testare statisticamente l'ipotesi nulla la distribuzione `Chi-quadrato`:

```
anova(M0, M1, test="Chisq")
```
[21]

[4] Per un approfondimento sulla funzione `glm` digitare il comando `help(glm)` nella finestra di lavoro del software R. Per un approfondimento sui modelli lineari generalizzati vedere Gill (2001) e Miceli (2001).

Tabella 6

Risultato del confronto tra il Modello 0 e il Modello 1

```
> M1<-(glm(panic ~ anx, family=binomial))
> M2<-(glm(panic ~ anx + age, family=binomial))
> anova(M0, M1, test="Chisq")

Analysis of Deviance Table

Model 1: panic ~ 1
Model 2: panic ~ anx
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        13    19.4081
2        12    13.9181  1   5.4901    0.0191
```

Il valore di probabilità che si riferisce al test statistico del confronto tra le devianze è p = .0191. Dal momento che è inferiore al valore di probabilità critico α = .05, questo ci porta a rifiutare l'ipotesi nulla (H_0 = la varianza spiegata dal Modello 1 è casuale) e ad accettare l'ipotesi alternativa (H_1 = la varianza spiegata dal Modello 1 è maggiore della varianza spiegata dal modello della sola intercetta, Modello 0).

Mediante la formula di McFadden [11] calcoliamo la percentuale di varianza spiegata dal Modello 1 che corrisponde a circa .28 (Tabella 7).

Tabella 7

Calcolo dell'indice di varianza spiegata (pseudo R^2) del Modello 1

```
> 5.4901 / 19.4081
[1] 0.2828767

> # Oppure in modo del tutto equivalente
> (M0$deviance-M1$deviance)/M0$deviance [5]
[1] 0.2828749
```

A questo punto, per interpretare il tipo di relazione che esiste tra il livello di ansia e la manifestazione di attacchi di panico analizziamo i parametri del Modello 1.

[5] In questo esempio, sono stati utilizzati i valori delle devianze residue disponibili negli oggetti M0 e M1. Per avere una descrizione delle caratteristiche di un oggetto e delle sue variabili è possibile utilizzare la funzione `str()`. Per maggiori informazioni digitare il comando `help(str)` nella finestra di lavoro di R.

Tabella 8

Parametri stimati del Modello 1

```
> summary(M1)

Call:
glm(formula = panic ~ anx, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.76093  -0.58098   0.05481   0.69059   1.92999

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.5196     7.6964  -2.016   0.0437 *
anx           0.6011     0.2945   2.041   0.0412 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19.408  on 13  degrees of freedom
Residual deviance: 13.918  on 12  degrees of freedom
AIC: 17.918

Number of Fisher Scoring iterations: 4
```

I risultati evidenziano che la relazione tra ansia e attacchi di panico è rappresentata dalla seguente funzione:

$$\ln\left(odds_{panico=1}\right) = -15.52 + 0.60 \cdot (ansia) \qquad [22].$$

Questo significa che se il livello di ansia è uguale a `0` il valore del logaritmo naturale dell'*odds* (*logit*) della variabile attacco di panico è uguale a `-15.52`. Poiché se il *logit* è inferiore a `0` il rapporto tra `Y = 1` e `Y = 0` è a favore del denominatore, questo significa che quando la variabile ansia è uguale a `0` la probabilità di osservare un valore di `Y = 1` è molto bassa (0.000000182). Per conoscere il valore esatto dell'*odds* basta calcolare l'esponenziale del parametro `a` che in R si può calcolare con la seguente formula:

```
> exp(-15.52)                    [23]
[1] 1.818652e-07                 [24]
```

Se invece consideriamo il parametro *b* possiamo osservare che la relazione tra la variabile ansia e la variabile attacchi di panico è una relazione di tipo positivo. Questo significa che per ogni incremento unitario del punteggio di ansia il valore del logaritmo naturale dell'*odds* (*logit*) della variabile attacco di panico aumenta di `.601` unità. Detto in altri termini all'aumentare dei

livelli di ansia aumenta la probabilità di manifestare un attacco di panico. Anche in questo caso, calcolando il valore esponenziale del parametro, che corrisponde a `1.822`, possiamo conoscere di quanto varia l'*odds* per ogni variazione unitaria della variabile indipendente ansia.

```
> exp(.601)          [25]
[1] 1.824124         [26]
```

Mediante i parametri ottenuti è quindi possibile conoscere la probabilità di osservare un attacco di panico per ogni specifico livello di ansia. Basta sostituire i valori osservati nella formula [6] e definire uno specifico livello per la variabile indipendente. Ad esempio, se un dato individuo ha un punteggio di ansia pari a `30` la sua probabilità di manifestare un attacco di panico corrisponde a:

$$P(Y=1) = \frac{e^{(-15.52+0.6\cdot(30))}}{1+e^{(-15.52+0.6\cdot(30))}} \qquad [27]$$

$$P(Y=1) = \frac{e^{(2.5134)}}{1+e^{(2.5134)}} = .9227 \qquad [28]^{6}$$

Mediante i parametri stimati, è quindi possibile calcolare per ciascuna osservazione la probabilità di manifestare un attacco di panico in base al punteggio sulla scala dell'ansia. Se stabiliamo che chi ha una probabilità stimata superiore a .5 sono individui che manifestano un attacco di panico, mediante le probabilità risultanti possiamo classificare le unità osservative in individui con attacco di panico e individui senza attacco di panico; e confrontare la classificazione ottenuta mediante i parametri del modello con i valori realmente osservati per la variabile dipendente. Questo ci consente di creare la tabella a due vie nota come tabella delle classificazione e calcolare ulteriori indici di adeguatezza del modello.

[6] Il risultato si può ottenere mediante la seguente funzione in R:

```
> (exp(-15.52+0.6*30))/(1+(exp(-15.52+0.6*30)))
[1] 0.9227278
```

Tabella 9

Costruzione della tabella delle classificazioni e calcolo delle percentuali di classificazione corrette

```
> # Questo comando consente di ottenere i valori di probabilità previsti
> # del Modello 1

> M1$fitted.values
        1         2         3         4         5         6
0.1552941 0.1552941 0.1552941 0.1552941 0.1552941 0.1552941
        7         8         9        10        11        12
0.6705879 0.7878432 0.6705879 0.7878432 0.7878432 0.7878432
       13        14
0.7878432 0.7878432

> # Questo commando arrotonda i valori previsti al numero intero
> # con 0 decimali
> panic.pre <- round(M1$fitted.values,0)
> panic.pre
 1  2  3  4  5  6  7  8  9 10 11 12 13 14
 0  0  0  0  0  0  1  1  1  1  1  1  1  1

> panic.oss <- panic
> # Crea la tabella delle classificazioni
> table.class <- table(panic.pre, panic.oss)
> table.class
         panic.oss
panic.pre 0 1
        0 5 1
        1 2 6

> # Questa serie di comandi estrae prima i valori della tabella
> # e poi esegue il calcolo delle percentuali (%)
> a <-table.class[1]
> b <-table.class[2]
> c <-table.class[3]
> d <-table.class[4]
> a/(a+b)             # % corrette valore 0 della VD
[1] 0.7142857
> d/(c+d)             # % corrette valore 1 della VD
[1] 0.8571429
> (a+d)/(a+b+c+d)     # % corrette (0+1) della VD
[1] 0.7857143
```

Riferimenti Bibliografici

Barbaranelli, C. (2003). *Analisi dei dati*. Milano: Led.

Baron, R.M., Kenny, D.A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations, 51(6), 1173-1182.

Berry, W.D., Feldman, S. (1985). *Multiple Regression in Practice* (Sage University Paper Series on Quantitative Applicatiosns in the Social Science). Newbury Park, CA: Sage.

Caudek, C., Luccio, R. (2001). *Statistica per psicologi*. Bari: Gius. Laterza & Figli Spa.

De Carlo, N., Robusto, E. (1996). *Teoria e tecniche di campionamento nelle scienze sociali*. Milano: LED.

Keppel, G., Saufley, W.H., Tokunaga, H. (2001). *Disegno sperimentale e analisi dei dati in psicologia*. Napoli: Edises.

Menard, S. (2001). *Applied Logistic Regression Analysis* (II Ed.) (Sage University Paper Series on Quantitative Applicatiosns in the Social Science). Thousand Oaks, CA: Sage.

Miceli, R. (2001). *Percorsi di ricerca e analisi dei dati*. Torino: Bollati Boringhieri editore S.r.l.

R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (URL http://www.R-project.org).

Appendice A

Alfabeto greco

Maiuscolo	Minuscolo		Maiuscolo	Minuscolo	
A	α	Alfa	N	ν	Nu
B	β	Beta	Ξ	ξ	Xi
Γ	γ	Gamma	O	ο	Omicron
Δ	δ	Delta	Π	π	Pi
E	ε	Ipsilon	P	ρ	Rho
Z	ζ	Zeta	Σ	σ	Sigma
H	η	Eta	T	τ	Tau
Θ	θ	Theta	Y	υ	Upsilon
I	ι	Iota	Φ	ϕ	Phi
K	κ	Kappa	X	χ	Chi
Λ	λ	Lambda	Ψ	ψ	Psi
M	μ	Mu	Ω	ω	Omega

Appendice B

Funzioni del software R utilizzate

help()

lm()

glm()

round()

summary()