Capitolo 5

I modelli adottati:
Regressione Logistica
ed
Event History Analysis

5.1 LA REGRESSIONE LOGISTICA

OBIETTIVI E CONTESTI APPLICATIVI

I modelli di regressione logistica nascono nell'ottica di spiegare la probabilità di possesso di un attributo, o di accadimento di un evento[1], in relazione ad una serie di possibili determinanti, di variabili esplicative: volendo fare un classico esempio, l'obiettivo potrebbe essere quello di individuare i principali fattori di rischio di una particolare malattia. Si tratta, come vedremo in seguito, di un caso particolare di analisi di regressione[2], dove la **variabile dipendente** è **dicotomica** (già in origine o resa tale ai fini dell'analisi) invece che quantitativa: convenzionalmente una modalità viene associata al valore 0 e l'altra al valore 1, per cui la sua distribuzione è ovviamente binomiale. La stima di Y nel continuo dovrà variare nell'intervallo [0,1] e non nei reali, assumendo il significato di probabilità che Y valga 1 o quindi, volendo, di ***rischio***.

I contesti applicativi sono molteplici, alcuni analoghi o comunque riconducibili a quelli della classica analisi di regressione, altri specifici di un'analisi di variabili dicotomiche. Tra i principali è importante ricordare:

- ✓ L' **analisi dei rischi** o la **ricerca di determinanti**: la prima di fatto vincolata ad ipotesi che restringono il campo alle variabili associate ad alte probabilità di riscontrare l'attributo o il fenomeno studiato; la seconda più orientata ad una ricerca esplorativa di fattori che in generale spieghino al meglio la variabilità della dipendente, non necessariamente correlati positivamente al fenomeno, fattori quindi sia di *rischio* che di *protezione*.
- ✓ La **stima di probabilità d'appartenenza** e la **discriminazione, l'assegnazione delle unità a specifici gruppi**[3]: la ricerca di una regola di classificazione basata su probabilità di riscontrare l'attributo discriminante e su fissati *valori soglia*.
- ✓ La **previsione**: la specificazione di un modello che non si limiti ad un'adeguata descrizione della variabile studiata nel dato contesto, bensì estendibile ad altri campioni con margini d'errore contenuti.

---

[1] Per ora consideriamo la probabilità indipendentemente dal tempo di attesa di accadimento dell'evento, che verrà introdotto nel paragrafo 5.2.

[2] Sono tutti modelli riconducibili alla classe dei ***modelli lineari generalizzati*** (McCullagh e Nelder _ [1989]), non condizionati alla natura di *Y* e al tipo di funzione (*funzione legame*) che la lega alla combinazione lineare delle esplicative, purché monotòna e differenziabile: $g(\mathrm{E}[Y|\underline{X}]) = \alpha + \underline{X}'\underline{\beta}$

[3] In tale contesto la regressione logistica può essere considerata una valida alternativa all'analisi discriminante.

IL MODELLO

Nel modello di regressione lineare la stima della variabile dipendente è ottenuta da una combinazione delle esplicative che teoricamente può assumere valori in tutto l'asse reale:

$$Y = \alpha + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon = \alpha + \underline{X}' \underline{\beta} + \varepsilon \qquad [4]\ \mathrm{E}[Y|\underline{X}] = \alpha + \underline{X}' \underline{\beta};$$

una scelta di questo tipo non sarebbe ovviamente opportuna nel nostro caso per modellare una probabilità,

$$\boxed{\lambda = \Pr(Y = 1|\underline{X}) = \mathrm{E}[Y|\underline{X}]},$$

che perde di significato al di fuori dell'intervallo [0,1].

Si ovvia a questo problema ricorrendo ad una trasformazione biunivoca di quest'ultima che estenda l'intervallo in questione a tutto il campo reale: sarà poi necessario ricorrere alla relativa inversa per ricondurre le stime a valori sensati nel campo della probabilità. Le alternative sono molteplici[5]; tra queste la più comune è l'utilizzo della ***funzione logistica***:

$$\boxed{g(\lambda) = \mathrm{logit}(\lambda) = \ln\left(\frac{\lambda}{1-\lambda}\right) = \alpha + \underline{X}' \underline{\beta}}\ [6].$$

Il rapporto tra la probabilità che si verifichi un dato evento e la probabilità che non si verifichi è chiamato ***odds*** e assume valori tra 0 [$\lambda = 0$] e $+\infty$ [$\lambda \to 1$]:

$$odds(\lambda) = \frac{\Pr(Y = 1|\underline{X})}{\Pr(Y = 0|\underline{X})} = \frac{\lambda}{1-\lambda} = e^{\alpha + \underline{X}'\underline{\beta}} = \exp\left(\alpha + \underline{X}' \underline{\beta}\right);$$

non rappresenta altro che il numero di *successi* per ogni *insuccesso* del fenomeno in esame.

La scelta di tale trasformazione per mettere in relazione la probabilità del fenomeno associato ad Y con le variabili predittive è sicuramente dovuta alla semplicità di interpretazione dei parametri stimati[7] ma anche al fatto che la sua inversa rispecchi l'andamento a "*sigmoide*" tipico della probabilità, con un lento avvicinarsi agli estremi 0 e 1:

$$\lambda = \frac{\exp\left(\alpha + \underline{X}' \underline{\beta}\right)}{1 + \exp\left(\alpha + \underline{X}' \underline{\beta}\right)} = \frac{1}{1 + \exp\left[-\left(\alpha + \underline{X}' \underline{\beta}\right)\right]}.$$

---

[4] *Funzione di regressione*

[5] Tutte funzioni continue e monotone crescenti. Tra quelle di maggior rilievo qui non prese in considerazione:

1. $g(\lambda) = \Phi^{-1}(\lambda)$ ***Funzione probit*** (*Normale inversa*)
2. $g(\lambda) = \ln(-\ln(1-\lambda))$ ***Funzione complementary log-log***

La *funzione logistica* e la *funzione probit* risultano approssimativamente lineari per valori di λ compresi tra 0.1 e 0.9; per bassi valori di λ la *logistica* e la *complementary log-log* presentano valori vicini a quelli di $\ln(\lambda)$; per λ molto vicino ad 1 invece la *complementary log-log* tende all'infinito molto più lentamente delle altre due.

[6] Non è presente alcun termine d'errore perché stiamo modellando la media condizionata (della dipendente).

[7] Vedi paragrafo 5.8.

5.2 ESTENSIONE ALL'EVENT HISTORY ANALYSIS [TEMPO DISCRETO]

IL RUOLO CRUCIALE DEL TEMPO

La regressione logistica in sé non prevede esplicitamente la variabile temporale tra i fattori che influenzano la probabilità del verificarsi di un fenomeno, ma non sono solo caratteristiche invariabili o eventi passati ad influenzare la storia e gli sviluppi futuri: anche il tempo di accadimento può avere un ruolo cruciale nel determinare la possibilità di sperimentare un evento; si parla in questo caso di "processo" al fine di evidenziarne la dimensione dinamica.

Diverse possono essere le cause di una possibile rilevante influenza:

- vi possono essere dei periodi in cui l'evento studiato non può oggettivamente aver luogo, in cui non vi è esposizione al rischio;
- durante la fase di rischio non è detto che questo sia costante e che non vi siano delle fasi di maggiore o minore propensione al realizzarsi del fenomeno;
- il rischio associato ad un evento può dipendere anche dal tempo di accadimento di eventi precedenti;
- il tempo storico o sociale in cui si colloca il processo può influenzare le sue caratteristiche ed il suo sviluppo (*effetti di periodo e di coorte*).

Nasce quindi spontaneo il bisogno di modelli che tengano conto della dimensione temporale e che impostino il fenomeno studiato come un processo caratterizzato dall'accadimento di particolari eventi; questi ultimi determinano l'uscita delle unità che lo sperimentano dallo stato d'interesse e quindi dall'osservazione, od eventualmente la transizione in un nuovo stato a sua volta preso in considerazione.

Gli strumenti descrittivi sui quali vengono strutturate le analisi non parametriche di processi di questo tipo sono la *tavola di eliminazione* e le sue estensioni alla *tavola ad eliminazione multipla* (dove l'uscita dallo stato di interesse può avvenire per più cause distinte) e alla *tavola multistato* (dove è prevista la transizione dei soggetti tra vari stati, con la possibilità di rientrare anche in quelli già sperimentati nel caso di eventi ripetibili).

A partire da questa impostazione e da queste forme rappresentative si sviluppa l'analisi parametrica, ovvero la costruzione di modelli statistici che permettano di rendere esplicita la dipendenza del processo da una serie di variabili esplicative secondo la logica della

regressione multipla: in questo contesto nascono negli anni '70 i modelli di ***survival analysis*** (analisi della sopravvivenza), come generalizzazione stocastica delle tavole di eliminazione. La variabile dipendente in tali modelli è la probabilità (o meglio il rischio) di accadimento dell'evento di interesse in t (nel caso in cui il tempo venga misurato in unità discrete) o tra t e $t+\Delta t$ (nel caso in cui venga considerato come variabile continua), mentre le esplicative possono essere tutti quei fattori (sia caratteristiche individuali o di contesto che eventi passati) che si ritiene possano influenzare l'oggetto di studio.
I modelli di *survival analysis* hanno il limite di non consentire l'analisi di processi *multiepisodio* (nel caso di eventi ripetibili, ovvero che possono essere sperimentati più volte nel corso della stessa storia individuale), *ad uscita multipla* e quindi *multistato*: da queste esigenze nascono le generalizzazioni nelle tecniche di ***event history analysis*** (**EHA**).

La scelta di modelli che prevedano una ripartizione del tempo in unità discrete può essere giustificata essenzialmente da due motivazioni:

- ✓ possono essere considerati approssimazioni di modelli a tempo continuo per convenienza o disponibilità di dati;
- ✓ il processo in esame ha di natura una scansione temporale discreta[8].

I modelli di *event history analysis* a tempo discreto non risulteranno altro che particolari estensioni (a seconda dei casi più o meno articolate) del modello di regressione logistica, in cui innanzitutto la dimensione temporale viene introdotta tra le variabili esplicative. La variabile dipendente rimane dicotomica (accadimento o meno dell'evento d'interesse), la stima del modello non subirà tanto variazioni formali quanto di significato, se non altro per il fatto che sarà funzione anche del tempo.

[8] Come nel caso degli studi universitari, la cui durata viene calcolata in numero di appelli di laurea trascorsi a partire dall'iscrizione.

IL MODELLO DI SURVIVAL ANALYSIS PER DATI AGGREGATI

L'organizzazione dei dati più prossima alla forma di una tavola di eliminazione è quella in cui questi vengono presentati in forma aggregata, una volta raggruppati gli individui secondo ogni diversa combinazione di variabili esplicative (compresa quella temporale[9]) che è venuta a presentarsi. Ogni osservazione registrata sarà a questo punto composta dalla variabile rappresentante la durata del tempo d'attesa presa in considerazione [t], da tutte le variabili che nella loro globalità esprimono le peculiarità di uno specifico gruppo e lo identificano univocamente [$\underline{x}_g$] nonché, sempre con riferimento all'unità temporale in questione, dal numero di individui con le caratteristiche date che risultano soggetti al rischio [n_{tg}] e quindi dal numero di eventi che tra essi si verificheranno [E_{tg}]; ognuna di queste osservazioni risulterà di fatto la realizzazione di un **processo binomiale** in cui avremo un certo numero di osservazioni (il numero di individui di un particolare gruppo in un dato momento), da cui un certo numero di successi (individui che sperimentano l'evento d'interesse ed escono dall'osservazione[10]) e per contro di insuccessi (individui che non sperimentano l'evento, la cui osservazione continua o viene censurata).

Il modello è il seguente:

$$\left[\left(E_{tg} / n_{tg}\right) \middle| \underline{X} = \underline{x}_g\right] \sim \textit{Binomiale}\left(n_{tg}, \lambda_{tg}\right)$$

$$\text{logit}\left(\lambda_{tg}\right) = \alpha_t + \underline{x}_g' \underline{\beta}$$

In un approccio di questo tipo, grazie ad una progressiva e costante ridefinizione dell'appartenenza dei singoli individui a specifici gruppi, la cui composizione non è vincolata a rimanere invariata, risultano immediate anche la gestione e la rappresentazione delle variabili *time-dependent*, ovvero di quelle caratteristiche individuali che possono variare con il passare del tempo; in tal caso aggiungeremo a $\underline{x}_g$ un indice temporale [$\underline{x}_{tg}$].

La dipendenza del rischio dal tempo di attesa dell'evento è qui modellata con i parametri α_t; questa scelta, non vincolandosi ad un particolare tipo di dipendenza, non richiede di formalizzare e rendere esplicito il tipo di legame che intercorre tra le due variabili, di supporne a priori un particolare andamento. Scelte opposte prevedono invece la possibilità di adottare modelli parametrici ipotizzando ad esempio un andamento lineare [$\alpha_t = \alpha_0 + \alpha_1 t$], quadratico [$\alpha_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2$] o log-lineare [$\alpha_t = \alpha_0 + \alpha_1 \ln(t)$].

---

[9] Ogni individuo verrà incluso in tutti gli aggregati per età di cui ha fatto parte nel corso della sua vita.

[10] In modelli *multiepisodio* o *multistato*, invece, l'individuo non esce dall'osservazione a seguito del verificarsi di un evento, a meno che non si tratti dell'entrata in uno strato assorbente, ovvero da cui non può più uscire.

IL MODELLO DI SURVIVAL ANALYSIS PER DATI INDIVIDUALI

Diversa opportunità è quella di strutturare i dati tenendo distinto ogni individuo dagli altri [***person-period data set***]: ogni record rappresenterà l'osservazione in una particolare unità temporale di uno specifico soggetto, il quale potrà aver o meno sperimentato l'evento, ed equivarrà alla realizzazione di un **processo bernoulliano**; ognuno di questi record sarà composto dalla variabile *t*, da tutte le variabili esplicative e dalla variabile che definirà il verificarsi o meno dell'evento di interesse [Y_{ti}]. Il modello è il seguente:

$$\left(Y_{ti}\middle|\underline{X}=\underline{x}_i\right)\sim Bernoulli\left(\lambda_{ti}\right)$$

$$\text{logit}\left(\lambda_{ti}\right)=\alpha_t+\underline{x}_i'\,\underline{\beta}$$

Il generico individuo *i* contribuirà all'analisi con t_i osservazioni: t_i*-1* insuccessi e *1* successo nel caso in cui in t_i sperimenti l'evento [$y_{t_i i}=1$], t_i insuccessi nel caso in cui in t_i la sua storia risulti censurata.

Un modello di questo tipo risulta sicuramente più flessibile del precedente ed ammette variabili esplicative anche di tipo continuo; rimane agevole gestire variabili *time-dependent*, potendo esse assumere anche in questo caso valori differenti in ogni periodo [$\underline{x}_{ti}$].

Il principale limite di quest'approccio potrebbe consistere invece nella vasta dimensione del data set su cui è impostato e di conseguenza in lunghi tempi di elaborazione che ciò potrebbe comportare.

Un esempio di dati organizzati in tal modo è riportato nella tabella seguente, costruita per l'analisi dell'evento "nascita del primo figlio" e considerando come evento origine il "compimento dei 15 anni":

Donna (i)	Età (t)	Istruzione (x_1)	Matrimonio (x_2)	Anticoncezionali (x_3)	Evento (y_{ti})
1	15	1	0	0	0
1	16	1	0	0	1
2	15	0	0	0	0
2	16	0	1	1	0
2	17	0	1	0	0
2	18	0	1	0	1
3	15	1	0	0	0
3	16	1	0	0	0
3	17	1	0	0	0
3	18	1	0	0	0
3	19	2	0	1	0
3	20	2	0	1	0
3	21	2	0	1	0
3	22	2	1	1	0
…	…	…	…	…	…

La prima colonna riporta il numero identificativo dell'individuo al quale l'osservazione appartiene; la seconda indica l'unità temporale (o durata del tempo di attesa) a cui essa si riferisce, che qui, visto l'evento origine, può essere rappresentata direttamente con l'età della donna; in quelle che seguono, dalla terza alla quinta, trovano spazio le altre esplicative, nel caso considerato tutte *time-dependent*; l'ultima, infine, è associata al verificarsi o meno dell'evento di interesse. Volendo quindi leggere i dati a disposizione, osserviamo che la prima donna ha un figlio a 16 anni, prima del matrimonio, la seconda ha il primo figlio due anni dopo il matrimonio, a seguito del quale ha interrotto l'uso di anticoncezionali, e la terza, che raggiunge un alto grado d'istruzione e si sposa relativamente tardi, a 22 anni fa uso di anticoncezionali e non ha ancora avuto figli.

5.3 L'IPOTESI DI RISCHI PROPORZIONALI (IN TERMINI DI ODDS)

L'IPOTESI IMPLICITA NEL MODELLO

Consideriamo il modello di survival analysis per dati individuali ed in particolare la relazione tra il rischio e le variabili esplicative già presentata nel precedente paragrafo, nel caso di variabili esplicative costanti nel tempo:

$$\text{logit}(\lambda_{ti}) = \ln\left(\frac{\lambda_{ti}}{1-\lambda_{ti}}\right) = \ln[odds(\lambda_{ti})] = \ln\left[\frac{\Pr(Y_{ti}=1|\underline{X}=\underline{x}_i)}{\Pr(Y_{ti}=0|\underline{X}=\underline{x}_i)}\right] = \alpha_t + \underline{x}_i'\,\underline{\beta}\,.$$

Il parametro α_t è indipendente dalle caratteristiche individuali introdotte nel modello: potrebbe quindi essere considerato una "base" per il rischio a cui è soggetta nel tempo *t* ogni entità della popolazione, il che può essere espresso ponendo

$$\alpha_t = \text{logit}(\lambda_{t0}),$$

da cui si ottiene:

$$\text{logit}(\lambda_{ti}) = \text{logit}(\lambda_{t0}) + \underline{x}_i'\,\underline{\beta}$$

$$\ln[odds(\lambda_{ti})] = \ln[odds(\lambda_{t0})] + \underline{x}_i'\,\underline{\beta}$$

$$odds(\lambda_{ti}) = odds(\lambda_{t0})\exp\left(\underline{x}_i'\,\underline{\beta}\right).$$

L'*odds* dell'individuo *i* al tempo *t* è scomposto quindi in due fattori, uno che dipende esclusivamente dalle caratteristiche individuali [$\exp\left(\underline{x}_i'\,\underline{\beta}\right)$], l'altro dalla dimensione temporale; quest'ultimo [$odds(\lambda_{t0})$] rappresenta il ***rischio di base*** relativo all'unità temporale *t*.

Specificare il modello nella forma proposta corrisponde di fatto ad assumere l'ipotesi che particolari caratteristiche individuali diano luogo, qualunque sia l'unità temporale considerata, a medesime variazioni percentuali del *rischio di base* ad essa relativo. In altre parole, qualunque sia l'unità temporale in esame, gli odds (volendo si potrebbe anche parlare di "rischi espressi in termini di odds") di diversi individui risultano **proporzionali** ad un rischio di base (e quindi tra loro) **secondo un fattore indipendente dal tempo**.

Supponiamo ad esempio di studiare l'evento "nascita del primo figlio" in relazione ad un'unica variabile esplicativa "matrimonio" [*X*], di natura dicotomica, pari a 1 nel caso in cui la donna sia sposata e a 0 nel caso sia nubile.

$$odds(\lambda_{ti}|X=0)=odds(\lambda_{t0})$$

$$odds(\lambda_{ti}|X=1)=odds(\lambda_{t0})\exp(\beta)$$

Il rapporto tra l'odds relativo ad una data categoria di individui e quello invece relativo alla categoria di riferimento prende il nome di *rapporto crociato* o ***odds ratio*** (OR)[11]; in questo caso, considerato il gruppo delle donne sposate,

$$\mathrm{OR}(X=1,t)=\frac{odds(\lambda_{ti}|X=1)}{odds(\lambda_{ti}|X=0)}=\frac{\Pr(Y_{ti}=1|X=1)}{\Pr(Y_{ti}=0|X=1)}\cdot\frac{\Pr(Y_{ti}=0|X=0)}{\Pr(Y_{ti}=1|X=0)}=\frac{odds(\lambda_{t0})\exp(\beta)}{odds(\lambda_{t0})}=$$

$$=\exp(\beta)=\mathrm{OR}(X=1).$$

Nel caso in cui $\exp(\beta)=1$ il rischio di sperimentare l'evento associato ad una donna sposata è lo stesso rispetto a quello associato ad una nubile, nel caso $\exp(\beta)<1$ il rischio per la prima sarebbe minore rispetto a quello per la seconda, viceversa se $\exp(\beta)>1$. Ciò che potrebbe destare perplessità a livello d'interpretazione è il fatto che comunque, in qualunque di questi casi, le donne sposate sarebbero costantemente sottoposte ad un rischio di "nascita del primo figlio" pari a $\exp(\beta)$ volte quello a cui sono sottoposte invece le nubili, indipendentemente dall'età considerata.

[11] Se la variabile *X* è continua, l'*odds ratio* solitamente considerato misura l'effetto sul rischio di una variazione unitaria di *X*. Vedi paragrafo 5.8 in "*Variabile esplicativa quantitativa*".

IL SUPERAMENTO DI TALE LIMITE

L'ipotesi di rischi proporzionali può essere facilmente superata introducendo tra le esplicative un fattore di dipendenza dal tempo, ad esempio la ***variabile d'interazione*** $t\underline{X}$.

Il modello assumerebbe la forma

$$\text{logit}(\lambda_{ti}) = \alpha_t + \underline{x}_i' \underline{\beta} + t\underline{x}_i' \underline{\gamma} = \alpha_t + \underline{x}_i' (\underline{\beta} + \underline{\gamma} \cdot t),$$

da cui

$$odds(\lambda_{ti}) = odds(\lambda_{t0}) \exp\left(\underline{x}_i' \underline{\beta} + t\underline{x}_i' \underline{\gamma} \right).$$

A questo punto, nell'esempio precedente (ora passato da una a due variabili esplicative) risulterebbe

$$\text{OR}(X = 1, t) = \frac{odds(\lambda_{t0}) \exp(\beta + \gamma \cdot t)}{odds(\lambda_{t0})} = \exp(\beta + \gamma \cdot t),$$

questa volta variabile in funzione dell'età considerata.

5.4 ANALISI PRELIMINARI PER LE VARIABILI ESPLICATIVE

Per scegliere in modo opportuno le potenziali variabili esplicative da introdurre nel modello[12], è fondamentale definire con chiarezza quali siano **gli scopi primari dell'analisi**: nel caso in cui quest'ultima sia prevalentemente orientata alla ricerca dei fattori determinanti (di rischio o di protezione) di un evento, come nel nostro caso, si presterà particolare attenzione alla natura delle variabili e si svilupperà lo studio delle loro interazioni; nel caso in cui lo scopo sia invece la costruzione di un algoritmo di classificazione, efficiente ma soprattutto estendibile anche ad altri campioni, si cercherà di isolare un gruppo di regressori più ristretto ma con forte potere discriminante.

I principali **accorgimenti** da avere nella fase di selezione delle esplicative sono i seguenti:

- le variabili selezionate dovranno rappresentare plausibili fattori determinanti della dipendente: andrà analizzata, come vedremo nel seguito del paragrafo, la significatività delle relazioni intercorrenti tra quest'ultima ed ogni singola variabile candidata ad entrare nel modello di regressione;
- dovranno essere prese in considerazione combinazioni (o perlomeno coppie) di variabili che potrebbero rivelarsi determinanti significative nonostante non lo siano gli apporti delle stesse considerate singolarmente;
- di tutte le variabili andranno specificate la natura, la scala e l'unità di misura;
- si dovrà verificare che nessuna delle modalità di una qualunque variabile presenti frequenza nulla perché ciò comporterebbe, come si vedrà nel paragrafo 5.5, di ottenere come stime dei parametri valori infiniti;
- per quanto riguarda le variabili quantitative, dovrà risultare accettabile l'ipotesi di dipendenza lineare tra $\text{logit}(\lambda)$ ed esse, che altrimenti dovranno essere trasformate in variabili ordinali;
- dev'essere rispettata la condizione di non collinearità tra i regressori;
- il numero di variabili prese in considerazione dovrà essere adeguato al numero di osservazioni a disposizione: una numerosità campionaria esigua rispetto al numero di esplicative potrebbe essere causa, come vedremo nel paragrafo successivo, di inesistenza o instabilità delle stime di massima verosimiglianza.

[12] Qui si segue un'ottica univariata, per un'ulteriore selezione in ottica multivariata si rinvia al paragrafo 5.7.

RILEVANZA DI UNA VARIABILE ESPLICATIVA DICOTOMICA

L'analisi della significatività della relazione intercorrente tra la variabile dipendente ed una variabile esplicativa dicotomica è costruita sulla base di una *tabella di contingenza* 2×2, detta ***tabella tetracorica***, nella quale vengono rappresentate le frequenze congiunte (assolute) delle due variabili.

		VARIABILE DIPENDENTE (Y)		
		1 (*evento sperimentato*)	**0 (*evento non sperimentato*)**	
VARIABILE ESPLICATIVA (X)	**1 (*caratteristica presente*)**	a	b	$a+b$
	0 (*caratteristica assente*)	c	d	$c+d$
		$a+c$	$b+d$	n

Vi sono più misure, calcolabili a partire da una tabella di questo tipo, per valutare la dipendenza tra le due variabili; dipenderà dalla natura e dall'obbiettivo dell'analisi quale di queste di volta in volta il ricercatore prediligerà per impostarvi il suo criterio di selezione.

1. Il ***coefficiente di correlazione di Bravais-Pearson***

   $$\phi = \frac{\hat{\text{cov}}(X,Y)}{\sqrt{\hat{\text{var}}(X)\cdot\hat{\text{var}}(Y)}} = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$ [13]

   Il coefficiente di correlazione misura l'intensità e la direzione della relazione (lineare) tra X e Y; assume valori nell'intervallo [-1,1]:

   - ✓ $\phi = 0$ quando X e Y sono tra loro *incorrelate*;
   - ✓ $\phi = 1$ quando tra X e Y si presenta una *perfetta correlazione positiva*;
   - ✓ $\phi = -1$ quando tra X e Y si presenta una *perfetta correlazione negativa*.

   Tale statistica viene presa in considerazione al fine di ottenere un primo eventuale segnale di una qualche forma di dipendenza dell'evento di interesse dal fattore considerato.

[13] Ricordando che $\hat{\text{var}}(X) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

2. Il ***rischio relativo***

$$R_{1/0} = \frac{\Pr(Y=1|X=1)}{\Pr(Y=1|X=0)} = \frac{a}{a+b} \Big/ \frac{c}{c+d} = \frac{a(c+d)}{c(a+b)}$$

È il rapporto tra il rischio associato ai soggetti in cui si riscontra la caratteristica presa in considerazione ed il rischio associato ai soggetti in cui invece tale caratteristica non si riscontra. Non è mai negativo ed assume valore:

- ✓ 1 nel caso in cui il possesso o meno dell'attributo X considerato non abbia il minimo effetto sul rischio di sperimentare l'evento;
- ✓ tanto più grande di 1 quanto più X è determinante come *fattore di rischio* per Y, in altre parole quanto più la presenza dell'attributo X è associata ad elevate probabilità di osservare l'evento;
- ✓ tanto più vicino a 0 quanto più X è determinante come *fattore di protezione* per Y.

Se studiamo un evento raro, $\Pr(Y=0|X=1) \cong 1$ e $\Pr(Y=0|X=0) \cong 1$; in tal caso il rischio relativo può essere stimato con il cosiddetto ***rapporto crociato*** (pari all'*odds ratio*)

$$\frac{a \cdot d}{b \cdot c} = \left(\frac{a}{a+b} \cdot \frac{d}{c+d} \right) \Big/ \left(\frac{b}{a+b} \cdot \frac{c}{c+d} \right) = \frac{\Pr(Y=1|X=1) \cdot \Pr(Y=0|X=0)}{\Pr(Y=0|X=1) \cdot \Pr(Y=1|X=0)} \cong$$

$$\cong \frac{\Pr(Y=1|X=1)}{\Pr(Y=1|X=0)}.$$

3. La ***sensibilità***

Misura la proporzione di soggetti, tra quelli che sperimentano l'evento, in cui si riscontra il fattore di rischio considerato:

$$Sen = \Pr(X=1|Y=1) = \frac{a}{a+c}.$$

4. La ***specificità***

Misura la proporzione di soggetti, tra quelli che non sperimentano l'evento, in cui non si riscontra effettivamente il fattore di rischio considerato:

$$Spe = \Pr(X=0|Y=0) = \frac{d}{b+d}.$$

RILEVANZA DI UNA VARIABILE ESPLICATIVA NOMINALE OD ORDINALE

Anche nel caso di variabili nominali od ordinali, l'analisi della significatività della relazione intercorrente tra esse e la dipendente può essere impostata su diversi indici e relativi test.

1. Il ***coefficiente χ^2 di Pearson***

   Una misura della dipendenza tra due variabili nominali od ordinali, rispettivamente con *k* e *h* modalità, è ottenibile con la seguente statistica:

   $$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{\left(n_{ij} - n_{ij}^{*}\right)^2}{n_{ij}^{*}} \underset{\text{sotto } H_0}{\sim} \chi^2_{kh-1},$$

   dove n_{ij} è la frequenza assoluta dell'osservazione congiunta dell'*i*-esima modalità della prima variabile e della *j*-esima modalità della seconda, mentre n_{ij}^{*}, calcolata come

   $$n_{ij}^{*} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n},$$

   è la corrispondente frequenza teorica che ci si aspetterebbe di osservare nell'ipotesi nulla [H_0] di indipendenza tra le due variabili. Nel caso di perfetta indipendenza, le frequenze osservate coinciderebbero con tali frequenze teoriche e l'indice risulterebbe nullo.

   Nel nostro caso, andando a verificare l'indipendenza di *Y* (con *2* sole modalità) da *X* (con *k* modalità), il coefficiente assume la forma

   $$\chi^2 = \sum_{i=1}^{k}\sum_{j=0}^{1}\frac{\left(n_{ij} - n_{ij}^{*}\right)^2}{n_{ij}^{*}} \underset{\text{sotto } H_0}{\sim} \chi^2_{2k-1}.$$

2. Il ***test di Mann-Whitney***

   Utilizzabile nel caso di variabile esplicativa ordinale, il test *U* di Mann-Whitney è un *test basato sui ranghi*, che qui non trattiamo nel dettaglio[14].

[14] Vedi "Piccolo Domenico _ ***Statistica*** _ il Mulino _ [1998]".

RILEVANZA DI UNA VARIABILE ESPLICATIVA QUANTITATIVA

Per lo studio della significatività di una variabile quantitativa, è possibile seguire uno dei due differenti approcci qui di seguito proposti:

1. L'applicazione di un modello di regressione logistica univariata

   Per studiare il legame tra Y e una variabile quantitativa, si può applicare un modello di regressione logistica univariato che preveda come unici predittori l'intercetta e l'esplicativa in questione. Lo studio della significatività si riconduce così ad una verifica d'ipotesi sulla significatività di un singolo parametro nella regressione logistica, trattata nel paragrafo 5.6.

2. Il ***test t di Student bilaterale a due campioni***

   Si può pensare di analizzare la rilevanza della variabile valutando se la sua distribuzione nel gruppo che sperimenta l'evento (di numerosità $n_{Y=1}$) sia significativamente diversa da quella osservata nel gruppo che non lo sperimenta (di numerosità $n_{Y=0}$): in tal caso le due variabili non sarebbero indipendenti.

   Uno dei modi possibili di condurre questo tipo di verifica è il confronto tra le medie di *X* nei due diversi gruppi: **diversità delle medie** implicherebbe diversa distribuzione.

   Nelle ipotesi di indipendenza delle osservazioni, di normalità della variabile considerata e di omoschedasticità, ovvero uguale varianza di X nei due gruppi, un test utilizzabile è il t di student bilaterale a due campioni, che qui, data

   $$H_0 : \mathrm{E}\left[X \middle| Y=1\right] = \mathrm{E}\left[X \middle| Y=0\right],$$

   assume la forma

   $$t = \frac{\bar{x}_{Y=1} - \bar{x}_{Y=0}}{\sqrt{\mathrm{var}(X)} \cdot \sqrt{1/n_{Y=1} + 1/n_{Y=0}}} \underset{\text{sotto } H_0}{\sim} t_{n-2} \,.$$

   $$t_{\text{oss}} = \frac{\bar{x}_{Y=1} - \bar{x}_{Y=0}}{\sqrt{\hat{\mathrm{var}}(X)} \cdot \sqrt{1/n_{Y=1} + 1/n_{Y=0}}} =$$

   $$= \frac{\bar{x}_{Y=1} - \bar{x}_{Y=0}}{\sqrt{\dfrac{(n_{Y=1}-1)\,\hat{\mathrm{var}}\left(X \middle| Y=1\right) + (n_{Y=0}-1)\,\hat{\mathrm{var}}\left(X \middle| Y=0\right)}{n-2}} \cdot \sqrt{1/n_{Y=1} + 1/n_{Y=0}}}$$

LA RELAZIONE LINEARE TRA LOGIT(λ_{ti}) E UNA VARIABILE QUANTITATIVA

Uno dei presupposti di base del modello di regressione logistica è la dipendenza lineare tra $\text{logit}(\lambda)$ e le variabili esplicative quantitative; per poter verificare tale ipotesi, si può adottare la seguente procedura:

- Si suddivide l'insieme di definizione della variabile quantitativa in classi di ampiezza regolare;
- per ogni classe si calcola la probabilità associata al verificarsi dell'evento;
- si calcola il logit delle probabilità trovate;
- si rappresenta su piano cartesiano il valore del logit in funzione del valore centrale delle classi.

Qualora non risulti rispettata l'ipotesi di linearità, la variabile quantitativa dev'essere trasformata in ordinale, identificando i valori in corrispondenza dei quali il legame si modifica e scegliendoli come estremi per le classi: ogni classe sarà associata ad un diverso effetto della variabile sul rischio di sperimentare l'evento.

5.5 LA STIMA DEI PARAMETRI CON DATI INDIVIDUALI[15]

Il processo di stima viene impostato sulla base del ***metodo di massima verosimiglianza***, secondo il quale vengono scelti come stime dei parametri i valori che massimizzano la probabilità che si aveva a priori di ottenere l'insieme di dati effettivamente osservato.

CALCOLO DELLE STIME NEL MODELLO CLASSICO DI REGRESSIONE LOGISTICA

Nel *modello classico di regressione logistica*, in cui non si fa alcuna differenza tra eventuali diversi tempi d'osservazione, il modello per il generico individuo si presenta come segue:

$$y_i = \mathrm{E}\left[\mathrm{Y_i}\middle|\underline{X} = \underline{x}_i\right] + \varepsilon_i = \lambda_i + \varepsilon_i = \frac{\exp\left(\alpha + \underline{x}_i' \underline{\beta}\right)}{1 + \exp\left(\alpha + \underline{x}_i' \underline{\beta}\right)} + \varepsilon_i .$$

Poiché la singola Y_i è dicotomica e può assumere i valori 0 e 1, ha una distribuzione bernoulliana di media $\mathrm{E}\left[\mathrm{Y_i}\middle|\underline{X} = \underline{x}_i\right] = \lambda_i$ [16]. Il contributo del singolo individuo alla funzione di verosimiglianza è quindi λ_i nel caso in cui egli sperimenti l'evento [$y_i = 1$], *1-* λ_i in caso contrario [$y_i = 0$]; la sua funzione di probabilità è di conseguenza

$$\mathrm{L}_i = f\left(y_i\middle|\underline{X} = \underline{x}_i; \alpha, \underline{\beta}\right) = \lambda_i^{\,y_i}\left(1 - \lambda_i\right)^{1-y_i} .$$

Assumendo ragionevolmente che ogni osservazione sia indipendente dalle altre, la verosimiglianza complessiva, ovvero la probabilità associata al campione osservato, non è altro che il prodotto dei contributi delle singole unità che compongono quest'ultimo:

$$\mathrm{L} = \prod_{i=1}^{n} \mathrm{L}_i .$$

Per ottenere la *stima di massima verosimiglianza* bisogna determinare l'α e il vettore $\underline{\beta}$ che massimizzano tale quantità, il che equivale a massimizzarne il logaritmo (trasformazione monotòna crescente), detto *log-verosimiglianza*:

$$\mathrm{l} = \ln(\mathrm{L}) = \sum_{i=1}^{n}\left[y_i \ln\lambda_i + (1 - y_i)\ln(1 - \lambda_i)\right] = \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\lambda_i}{1 - \lambda_i}\right) + \ln(1 - \lambda_i)\right].$$

---

[15] Le lettere maiuscole rappresentano le variabili casuali, quelle minuscole realizzazioni di esse, ovvero valori da queste assunti.

[16] Il termine d'errore è una variabile casuale dicotomica:

$$\varepsilon_i = \begin{cases} -\lambda_i & \text{con probabilità } 1 - \lambda_i \text{ ovvero se } y_i = 0 \\ 1 - \lambda_i & \text{con probabilità } \lambda_i \text{ ovvero se } y_i = 1 \end{cases} \qquad \mathrm{E}[\varepsilon_i] = -\lambda_i(1 - \lambda_i) + (1 - \lambda_i)\lambda_i = 0$$

Ponendo uguali a 0 le derivate parziali rispetto ad α e ad ogni componente di $\underline{\beta}$, si ottengono le cosiddette *equazioni di verosimiglianza*; nel nostro caso esse non saranno lineari nei parametri e la loro risoluzione richiederà l'applicazione di metodi iterativi, comunque implementati in molti pacchetti software statistici.

CALCOLO DELLE STIME NEL MODELLO DI SURVIVAL ANALYSIS

Nel momento in cui introduciamo la dimensione temporale, il modello per il generico individuo si trasforma in

$$y_{ti} = \mathrm{E}\left[\mathrm{Y}_{ti} \middle| \underline{X} = \underline{x}_i\right] + \varepsilon_{ti} = \lambda_{ti} + \varepsilon_{ti} = \frac{\exp\left(\alpha_t + \underline{x}_i' \underline{\beta}\right)}{1 + \exp\left(\alpha_t + \underline{x}_i' \underline{\beta}\right)} + \varepsilon_{ti}\,.$$

La funzione di probabilità del singolo individuo, se ipotizzassimo che egli non abbia ancora sperimentato l'evento, risulterebbe

$$f\left(y_{ti} \middle| \underline{X} = \underline{x}_i, Y_{1i} = \ldots = Y_{(t-1)i} = 0; \alpha_t, \underline{\beta}\right) = \lambda_{ti}^{\,y_{ti}} \left(1 - \lambda_{ti}\right)^{1-y_{ti}},$$

con $t \le t_i$ nel momento in cui con t_i indicassimo l'unità temporale in cui l'*i*-esimo individuo esce dall'osservazione per aver sperimentato l'evento o perché soggetto a censura[17].

La realizzazione al tempo *t* è di fatto vincolata alla mancata realizzazione precedente dell'evento, per cui la verosimiglianza associata al vettore delle osservazioni relative al generico individuo fino all'unità *t* assume necessariamente la forma:

$$\mathrm{L}_{ti} = f\left(y_{1i}, \ldots, y_{ti} \middle| \underline{X} = \underline{x}_i; \alpha_t, \underline{\beta}\right) = \left[\Pr(T_i > t)\right]^{1-y_{ti}} \left[\Pr(T_i = t)\right]^{y_{ti}} = \left[\prod_{j=1}^{t}\left(1 - \lambda_{ji}\right)\right]^{1-y_{ti}} \left[\lambda_{ti}\prod_{j=1}^{t-1}\left(1 - \lambda_{ji}\right)\right]^{y_{ti}} =$$

$$= \left[\prod_{j=1}^{t}\left(1 - \lambda_{ji}\right)\right]^{1-y_{ti}} \left[\frac{\lambda_{ti}}{1 - \lambda_{ti}}\prod_{j=1}^{t}\left(1 - \lambda_{ji}\right)\right]^{y_{ti}} = \left(\frac{\lambda_{ti}}{1 - \lambda_{ti}}\right)^{y_{ti}} \left[\prod_{j=1}^{t}\left(1 - \lambda_{ji}\right)\right]^{18} \qquad [\text{con } t \le t_i\,].$$

---

[17] Come già notato in passato, nel caso di *processi multiepisodio,* ovvero di eventi ripetibili, o *multistato*, il verificarsi dell'evento non implicherebbe la fine dell'osservazione, condizionata esclusivamente alla censura o all'entrata in uno stato assorbente.

[18] Nel caso di *processi multiepisodio*, i fattori relativi al passato non avrebbero necessariamente forma $(1-\lambda_{ji})$, in quanto la realizzazione presente dell'evento in esame non sarebbe vincolata alla mancata realizzazione precedente.
Potrebbe inoltre risultare necessario, in modelli di questo tipo, introdurre dei fattori di dipendenza del rischio dall'insieme di eventuali realizzazioni precedenti dell'evento in questione, ad esempio costruendo una variabile esplicativa che rappresenti l'ordine dell'episodio. In alternativa, con vantaggi in termini di semplicità formale e chiarezza, un differente approccio per studiare *eventi ripetuti* è il cosiddetto ***approccio piecemeal***, che consiste nell'analizzare gli episodi distintamente per ordine.

Avendo posto t_i come unità temporale in cui l'*i*-esimo individuo esce dall'osservazione, la verosimiglianza ad egli associata è

$$\mathrm{L}_i = \mathrm{L}_{t_i i} = \left(\frac{\lambda_{t_i i}}{1-\lambda_{t_i i}} \right)^{y_{t_i i}} \left[\prod_{j=1}^{t_i} \left(1-\lambda_{ji}\right) \right],$$

da cui la seguente *verosimiglianza* complessiva e relativa *log-verosimiglianza*:

$$\mathrm{L} = \prod_{i=1}^{n} \mathrm{L}_i = \prod_{i=1}^{n} \left\{ \left(\frac{\lambda_{t_i i}}{1-\lambda_{t_i i}} \right)^{y_{t_i i}} \left[\prod_{j=1}^{t_i} \left(1-\lambda_{ji}\right) \right] \right\}.$$

$$1 = \ln(\mathrm{L}) = \sum_{i=1}^{n} \left[y_{t_i i} \ln\left(\frac{\lambda_{t_i i}}{1-\lambda_{t_i i}} \right) + \sum_{j=1}^{t_i} \ln\left(1-\lambda_{ji}\right) \right] = \sum_{i=1}^{n} \sum_{j=1}^{t_i} \left[y_{ji} \ln\left(\frac{\lambda_{ji}}{1-\lambda_{ji}} \right) + \ln\left(1-\lambda_{ji}\right) \right].$$

Una volta stimati i parametri, è ovviamente possibile ricavare da essi le stime delle probabilità condizionate di realizzazione dell'evento di interesse per ciascun individuo in ognuna delle unità temporali:

$$\hat{y}_{ti} = \hat{\lambda}_{ti} = \frac{\exp\left(\hat{\alpha}_t + \underline{x}_i' \, \underline{\hat{\beta}} \right)}{1+\exp\left(\hat{\alpha}_t + \underline{x}_i' \, \underline{\hat{\beta}} \right)} = \frac{1}{1+\exp\left[-\left(\hat{\alpha}_t + \underline{x}_i' \, \underline{\hat{\beta}} \right) \right]};$$

si tratta di probabilità teoriche date dal modello e generalmente non coincideranno con quelle osservate, calcolate sulla base della tavola di eliminazione.

PROBLEMI DI CALCOLO DELLE STIME

In alcuni casi, particolari caratteristiche dei dati in esame possono causare problemi di natura computazionale che compromettono l'accuratezza dei risultati.

Sintomi di problemi di questo tipo possono essere:

- ✓ errori standard dei parametri troppo elevati rispetto al valore a cui si riferiscono;
- ✓ stime che aumentano rapidamente con il numero di iterazioni necessarie al calcolo delle stesse.

Le **cause** possono essere di diverso tipo:

- La **frequenza nulla di una modalità**: questo fatto determina un rapporto crociato pari a 0 o a $+\infty$ e di conseguenza, come vedremo nel paragrafo 5.8, relative stime pari a $-\infty$ e $+\infty$. In tal caso si può pensare di inglobare la modalità che presenta frequenza nulla in un'altra o alternativamente di aggiungere 0.5 alla frequenza di ogni diversa modalità.
- Una **perfetta discriminazione**, sulla base delle variabili esplicative, del gruppo di unità per cui Y=0 da quello per cui Y=1, che si traduce in una funzione di verosimiglianza monotòna e di conseguenza nell'inesistenza di un stima di massima verosimiglianza. L'origine del problema potrebbe consistere ad esempio in un'esigua numerosità campionaria rispetto al numero di variabili esplicative considerate, cosa che andrebbe considerata in fase di selezione di queste.
- Situazioni di **collinearità** delle variabili esplicative, nel momento in cui una è esprimibile come combinazione lineare delle altre; anche questo problema è ovviabile in fase di selezione delle variabili.

5.6 LA VERIFICA DEL MODELLO

Un buon modello, per poter esser considerato tale, dovrebbe soddisfare tre requisiti che potremmo considerare fondamentali:

- ✓ ***Bontà dell'adattamento***: le stime ottenute devono essere il più possibile "vicine" ai valori reali osservati (qui in termini di λ più che di *Y*); in altre parole, il modello dovrebbe garantire un errore di stima il più piccolo possibile.
- ✓ ***Parsimonia***: il numero di variabili esplicative dovrebbe essere possibilmente esiguo, per far sì che il modello risulti il meno "costoso" possibile, più stabile e facilmente interpretabile.
- ✓ ***Capacità previsiva***: il modello, per quanto valido rispetto alle osservazioni su cui viene stimato, dovrebbe risultare estensibile ad altri insiemi di dati, ovvero godere di possibili generalizzazioni.

VERIFICA DELLA SIGNIFICATIVITÀ DEI SINGOLI PARAMETRI

Accettare la *significatività* di un singolo parametro equivale a rifiutare l'ipotesi che esso sia nullo. Facendo riferimento alla *normalità delle stime di massima verosimiglianza*, una *statistica test* adeguata a tale verifica d'ipotesi può essere

$$z = \frac{\hat{\beta}_k}{\sigma\left(\hat{\beta}_k\right)} \underset{\text{sotto } H_0}{\sim} N(0;1),$$

dove $\sigma\left(\hat{\beta}_k\right)$ è lo *scarto quadratico medio* dello stimatore di β_k.

Nel caso di *ipotesi alternativa* bilaterale,

$$H_0 : \beta_k = 0 \qquad H_1 : \beta_k \neq 0,$$

usualmente si ricorre al quadrato della precedente, ottenendo il cosiddetto ***test di Wald***:

$$w = \left[\frac{\hat{\beta}_k}{\sigma\left(\hat{\beta}_k\right)}\right]^2 \underset{\text{sotto } H_0}{\sim} \chi_1^2.$$

Il valore osservato per tale statistica risulterà

$$w_{\text{oss}} = \left[\frac{\hat{\beta}_k}{\hat{\sigma}\left(\hat{\beta}_k\right)}\right]^2 = \left[\frac{\hat{\beta}_k}{SE\left(\hat{\beta}_k\right)}\right]^2,$$

dove $SE\left(\hat{\beta}_k\right)$ indica lo *standard error* del parametro, ovvero una stima dello scarto quadratico medio del relativo stimatore.

L'accettazione o meno di H_0 dipende allora dal cosiddetto *p-value* (*livello di significatività osservato*), pari alla probabilità di osservare dei valori di *W* più grandi di quello ottenuto:

$$p-value = \Pr\{W \geq w_{oss}\};$$

valori man mano più piccoli di tale quantità corrispondono a situazioni sempre più difficilmente riconducibili alla distribuzione attesa sotto H_0.

Scelta quindi un'opportuna soglia (*livello di significatività* δ, ad esempio il 5%) per la probabilità che, rifiutata H_0, questa risultasse invece vera (*errore di primo tipo*), ne deriva il seguente criterio di scelta:

$p-value < \delta \Rightarrow H_0$ è rifiutata [il parametro viene considerato significativo];

$p-value > \delta \Rightarrow H_0$ è accettata [il parametro viene considerato non significativo].

VERIFICA DELLA SIGNIFICATIVITÀ DI GRUPPI DI PARAMETRI

La generalizzazione della verifica d'ipotesi appena considerata è la verifica della significatività di un gruppo di parametri del modello stimato, per la quale viene utilizzato il ***test del rapporto di verosimiglianza***.

Vengono confrontati due modelli, detti ***modelli annidati*** *o nidificati* (*nested models*), uno dei quali (*modello ridotto*) presenta solo un sottoinsieme delle variabili esplicative contenute nell'altro (*modello completo*). L'ipotesi nulla è l'equivalenza dei due modelli, ovvero l'ipotesi che i parametri relativi a tutte le variabili non comprese nel modello più parsimonioso (quelli di cui, nel nostro caso, vogliamo verificare la significatività) siano nulli.

Il test[19] assume la forma:

$$T = -2\ln\left[\frac{\max(L_{\text{modello ridotto}})}{\max(L_{\text{modello completo}})}\right] = -2\left[\max(l_{\text{modello ridotto}}) - \max(l_{\text{modello completo}})\right] \underset{\text{sotto } H_0}{\sim} \chi^2_q,$$

dove *q* è la differenza di parametri tra i due modelli, il numero di parametri esclusi nel modello ridotto.

Grandi valori di *T* (corrispondenti a piccoli valori del p-value) portano al rifiuto di H_0 e alla scelta del modello più ampio, in tal caso infatti l'aumento di variabili esplicative viene compensato da un consistente miglioramento dell'adattamento (rispecchiato da un consistente aumento della massima verosimiglianza ottenibile).

[19] Il *test del rapporto di verosimiglianza* può essere anche chiamato, come vedremo nelle prossime pagine, ***extradevianza*** ed essere espresso come differenza tra le *devianze* dei due modelli.

VERIFICA DELLA BONTÀ DEL MODELLO IN TERMINI DI ADATTAMENTO

Il modello a cui corrisponde il massimo valore di verosimiglianza ottenibile è ovviamente quello in cui il numero di parametri coincide con il numero *n* di osservazioni a disposizione (*modello saturato*); in tal caso, infatti, riusciamo ad ottenere delle stime perfettamente coincidenti con i valori osservati[20], il perfetto adattamento del modello ai dati disponibili. Un modello di questo tipo non risulterà però né parsimonioso né tanto meno flessibile ad un'estensione ad altri insiemi di dati; è un modello eccessivamente dispendioso e senza dubbio tremendamente rigido.

Se il modello considerato (*p* parametri esclusa l'intercetta) può comunque essere ritenuto un buon modello, la massima verosimiglianza non dovrà subire eccessive diminuzioni rispetto a quella associata al modello saturato: il confronto viene anche in questo caso impostato sul test del rapporto di verosimiglianza, che assume la forma

$$D = -2\ln\left[\frac{\max(\mathrm{L}_{\text{modello considerato}})}{\max(\mathrm{L}_{\text{modello saturato}})}\right] \underset{\text{sotto } H_0}{\sim} \chi^2_{n-(p+1)} = \chi^2_{n-p-1}$$

e prende il nome di ***devianza***. Valori piccoli di tale quantità indicheranno una sostanziale equivalenza tra i due modelli in termini di verosimiglianza e di conseguenza un buon adattamento del modello da noi considerato (accettazione dell'ipotesi nulla di non significatività dei parametri tralasciati).

VERIFICA DELLA BONTÀ DEL MODELLO IN TERMINI DI PARSIMONIA

La bontà di un modello non può essere però valutata esclusivamente in termini di adattamento, che di certo risulterà migliore quanto più vicini saremo al modello saturato. La qualità complessiva dev'essere apprezzata nell'**equilibrio tra un buon adattamento e una sostanziale parsimonia** nel numero di esplicative introdotte nella *funzione di regressione*.

Un metodo per valutare questo secondo aspetto è il confronto del modello considerato con il modello avente la sola intercetta, ovvero questa volta il modello più scarno. Vediamo come il test, sempre costruito sul rapporto di verosimiglianza, possa però essere espresso anche come differenza delle devianze dei due modelli e da qui chiamato ***extradevianza***:

$$G = -2\ln\left[\frac{\max(\mathrm{L}_{\text{modello con sola intercetta}})}{\max(\mathrm{L}_{\text{modello considerato}})}\right] = D_{\text{modello con sola intercetta}} - D_{\text{modello considerato}} \underset{\text{sotto } H_0}{\sim} \chi^2_{n-1} - \chi^2_{n-p-1} \sim \chi^2_p \,.$$

[20] Si pensi ad esempio di associare ad ogni singola osservazione un solo diverso parametro e di annullare di volta in volta tutti gli altri: la sua stima, vincolata ad un solo dato, coinciderà con il valore che permette di ottenere esattamente il valore osservato.

ANALISI DEI RESIDUI

Portato a termine il processo di stima, risulta necessario verificare le ipotesi avanzate e l'accuratezza delle stime ottenute: le principali valutazioni qualitative muovono solitamente dall'analisi grafica dei *residui* restituiti dal modello, misure della distanza tra i valori stimati e quelli realmente osservati:

$r_{ti} = y_{ti} - \hat{y}_{ti}$.

In primo luogo, sarebbe importante verificare che i residui presentino una distribuzione normale, o comunque una distribuzione sufficientemente simmetrica. In realtà, per la natura bernoulliana della singola realizzazione della variabile dipendente considerata e per il fatto che le stime sono comprese tra i due estremi 0 e 1, in molte situazioni questo risultato è strutturalmente inottenibile. Cerchiamo di capirne il motivo analizzando distintamente due situazioni:

- Modello con presenza di variabili esplicative continue

  In tal caso, ad ogni diversa combinazione di valori assunti dalle variabili esplicative corrisponde generalmente una sola osservazione, di valore 0 o 1. Un buon modello presenterà delle stime piuttosto vicine a questi due valori e dei residui generalmente piccoli; nonostante ciò, dato che le stime sono comprese tra 0 e 1, tutti i residui associati alle osservazioni pari ad 1 saranno per loro natura positivi e, al contrario, tutti quelli associati alle osservazioni pari a 0 negativi. In altre parole, la distribuzione dei residui rispetto allo 0 dipende dalla natura dei dati a disposizione, dal rapporto tra "successi" e "insuccessi" sperimentati: risulta quindi ovvio che l'ipotesi di una loro distribuzione simmetrica, e a maggior ragione normale, non potrà che essere rifiutata.

  La presenza di variabili casuali continue nel modello, da cui deriva il problema strutturale appena descritto, non è rara: è necessario di conseguenza cercare di valutare i residui diversamente, in modo che possano effettivamente dar origine ad una distribuzione simmetrica[21].

  Una possibile soluzione è quella di raggruppare le unità in base a valori analoghi assunti dalle variabili esplicative, in modo da considerare come residui non tanto i singoli valori, quanto una loro media calcolata nei diversi gruppi.

[21] Per ovviare a questo tipo di problema, una soluzione alternativa a quella presentata di seguito è ottenuta ricorrendo ai ***residui di Anscombe***, qui non trattati.

Un possibile tipo di residui che rientra in quest'ottica è rappresentato dai ***residui di Pearson*** [$_Pr$], che asintoticamente si distribuiscono come una normale standardizzata, nonostante nella pratica sia difficile ottenere questo risultato per variabili dicotomiche:

$$ {}_P r_{tj} = \frac{\sum_{i=1}^{n_{tj}} \left(y_{tji} - \hat{\mu}_{tj} \right)}{\sqrt{\hat{\text{var}}\left(\sum_{i=1}^{n_{tj}} Y_{tji} \right)}} = \frac{\sum_{i=1}^{n_{tj}} y_{tji} - n_{tj}\hat{\mu}_{tj}}{\sqrt{n_{tj}\ \hat{\text{var}}\left(Y_{tji}\right)}} = \frac{\sum_{i=1}^{n_{tj}} y_{tji} - n_{tj}\hat{\lambda}_{tj}}{\sqrt{n_{tj}\hat{\lambda}_{tj}\left(1-\hat{\lambda}_{tj}\right)}}, $$

dove, considerato il gruppo *j* al tempo *t*, n_{tj} è il numero di unità, y_{tji} la generica osservazione e $\hat{\mu}_{tj} = \hat{\lambda}_{tj}$ la stima della media di *Y*, ovvero la proporzione di successi stimata dal modello.

La somma dei quadrati dei residui di Pearson fornisce la nota statistica χ^2.

- <u>Modello con variabili esplicative esclusivamente discrete</u>

  La situazione in questo caso è diversa: ogni incrocio tra le modalità assunte dalle diverse variabili esplicative raggruppa più osservazioni, plausibilmente non tutte associate alla stessa realizzazione della dipendente[22]. Di fatto, determinati i valori dei parametri, non saremo andati a stimare i valori della singole osservazioni, bensì le medie relative ai diversi insiemi di unità, la proporzione di "successi" che ci aspettiamo in corrispondenza di date combinazioni di modalità delle variabili esplicative: i valori reali corrispondenti alle stime variano ora in tutto l'intervallo [0,1], non coincidono esclusivamente con gli estremi. I residui non saranno necessariamente e incondizionatamente positivi o negativi a seconda del valore della relativa osservazione; quanto più numerosi risulteranno i gruppi, tanto più la distribuzione dei residui risulterà approssimabile ad una normale[23].

Qui di seguito, nelle descrizioni delle possibili analisi conducibili, con $\hat{y}_{ti}$ verrà indicato indistintamente il valore stimato dal modello per la singola unità o per la media riferita ad un gruppo di unità.

---

[22] Per ogni "combinazione di attributi", generalmente ci saranno sia individui che avranno sperimentato l'evento sia altri che non l'avranno sperimentato.

[23] La distribuzione binomiale è approssimabile asintoticamente alla normale.

Tra le possibili analisi che potrebbero essere prese in considerazione, di particolare importanza potrebbero risultare le seguenti:

- Considerata la *k*-esima esplicativa, la rappresentazione grafica dei punti $(x_{ki}, \hat{y}_i)$, grazie alla quale si va a valutare la correttezza dell'ipotesi di dipendenza lineare tra $\text{logit}[\Pr(Y=1|\underline{X})] = \text{logit}(\lambda)$ e X_k; nel caso in cui essa non risulti soddisfatta, l'andamento grafico può eventualmente suggerire quale relazione possa esprimere la dipendenza in modo migliore.

- Se la numerosità campionaria non è troppo elevata, la costruzione di un grafico dei residui relativamente a tutte le unità statistiche elencate in ascissa: dato che in un buon modello i residui non dovrebbero discostarsi troppo dallo 0, un'ispezione di questo tipo dovrebbe permettere di individuare eventuali ***valori anomali*** (*outliers*), tenendo comunque presente che potrebbe trattarsi solo di valori che il modello considerato non riesce a spiegare.

- La rappresentazione grafica dei punti $(r_i, \hat{y}_i)$, che in un buon modello dovrebbero risultare casualmente distribuiti attorno all'asse delle ascisse; nel caso in cui invece presentino un particolare andamento, ad esempio un sistematico aumento al crescere dei valori stimati, potrebbe essere inopportuna la scelta della logistica come *funzione legame* o inappropriata la scala di alcune variabili esplicative.

- La rappresentazione normale (*normal plot*), che prevede in ascissa i residui del modello e in ordinata altrettante realizzazioni casuali di una normale standardizzata: i residui che possono essere considerati normali sono quelli corrispondenti ai punti non eccessivamente lontani dalla diagonale principale, gli altri potrebbero corrispondere ad eventuali valori anomali. Il mancato allineamento su tale diagonale può denotare l'inadeguatezza della logistica come funzione legame.

5.7 LA SELEZIONE DELLE VARIABILI ESPLICATIVE IN UN'OTTICA MULTIVARIATA

Nel paragrafo 5.4, "*Analisi preliminari per le variabili esplicative*", è già stata presentata una serie di fondamentali criteri per la selezione di un gruppo di potenziali variabili esplicative in un'ottica univariata, in altre parole per una scelta basata sull'analisi di ogni variabile considerata singolarmente[24]; in un contesto multivariato, risulta però importante in seguito valutare nel suo complesso il gruppo di regressori così costituito e strutturare dei processi di selezione basati su criteri di buon adattamento e parsimonia del modello costruito.
Vengono presentati qui di seguito due diversi metodi orientati a tal fine: il metodo *stepwise* e il metodo *best subsets*.

IL METODO STEPWISE

Il metodo *stepwise* è impostato su una selezione progressiva delle esplicative, prevede di introdurle una alla volta nel modello secondo la loro rilevanza in relazione a tutte le concorrenti. Si ponga di aver preselezionato un insieme di p variabili candidate ad entrare nel modello come predittive (non collineari); l'algoritmo del processo di selezione si presenta come segue:

1) Viene preso come modello di partenza quello con l'intercetta come unico regressore e si analizzano i p modelli ottenibili inserendo nell'equazione di regressione una sola delle variabili a disposizione: la prima di queste ad esser selezionata è quella da cui si ottiene il maggior aumento della massima verosimiglianza rispetto a quella associata al modello iniziale. Quest'aumento viene valutato con l'*extradevianza* G, presentata nel paragrafo precedente, che in questo caso, sotto l'ipotesi di equivalenza dei due modelli, si distribuisce come un χ^2_1: i valori maggiori della statistica corrispondono ai valori minori del p-value.
   Viene comunque prefissata una soglia massima che il livello di significatività del test non deve superare perché il modello ottenuto possa essere ritenuto significativamente migliore del precedente; in caso contrario il processo termina. È

[24] In realtà è già stato introdotto il primo criterio di natura multivariata quando nel paragrafo 5.4 si è parlato della condizione di non collinearità delle esplicative.

consigliabile non adottare una soglia troppo bassa per non scartare variabili che potrebbero rivelarsi di fondamentale apporto congiuntamente ad altre: solitamente si opta per valori compresi tra 0.15 e 0.20[25].

2) Si adotta un procedimento analogo a quello descritto al punto precedente, adottando però come modello di partenza il modello che oltre all'intercetta contiene la prima variabile selezionata (in seguito l'insieme delle *k* selezionate) e valutando i diversi effetti ottenibili dall'introduzione di una delle *p-1* (in seguito *p-k*) variabili ancora a disposizione.
   Si passa al punto successivo dopo la terza selezione.

3) Dopo la terza selezione, il processo ammette che una delle variabili temporaneamente incluse nell'insieme delle esplicative possa riuscirne (*eliminazione backward*) se il miglioramento da essa apportato al modello si dimostra in seguito non rilevante.
   Per evitare un ciclico processo di continua inclusione ed esclusione di una stessa variabile, il valore minimo del p-value per cui venga accettata l'ipotesi di equivalenza tra modello completo e ridotto, ovvero di non significatività della variabile in questione, deve essere ovviamente non minore di quello fissato come soglia massima per l'inclusione tra le esplicative.

4) Il ciclo formato dal secondo e dal terzo punto si interrompe quando viene a verificarsi una prefissata condizione di arresto, basata su regole statistiche o su condizioni dettate dall'analisi intrapresa. Alcune di queste potrebbero essere le seguenti:
   - tutte le variabili sono entrate nel modello;
   - tutte le variabili il cui livello di significatività rispetta quello prefissato sono entrate nel modello;
   - pur avendo a disposizione altre variabili esplicative, il modello è soddisfacente dal punto di vista della classificazione delle unità.

[25] Hosmer e Lemeshow [1989].

Perché un processo di questo tipo risulti efficiente deve essere accompagnato da necessari accorgimenti e da adeguate scelte, in particolare:

- ✓ L'insieme delle variabili esplicative selezionate deve risultare, oltre che giustificato dal punto di vista statistico, sensato da un punto di vista logico considerata la natura del fenomeno studiato. Può essere ad esempio opportuno introdurre nel modello alcune variabili indipendentemente dal loro livello di significatività: è il caso di sesso ed età, ad esempio, se si desidera i risultati non dipendano dalla struttura della popolazione di riferimento, o di variabili a cui siano imputabili effetti altrimenti attribuiti ad altre. Bisogna sempre ricordare però che un numero eccessivo di variabili può tradursi in costi e tempi più elevati, nonché in una maggior incertezza delle stime ottenute, soprattutto in caso di esigua numerosità campionaria.
- ✓ Nel caso in cui vengano selezionate solamente alcune di una serie di variabili dummy costruite sulle modalità di un'unica variabile nominale, può essere sensato forzare l'inclusione nel modello di quelle escluse per non perdere o snaturare il significato della variabile originaria.
- ✓ Alcune variabili potrebbero risultare significative solamente se considerate contemporaneamente ad altre: in tal caso il metodo stepwise potrebbe non trovare mai la migliore combinazione tra le esplicative a disposizione, perché alcune di esse potrebbero non venir mai scelte in un processo di selezione "in avanti" a causa della loro bassa significatività. Per superare problemi di questo tipo risulta più appropriato un processo di selezione all'indietro o metodi che considerino progressivamente gruppi di variabili, ad esempio il *best subsets*.

Se uno dei vantaggi del metodo best subsets può essere quello appena segnalato di non trascurare variabili che non risultano significative se non considerate congiuntamente ad altre, il metodo *stepwise* ha il grande pregio di permettere a chi sta effettuando l'analisi di seguire progressivamente il processo di selezione e di poter eventualmente intervenire con scelte opportune: tutto ciò porta ovviamente ad una miglior comprensione della rilevanza dei singoli predittori.

IL METODO BEST SUBSETS

Si ponga anche in questo caso di aver a disposizione, dopo il processo di preselezione, un insieme di p potenziali variabili esplicative (sensate da un punto di vista logico e non collineari); il metodo *best subsets* parte da quest'ultimo e cerca di individuarne il miglior sottoinsieme in termini di capacità esplicativa sotto particolari vincoli di parsimonia.

Il criterio di selezione è basato su una misura standardizzata della varianza residua del modello, l'***indice di Mallows*** $_kC$[26], dove k è il numero di variabili esplicative prese di volta in volta in considerazione; il valore minimo (fissato a $p+1$) è ottenuto ovviamente in corrispondenza del modello più vasto, con l'intercetta e tutte le p esplicative a disposizione: un banale criterio di minimizzazione, che opterebbe per quest'ultimo, non potrebbe di conseguenza risultare adeguato, dato che implicherebbe evidenti problemi in termini di parsimonia. Tra i sottoinsiemi di dimensione k verrà ovviamente preferito quello in corrispondenza del quale l'indice assumerà valore minore; per quanto riguarda invece la scelta di tale dimensione, si procederà al suo incremento fino a quando si riterrà compensato da una sufficiente riduzione di $_kC$.

Come già visto nell'ambito del metodo precedente, anche in questo caso potrebbe risultare necessario, o quantomeno opportuno, introdurre forzatamente nel modello alcune variabili, indipendentemente dalla loro significatività statistica nel sottoinsieme considerato; questo nei casi in cui:

- siano oggetto di particolare attenzione per la natura del fenomeno studiato;
- permettano di assorbire gli effetti che una particolare struttura della popolazione di interesse può avere sulle stime dei parametri quando si riscontra una significativa variabilità tra diversi gruppi di unità;
- ad esse sia imputabile parte dell'effetto altrimenti attribuito ad altre esplicative;
- facciano parte di una serie di dummy, di cui altre già introdotte nel modello, associate alle diverse modalità di un'unica variabile nominale.

Per cercare di superare i limiti dell'uno o dell'altro metodo e di sfruttarne allo stesso tempo i rispettivi vantaggi, è possibile costruire un processo di selezione misto combinando lo *stepwise* con il *best subsets*: con quest'ultimo verrà selezionato un gruppo abbastanza ampio di predittori che poi verrà ulteriormente scremato secondo le procedure previste dal primo.

[26] Proposto nel 1975.

5.8 INTERPRETAZIONE DEI PARAMETRI

Selezionate le variabili, portato a termine il processo di stima e accertato che il modello goda di un buon adattamento ai dati, è fondamentale saper leggere in modo corretto e completo i risultati ottenuti, saper interpretare le stime dei singoli parametri.
Per poter comprendere pienamente i modelli più complessi[27], è opportuno innanzitutto considerare quelli con una sola variabile predittiva, nei diversi casi che si possono presentare a seconda che quest'ultima sia di natura dicotomica, nominale o quantitativa.

VARIABILE ESPLICATIVA DICOTOMICA

Supponiamo di considerare un modello di survival analysis in cui la sola variabile esplicativa sia dicotomica e possa assumere i valori 0 e 1, in altre parole sia una *variabile dummy* [D_1].

$$\text{logit}\left[\Pr\left(Y_{ti}=1\middle|D_1\right)\right]=\alpha_t+\beta_1 D_1$$

Il coefficiente di regressione, ovvero il parametro β_1, misura l'aumento di $\text{logit}(\lambda_{ti})$ determinato dal possesso dell'attributo D_1:

$$\text{logit}\left(\lambda_{ti}\middle|D_1=1\right)-\text{logit}\left(\lambda_{ti}\middle|D_1=0\right)=\ln\left[\frac{odds\left(\lambda_{ti}\middle|D_1=1\right)}{odds\left(\lambda_{ti}\middle|D_1=0\right)}\right]=$$

$$=\ln\left[\text{OR}\left(D_1=1,t\right)\right]=\ln\left[\frac{\Pr\left(Y_{ti}=1\middle|D_1=1\right)}{\Pr\left(Y_{ti}=0\middle|D_1=1\right)}\cdot\frac{\Pr\left(Y_{ti}=0\middle|D_1=0\right)}{\Pr\left(Y_{ti}=1\middle|D_1=0\right)}\right]=\left(\alpha_t+\beta_1 1\right)-\left(\alpha_t+\beta_1 0\right)=\beta_1\cdot$$

Il coefficiente in questione rappresenta quindi il logaritmo naturale dell'odds ratio relativo alla variabile D_1.
Come già visto nel paragrafo 5.4, se è ragionevole assumere che l'evento studiato sia raro, $\Pr\left(Y_{ti}=0\middle|D_1=1\right)$ e $\Pr\left(Y_{ti}=0\middle|D_1=0\right)$ assumono valori prossimi a 1; in tal caso l'*odds ratio* (o *rapporto crociato*), pari a $\exp(\beta_1)$, può essere ritenuto una buona approssimazione, e di conseguenza una buona stima, del *rischio relativo* (il rapporto tra il rischio associato all'evento in presenza della caratteristica rappresentata dalla dummy D_1 e quello in sua assenza):

$$\exp(\beta_1)=\frac{\Pr\left(Y_{ti}=1\middle|D_1=1\right)}{\Pr\left(Y_{ti}=0\middle|D_1=1\right)}\cdot\frac{\Pr\left(Y_{ti}=0\middle|D_1=0\right)}{\Pr\left(Y_{ti}=1\middle|D_1=0\right)}\cong\frac{\Pr\left(Y_{ti}=1\middle|D_1=1\right)}{\Pr\left(Y_{ti}=1\middle|D_1=0\right)}=R_{1/0}\cdot$$

[27] Non stiamo comunque prendendo in considerazione modelli in cui venga superata l'ipotesi di rischi proporzionali.

VARIABILE ESPLICATIVA NOMINALE

Consideriamo in questo caso una variabile X_1 a $k+1$ modalità; una volta scelta tra queste una modalità base e denominatala modalità *0*, per ognuna delle rimanenti (dalla *1* alla k) viene costruita una variabile dummy D_i, che assumerà valore 1 (mentre contemporaneamente tutte le altre saranno nulle) quando nell'unità considerata si presenterà la caratteristica ad essa associata; la modalità base corrisponderà invece alla situazione in cui tutte le dummy assumono valore nullo. Il modello di regressione sarà dunque il seguente:

$$\text{logit}\left[\Pr\left(Y_{ti}=1\middle|X_1\right)\right]=\text{logit}\left[\Pr\left(Y_{ti}=1\middle|\underline{D}\right)\right]=\alpha_t+\beta_{11}D_1+\beta_{12}D_2+\ldots+\beta_{1k}D_k$$

$$\text{logit}\left(\lambda_{ti}\middle|D_i=1\right)-\text{logit}\left(\lambda_{ti}\middle|\underline{D}=\underline{0}\right)=\ln\left[\text{OR}\left(D_i=1,t\right)\right]=\ln\left[\frac{\Pr\left(Y_{ti}=1\middle|D_i=1\right)}{\Pr\left(Y_{ti}=0\middle|D_i=1\right)}\cdot\frac{\Pr\left(Y_{ti}=0\middle|\underline{D}=\underline{0}\right)}{\Pr\left(Y_{ti}=1\middle|\underline{D}=\underline{0}\right)}\right]=$$

$$=\left(\alpha_t+\beta_{11}0+\ldots+\beta_{1i-1}0+\beta_{1i}1+\beta_{1i+1}0+\ldots+\beta_{1k}0\right)-\left(\alpha_t+\beta_{11}0+\ldots+\beta_{1k}0\right)=\beta_{1i}\,.$$

I singoli parametri β_{1i} rappresentano in questo caso il logaritmo naturale dell'odds ratio relativo all'*i*-esima modalità della variabile X_1, considerata come *modalità di riferimento* la modalità *0*.

VARIABILE ESPLICATIVA QUANTITATIVA

Consideriamo infine il caso di una variabile esplicativa quantitativa.

$$\text{logit}\left[\Pr\left(Y_{ti}=1\middle|X_1\right)\right]=\alpha_t+\beta_1X_1$$

Diversamente da quanto visto con le variabili dicotomiche o nominali, qui viene solitamente considerato come odds ratio il fattore di variazione dell'odds associato ad un incremento unitario della variabile esplicativa in questione, ed è questa quantità che in questo caso viene misurata dal coefficiente di regressione:

$$\text{logit}\left(\lambda_{ti}\middle|X_1=x+1\right)-\text{logit}\left(\lambda_{ti}\middle|X_1=x\right)=\ln\left[\frac{odds\left(\lambda_{ti}\middle|X_1=x+1\right)}{odds\left(\lambda_{ti}\middle|X_1=x\right)}\right]=\ln\left[\text{OR}\left(X_1,t\right)\right]=$$

$$=\ln\left[\frac{\Pr\left(Y_{ti}=1\middle|X_1=x+1\right)}{\Pr\left(Y_{ti}=0\middle|X_1=x+1\right)}\cdot\frac{\Pr\left(Y_{ti}=0\middle|X_1=x\right)}{\Pr\left(Y_{ti}=1\middle|X_1=x\right)}\right]=\left[\alpha_t+\beta_1\left(x+1\right)\right]-\left[\alpha_t+\beta_1x\right]=\beta_1\cdot$$

Considerare un incremento unitario della variabile esplicativa potrebbe però in date situazioni non avere un senso logico se riportato alla realtà della situazione presa in esame, magari per la natura stessa di tale variabile o per gli interessi specifici dell'analisi; potrebbe risultare più sensato valutare l'impatto di un incremento di *c* unità. In tal caso

$$\text{logit}\left(\lambda_{ti}\middle|X_1 = x + c\right) - \text{logit}\left(\lambda_{ti}\middle|X_1 = x\right) = c\beta_1$$

e l'odds ratio che si otterrebbe,

$$\varphi(c) = \exp(c\beta_1) = \frac{\Pr\left(Y_{ti} = 1\middle|X_1 = x + c\right)}{\Pr\left(Y_{ti} = 0\middle|X_1 = x + c\right)} \cdot \frac{\Pr\left(Y_{ti} = 0\middle|X_1 = x\right)}{\Pr\left(Y_{ti} = 1\middle|X_1 = x\right)} \cong \frac{\Pr\left(Y_{ti} = 1\middle|X_1 = x + c\right)}{\Pr\left(Y_{ti} = 1\middle|X_1 = x\right)},$$

andrebbe a valutare l'aumento di rischio corrispondente a tale incremento di *c* unità.

INTERPRETAZIONE DEI PARAMETRI IN UN MODELLO MULTIVARIATO

L'introduzione di ulteriori variabili esplicative nel modello, nel caso in cui tra esse vi sia qualche associazione o vi siano degli effetti d'interazione nel determinare *Y*, può comportare delle complicazioni a livello di interpretazione delle stime ottenute.

Se le variabili non interagiscono e sono indipendenti, ovvero non presentano alcuna forma d'associazione[28], la distribuzione di ciascuna non subisce cambiamenti al variare delle altre e ciò implica che la relazione tra essa e la dipendente *Y* non ne viene minimamente influenzata, né in intensità né in direzione: le stime dei parametri, tranne quella dell'intercetta, risultano in questo caso identiche a quelle ottenibili nei singoli modelli univariati.

Qualche problema può invece sorgere per l'eventuale verificarsi di una delle due situazioni seguenti:

1. Le variabili esplicative sono in qualche modo associate tra loro: la distribuzione di una o più d'una di esse varia in funzione dei valori assunti da altre: si parla in questo caso di *confondimento* in quanto l'analisi della relazione diretta tra Y e alcuni regressori viene confusa e fuorviata da altri.
2. Due o più variabili esplicative presentano *interazione* nel determinare il rischio di sperimentare l'evento.

Per non appesantire eccessivamente la notazione, al fine di facilitare la comprensione, consideriamo come esempio un modello con due sole variabili predittive.

[28] Tale ipotesi andrebbe tradotta in realtà in una condizione di ortogonalità dei regressori.

1. Confondimento

La distribuzione di una esplicativa [X_1] varia secondo i valori assunti da un'altra [X_2]: l'introduzione di questa seconda variabile nel modello depura la relazione diretta tra Y e X_1 che si osserverebbe in un modello univariato, andando di fatto ad isolare gli effetti attribuibili separatamente ai due regressori (sempre nel caso non vi siano effetti d'interazione).

Le rette che rappresentano la relazione tra $\text{logit}(\lambda_{ti})$ e X_1, condizionatamente ai livelli di X_2, saranno tra loro parallele, ovvero avranno diversa intercetta ma lo stesso **coefficiente angolare**, seppur **diverso da quello stimabile nel caso univariato**: β_1 misura ora l'effetto di X_1 su Y al netto degli effetti di X_2.

$$\text{logit}\left[\Pr\left(Y_{ti}=1 \middle| X_1=x+1, X_2=x_2\right)\right]-\text{logit}\left[\Pr\left(Y_{ti}=1 \middle| X_1=x, X_2=x_2\right)\right]=$$

$$=\left[\alpha_t+\beta_1(x+1)+\beta_2 x_2\right]-\left[\alpha_t+\beta_1 x+\beta_2 x_2\right]=\beta_1$$

Nel caso vi fossero ulteriori variabili confondenti, andrebbero anch'esse introdotte nell'equazione di regressione e β_1 rappresenterebbe, analogamente a quanto visto ora, la variazione di $\text{logit}(\lambda_{ti})$ corrispondente ad un incremento unitario di X_1 al netto degli effetti di tutte le altre variabili introdotte nel modello.

2. Interazione

Si parla di interazione tra due variabili esplicative se queste, nel determinare la variabile dipendente, oltre ai loro eventuali contributi additivi hanno un effetto moltiplicativo, sia esso positivo (si parla in tal caso di effetto "più che additivo" o di *sinergia*) o negativo (effetto "meno che additivo" o *inibizione*)[29].

$$\text{logit}\left[\Pr\left(Y_{ti}=1\middle|X_1,X_2\right)\right]=\alpha_t+\beta_1X_1+\beta_2X_2+\gamma\ X_1X_2$$

In questa situazione, le rette che descrivono il legame tra $\text{logit}(\lambda_{ti})$ e X_1 avranno **diverso coefficiente angolare a seconda del livello di X_2 considerato**; ciò si traduce nel fatto che l'odds ratio di una variabile dipende dal livello di quelle con essa interagenti:

$$\ln\left[\text{OR}\left(X_1\middle|X_2=x_2\right)\right]=\text{logit}\left[\Pr\left(Y_{ti}=1\middle|X_1=x+1,X_2=x_2\right)\right]-\text{logit}\left[\Pr\left(Y_{ti}=1\middle|X_1=x,X_2=x_2\right)\right]=$$
$$=\left[\alpha_t+\beta_1(x+1)+\beta_2x_2+\gamma(x+1)x_2\right]-\left[\alpha_t+\beta_1x+\beta_2x_2+\gamma\ x\ x_2\right]=\beta_1+\gamma\ x_2\,.$$

Ogni parametro si riferisce esclusivamente all'effetto additivo del regressore di cui è coefficiente: non esprime (diversamente da quanto avveniva nei casi precedenti) l'effetto complessivo di una variabile quale fattore di rischio, dovuto invece ad un insieme di contributi dati da più regressori; le singole stime non corrispondono quindi al logaritmo dell'odds ratio delle diverse variabili, perché quest'ultimo deve tener conto anche degli effetti d'interazione.

La varianza della stima del logaritmo dell'odds ratio di una variabile, di conseguenza, non è più data dalla varianza della stima di un unico parametro a lei associata, ma assume forma:

$$\text{var}\left\{\ln\left[\hat{\text{OR}}\left(X_1\middle|X_2=x_2\right)\right]\right\}=\text{var}\left(\hat{\beta}_1\right)+\text{var}(\hat{\gamma})x_2{}^2+2\,\text{cov}\left(\hat{\beta}_1,\hat{\gamma}\right)x_2\,.$$

Ovviamente, anche in questo caso tutte le considerazioni fatte si possono estendere all'interazione di più di due regressori.

[29] Si ricordi a tal proposito il caso della dimensione temporale, già analizzato nel paragrafo 5.3.