****

***Frontespizio***

**Università degli studi di Roma**

**Unitelma Sapienza**

**Dipartimento di Scienze giuridiche ed economiche**

**Corso di Laurea TRIENNALE in SCIENZE DELL’ECONOMIA E DELLA GESTIONE AZIENDALE**

**Tesi di Laurea in**

**[Statistica]**

**Classificazione binaria tramite la regressione logistica: caso di studio sul *cross-selling* bancario**

Relatore:

Ch.mo Prof. **Pasquale Sarnacchiaro**

Candidato:

Americo Costantini

Anno Accademico 2016/2017

INTRODUZIONE

# Obiettivo dell’elaborato

Lo scopo di questo elaborato è la creazione di un modello predittivo volto alla classificazione delle unità di analisi secondo due possibili uscite, genericamente “evento” e “non evento” oppure “presenza dell’attributo” e “assenza dell’attributo”. A tal fine si è deciso di utilizzare la regressione logistica. Un approccio scientificamente ottimale non contempla la scelta a priori del modello, bensì l’utilizzo di più tecniche differenti e poi la selezione (o la combinazione) del modello più performante secondo le metriche di *performance* prescelte[[1]](#footnote-1). Nel nostro caso la scelta è stata fatta a priori perché un approccio come quello su descritto va oltre i limiti delle ambizioni dell’elaborato che si sta redigendo, inducendo quindi chi scrive a individuare a priori un singolo metodo con il quale svolgere l’analisi dei dati. Data questa limitazione, la regressione logistica è sembrata la scelta più idonea in quanto tra i metodi di classificazione binaria è quello che più permetteva un approfondimento di concetti di statistica inferenziale, facendo così scampare il pericolo di una eccessiva robotizzazione (leggi: scarsa interpretabilità) del modello predittivo.

La struttura dello scritto consiste di due capitoli iniziali di presentazione degli strumenti statistici poi utilizzati nel terzo capitolo, quello centrale, di svolgimento di una analisi dei dati originale. Alla luce di ciò si è limitato l’approfondimento di dimostrazioni matematiche delle formule presenti a quanto era strettamente necessario per la comprensione dei concetti al fine del loro utilizzo nell’analisi stessa.

Lo scopo dell’analisi è quello di stimare un modello predittivo sui dati raccolti da una banca durante una campagna promozionale per la sottoscrizione di un nuovo contratto rivolta a un campione dei suoi clienti. L’obiettivo sarà individuare quelle caratteristiche che permettano di definire in anticipo, su tutta la popolazione dei clienti, quali saranno favorevoli alla sottoscrizione del nuovo contratto e discriminarli da quelli contrari, in modo da canalizzare in maniera efficiente lo sforzo dei venditori della banca[[2]](#footnote-2).

# Notazione matematica[[3]](#footnote-3)

Si denoterà con il numero di osservazioni del dataset, vale a dire la dimensione del campione. Con il numero di variabili indipendenti presenti nel dataset, quindi il numero di parametri. Con lettere maiuscole, , ecc. verranno indicate genericamente le variabili casuali. Data una matrice che in riga ha le misurazioni per ogni i-esima osservazione e in colonna le misurazioni del j-esimo parametro per tutte le unità osservate, allora si denoterà con la misurazione del j-esimo parametro relativa alla i-esima osservazione, con il vettore contenente le misurazioni per la i-esima unità e con il vettore delle misurazioni del j-esimo parametro. Inoltre con verrà indicata la realizzazione i-esima della variabile casuale . Con si indicherà la probabilità che un evento accada, mentre con si rappresenterà una funzione dipendente da alcuni parametri la cui uscita è la probabilità che l’attributo sia presente. Ad esempio, significherà: la probabilità che la variabile casuale abbia una realizzazione pari a per la i-esima osservazione dati certi valori delle variabili indipendenti, mentre sarà la funzione che, dati certi valori delle variabili indipendenti, avrà come uscita la probabilità che. Con si denoterà il vero coefficiente del j-esimo parametro, ma soprattutto con il relativo coefficiente stimato. Anche starà per la stima del valore della variabile di risposta per la i-esima osservazione. In generale il simbolo del “cappello” (*hat*, in inglese) starà a significare che quel valore è stimato e non osservato. Con si indicherà l’insieme dei coefficienti.

Quanto al lessico, la variabile dipendente potrà alternativamente essere chiamata variabile di uscita, variabile di risposta, mentre quelle indipendenti verranno indistintamente denotate come parametri, regressori, variabili esplicative.

CAPITOLO PRIMO

I MODELLI PREDITTIVI E LA CLASSIFICAZIONE BINARIA

# La stima di una funzione di apprendimento statistico

### L’obiettivo della stima della funzione

### L’accuratezza della stima

llllll

### Varianza e distorsione del modello

# Categorie di algoritmi di classificazione binaria

### Regressione contro classificazione

### Metodi parametrici e non parametrici

### Metodi supervisionati e non-supervisionati

### Obiettivo: predizione o inferenza

# Metodi di ricampionamento e potere predittivo

### Cross validation

### K-fold cross validation

# La selezione delle variabili

### Dummy variables

### Esplorazione tramite l’information value

### L’adjustment

### Tecniche di selezione automatica delle variabili

# Metodi di valutazione delle performance discriminatorie

### Curva ROC e AUC

### Matrice di confusione

### Indice di Gini

CAPITOLO secondo

LA REGRESSIONE LOGISTICA

# La funzione di ipotesi nella regressione logistica

### Perché non la regressione lineare

L’intento della classificazione binaria è predire, in base ai valori assunti da variabili indipendenti, l’uscita di una variabile dicotomica che può assumere soltanto due valori:

*Y*=

Un tipico approccio per la predizione è quello di stimare la probabilità che assuma il valore in base ai valori del vettore delle variabili indipendenti, . Come noto, la probabilità che un evento accada può assumere soltanto valori tra e *[[4]](#footnote-4)*, quindi a questo scopo la regressione lineare è inadeguata, in quanto modello volto a predire il valore di una variabile dipendente quantitativa che non può essere sottoposta a una restrizione di intervallo di questo tipo[[5]](#footnote-5).

Inoltre una variabile casuale dicotomica ha una distribuzione di *Bernoulli*, che prevede una varianza pari a . Tale varianza dipende da un valore non costante (perché l’assunto di cui sopra è che , quindi che la probabilità che l’attributo sia presente dipenda dai valori delle variabili indipendenti), e quindi non è costante. Se ne conclude quindi ulteriormente l’inadeguatezza dell’utilizzo della regressione lineare, che regge sull’assunto della omoschedasticità e quindi della omogeneità della varianza per tutte le combinazioni di valori dei regressori[[6]](#footnote-6).

Insomma, l’obiettivo è quello di individuare una funzione che a partire dal generico vettore dei valori delle variabili indipendenti ***X*** fornisca una probabilità che :

[[7]](#footnote-7)

e quindi:

che non rappresenta altro che la media condizionata a di .

### La funzione logistica

Una funzione adatta allo scopo di legare ***X*** a è la cosiddetta funzione logistica[[8]](#footnote-8):

dove è anch’essa una funzione ancora da specificare. Una funzione del genere, detta “funzione logistica”, assume la seguente forma:



La funzione logistica passa sempre per il punto di coordinate , dove interseca l’asse delle ordinate. Tale forma assunta dalla funzione logistica si chiama sigmoide. Una delle più importanti proprietà di questa funzione è che qualunque valore assuma , il valore di sarà sempre compreso nell’intervallo . Se tende a allora il valore di uscita tenderà a , in caso contrario a . Questo è esattamente quello che ci aspettiamo da una funzione la cui uscita deve essere una probabilità. può quindi essere visto come una sintesi dei vari fattori che determinano la probabilità dell’accadimento di un evento[[9]](#footnote-9).

### La trasformazione in relazione lineare

La funzione logistica può assumere una forma molto più interpretabile se passiamo dalla considerazione della probabilità a quella dell’ODDS:

ODDS =

L’ODDS è quindi il rapporto tra probabilità complementari ed esprime quanto le probabilità a favore sono maggiori (o minori) di quelle contro[[10]](#footnote-10).

L’ODDS è una misura che che si muove nell’intervallo . Tuttavia, se all’ODDS applichiamo il logaritmo naturale il risultato è una quantità che va da . Otteniamo così il cosiddetto LOGIT:

Se vediamo la probabilità che un evento accada (leggi: che un attributo sia presente, cioè che ) come funzione di regressori , è matematicamente dimostrabile[[11]](#footnote-11) questa equazione:

= z

A questo punto, se esprimiamo z come una combinazione lineare delle variabili indipendenti abbiamo realizzato la trasformazione della funzione logistica in una funzione lineare, tramite il link del LOGIT[[12]](#footnote-12), e possiamo modellare la probabilità che un evento a due uscite accada come:

dove quindi il LOGIT è una funzione lineare che dipende da variabili indipendenti e una intercetta, quindi parametri.

# La stima e l’interpretazione dei coefficienti

### Analisi campionaria

Poiché l’analisi statistica si rivolge quasi sempre a campioni estratti dalla popolazione e non alla totalità di essa, il valore dei coefficienti potrà essere soltanto stimato, e non calcolato. Se anche sia vera l’assunzione che la vera funzione che spiega il fenomeno è la funzione logistica, lo studioso non potrà altro che stimare a partire da un campione i coefficienti, e quindi predire l’uscita della variabile dipendente soltanto a partire da quelle stime campionarie, sui cui poi verranno applicate le tecniche di inferenza statistica per estendere quanto più possibile le conclusioni all’intera popolazione, secondo un certo grado di incertezza.

### La stima tramite il metodo della massima verosimiglianza

Esistono varie tecniche per stimare i coefficienti di un modello. Nella regressione lineare il più usato è il metodo OLS, *ordinary least squares*, il metodo dei minimi quadrati. Nella regressione logistica, poiché viene violata l’assunzione di omoschedasticità, questa tecnica non si può usare[[13]](#footnote-13). Si usa invece il metodo della massima verosimiglianza. In termini generici, si tratta di creare una funzione con i parametri del modello di studio (nel nostro caso quello logistico), la cui uscita è la probabilità che venga estratto dalla popolazione proprio il campione che stiamo analizzando; poi, tramite tecniche matematiche, risolvere l’equazione e dare un valore ai coefficienti col fine di massimizzare questa probabilità. La massima verosimiglianza è una tecnica ampiamente usata, e invero la stima tramite OLS nella regressione lineare conduce ai medesimi risultati[[14]](#footnote-14). I dettagli matematici della funzione di massima verosimiglianza sono al di là dello scopo di questo elaborato, basti aggiungere che la si può risolvere solo con metodi numerici iterativi. Asintoticamente, sotto condizioni non particolarmente restrittive, gli stimatori di massima verosimiglianza sono corretti, normodistribuiti ed efficienti[[15]](#footnote-15).

### Il significato dei coefficienti nel modello di regressione logistica

Ci sono differenti modi di interpretare il significato di un coefficiente in una regressione logistica. Per semplificare, immaginiamo uno scenario con una sola variabile. La funzione sarebbe la seguente:

è il parametro intercetta, ovvero il valore del LOGIT quando tutte le variabili indipendenti assumono valore pari a 0. rappresenta l’aumento del LOGIT per un aumento unitario di , proprio come nella regressione lineare. L’aumento del LOGIT significa generalmente un aumento delle probabilità a favore, anche se rimane una interpretazione non troppo concreta. Se tuttavia facciamo una trasformazione esponenziale, allora l’equazione diventa

Allora un aumento unitario di si risolverebbe così:

e quindi rappresenterebbe l’esponente di *e* tale per cui un aumento unitario di comporterebbe un aumento moltiplicativo e non additivo di sull’ODDS.

Nel mondo reale di solito i fenomeni sono multivariati, e quindi il modello logistico conterrà coefficienti. Proprio come nella regressione lineare, il valore di ogni è stimato come *adjusted*, vale a dire che viene stimato tenendo conto di tutti gli altri parametri e che deve essere interpretato numericamente come il valore che il coefficiente assume mantenendo tutti gli altri coefficienti costanti[[16]](#footnote-16).

# Test di ipotesi sui coefficienti del modello

Nella costruzione di un modello predittivo tramite la regressione logistica l’obiettivo di un test di ipotesi è capire quanto probabile possa essere, sotto certe assunzioni, che un regressore risulti utile alla predizione solo in ragione della casualità con cui è stato estratto il campione analizzato. Si tratta di una fondamentale tecnica di inferenza statistica che ha l’ambizione di estendere, sotto dei gradi di incertezza definiti, le conclusioni campionarie all’intera popolazione[[17]](#footnote-17). Esistono due approcci fondamentali al test di ipotesi dei coefficienti nella regressione logistica: il test rapporto di verosimiglianza e il test di Wald. Entrambi assumono come ipotesi nulla che alcuni coefficienti del modello siano pari a zero; se l’ipotesi viene rigettata, significa che il regressore associato a quel coefficiente è utile al modello.

Vediamoli entrambi nel dettaglio.

### Il test di rapporto di massima verosimiglianza

Il test rapporto di massima verosimiglianza confronta il valore assunto dalla funzione di massima verosimiglianza per due modelli: il modello cosiddetto *full*, esteso, che contiene più parametri, e il modello ridotto, che ne contiene di meno[[18]](#footnote-18). I parametri assenti dal secondo – perché assunti come aventi coefficienti pari a zero – sono quelli su cui si vuole testare l’ipotesi nulla. La statistica test con cui si rigetta o meno l’ipotesi viene costruita come differenza tra i due valori di massima verosimiglianza (la funzione , *likelihood* in inglese), a cui poi si applica il logaritmo naturale moltiplicato per due e cambiato di segno. Per le proprietà dei logaritmi viene che:

dove LRT sta per *likelihood ratio test* e per *likelihood* (verosimiglianza). Si può dimostrare che, sotto l’assunzione di un campione abbastanza grande, questa statistica test ha una distribuzione con gradi di libertà pari al numero di parametri di differenza tra i due modelli[[19]](#footnote-20). Se denominiamo con l’insieme dei parametri presenti nel modello più esteso ma non in quello ridotto, allora la nostra ipotesi nulla è:

:

o anche

:

e cioè che nessuno dei parametri aggiuntivi porti un contributo al modello. Se invece questo dovesse accadere, possiamo aspettarci che ; più piccolo sarà questo valore, più grande sarà allora . Alti valori della variabile casuale comportano *p-value* molto bassi; significa quindi che la probabilità che un alto valore di (e quindi l’evidenza che i coefficienti dei parametri aggiuntivi sono diversi da ) sia dovuto al caso sotto l’ipotesi nulla (e cioè che i coefficienti dei parametri aggiuntivi sono uguali a ) è molto bassa; se è inferiore al livello di significatività del test (usualmente ) allora si avrà il rigetto di : i coefficienti aggiuntivi sono utili al modello.

### Il test di Wald

Il test di Wald invece è costruito dividendo il coefficiente stimato per il suo errore standard stimato. Questa statistica test si può dimostrare abbia una distribuzione normale standardizzata se il campione è abbastanza grande[[20]](#footnote-21). Basterà quindi utilizzare la statistica test [[21]](#footnote-22)per rigettare o meno l’ipotesi nulla che il coefficiente sia diverso da zero.

### Test multipli e correzione di Bonferroni

Per completezza, andrebbe aperta una digressione sull’incremento della probabilità di errore di tipo I nel test di ipotesi quando si effettuano test multipli (come nel caso di più coefficienti da testare). L’errore di tipo I è la probabilità di rigettare l’ipotesi nulla quando questa è vera, ed è pari a . La probabilità di non commettere nessun errore è quindi pari a 0.95. Questa probabilità è chiamata FWER (*family wise error rate*)[[22]](#footnote-23) e si può esprimere come . Matematicamente si esprime come:

[[23]](#footnote-24)

dove è il numero di test eseguiti. Al crescere di questo cresce il FWER.

Esistono procedure di aggiustamento del livello di significatività che tengono conto di questo fenomeno, la più famosa delle quali è il metodo Bonferroni, che consiste nel diminuire in livello di significatività del test dividendo quello stabilito per T[[24]](#footnote-25). Il rischio del metodo Bonferroni è che rendendo molto difficile il rigetto dell’ipotesi nulla possa abbattere la potenza del test[[25]](#footnote-26), cioè la probabilità di rigettare l’ipotesi nulla quando questa è falsa.

# Valutazione della bontà di adattamento ai dati del modello

Esiste una differenza sostanziale tra statistiche test atte a testare la bontà di adattamento ai dati di un modello e misure del potere predittivo dello stesso modello[[26]](#footnote-27). Il potere predittivo di un modello misura quanto il modello predice in maniera corretta la variabile di risposta grazie ai valori delle variabili indipendenti. I test statistici sulla bontà dell’adattamento ai dati invece testano se un modello differente si adatti ai dati meglio di quello appena stimato. Capita che un modello con scarso potere predittivo si adatti bene ai dati (in tal caso sarà un fenomeno con tantissimo rumore e quindi con un termine di errore nella funzione vera molto grande) o che un modello con scarso potere predittivo si adatti bene ai dati (in tal caso siamo di fronte a *overfitting*).

Come misure del potere predittivo di una regressione logistica sono utilizzate, tra le altre, differenti versioni di e l’AUC (*area under the curve*)[[27]](#footnote-28). È fuori dallo scopo di questo elaborato approfondire le argomentazioni a favore di una certa versione dell’ piuttosto che di un’altra (ne sono state individuate almeno 12, di cui le più importanti quella di Mc Fadden e quella di Cox e Snell). Nella nostra ricerca useremo l’AUC, che trova un consenso molto meno controverso tra gli studiosi. Questo paragrafo vuole trattare invece le statistiche di bontà di adattamento ai dati di un modello.

### I residui

Stimato un modello, si può dire che si adatti bene ai dati se, trovata una misura dello scarto tra valore stimato e valore osservato, la misura complessiva per le unità statistiche analizzate è molto piccola, e se il contributo dell’i-esimo confronto a tale misura rientra nei limiti della variabilità intrinseca del fenomeno, descritta dal termine d’errore del modello[[28]](#footnote-29).

Nella regressione lineare la bontà di adattamento ai dati osservati è una funzione dei residui, definiti come la differenza tra valore reale e valore atteso per ogni osservazione i-esima . Nella regressione logistica il valore stimato della variabile di uscita da cui calcolare il residuo non è associato alla i-esima osservazione, ma al k-esimo gruppo di osservazioni accomunate dalla medesima combinazione di valori delle varabili indipendenti (*covariate pattern*)[[29]](#footnote-30).

Esistono due differenti metodologie per calcolare i residui di un modello di regressione logistica, il residuo di *Pearson* e il residuo di *deviance*.

Sia il numero di *covariate pattern* delle variabili indipendenti osservate per le *n* unità statistiche; sia il numero di unità statistiche presenti in ogni g-esimo covariate pattern ; siano e rispettivamente il numero osservato e il numero stimato di unità statistiche appartenenti al k-esimo *covariate pattern* per cui , cioè:

Allora il singolo k-esimo residuo di *Pearson* si definisce come:

cioè come residuo diviso la sua deviazione standard, e una misura sintetica della bontà di adattamento come somma dei quadrati dei residui:

=

che per campioni sufficientemente grandi e per si distribuisce come una variabile casuale con gradi di libertà (sotto l’ipotesi nulla che il modello stimato sia quello corretto)[[30]](#footnote-31). Perciò valori bassi della statistica , che conducono a *p-value* molto alti, porteranno all’impossibilità di rigettare l’ipotesi nulla e quindi di considerare il modello stimato adattato ai dati.

Per spiegare il residuo di *deviance* utilizzeremo invece un altro approccio[[31]](#footnote-32). Un modello saturo è un modello che contiene tanti parametri quante sono le osservazioni, e non genera residui poiché le sue predizioni sono sempre esatte. La devianza è una statistica test che confronta le performance di un modello saturo con quelle del modello da testare. Si può dimostrare[[32]](#footnote-33) che la devianza è una statistica test risultante dalla differenza tra il valore di massima verosimiglianza del modello saturo e il valore di massima verosimiglianza del modello da testare:

Questo test assume la forma di un test rapporto di massima verosimiglianza, e quindi si potrebbe ragionevolmente concludere che la statistica associata si distribuisca, sotto l’ipotesi nulla, come una variabile casuale con ) gradi di libertà. Data l’ipotesi nulla:

:

dove racchiude tutte le stime dei coefficienti presenti nel modello saturo ma non in quello stimato, allora un *p-value* sotto il livello di significatività (quindi una statistica molto grande, quindi un rapporto tra le verosimiglianze molto vicino a , quindi i coefficienti dei parametri aggiuntivi presenti nel modello saturo pari a ) ci indurrebbe a rigettare l’ipotesi nulla e quindi a concludere che il modello saturo non fornisce un contributo significativamente migliore del nostro modello. La bontà di adattamento infatti testa non il potere predittivo del modello, ma se un modello con più termini di interazione (piuttosto che con una funzione link differente) si adatti meglio ai dati in termini di entità dei residui[[33]](#footnote-34).

Ora, per le due statistiche sulla bontà di adattamento finora viste, la statistica di *Pearson* e la devianza , l’approssimazione della statistica test alla variabile casuale si verifica se e solo se il numero dei gruppi di soggetti accomunati dalla stessa combinazione di valori delle variabili indipendenti è molto inferiore rispetto alla grandezza del campione[[34]](#footnote-35) e se la numerosità degli eventi e dei non eventi presenti in ogni singolo *covariate pattern* è almeno. Con molte variabili continue presenti nel dataset, questo assunto viene facilmente violato e i test non sono più affidabili.

### La statistica Hosmer – Lemeshow (HL)

Una ulteriore statistica test per la bontà di adattamento ai dati è stata proposta dagli studiosi Hosmer e Lemeshow[[35]](#footnote-37), per ovviare ai limiti di statistiche test che si basato sul numero di *covariate pattern* .

Hosmer e Lemeshow hanno proposto il raggruppamento delle osservazioni in base ai valori stimati della variabile dipendente . In particolare, i valori predetti sono ordinati dal più piccolo al più grande, e poi separati in gruppi di approssimativamente uguale dimensione. Dieci gruppi è la raccomandazione standard.

Per ciascun gruppo, si calcola il numero osservato di eventi () e non eventi () – presenza o assenza di attributo, così come il numero atteso di eventi (e non eventi (). Il numero atteso di eventi è solo la somma delle probabilità previste per tutti le osservazioni del gruppo. E il numero atteso di non eventi è la dimensione del gruppo meno il numero atteso di eventi.

La statistica HL risultante

si distribuisce come La variabile casuale con gradi di libertà e può essere utilizzata per confrontare i conteggi osservati con conteggi attesi. Se il modello da testare fosse un modello saturo, la statistica HL sarebbe nulla; sotto l’ipotesi nulla che il modello è corretto, un basso valore di HL e quindi un alto *p-value* ci indurrebbero a non poter rigettare l’ipotesi nulla che il modello è corretto[[36]](#footnote-38). Per quanto anch’essa oggetto di critiche la statistica HL è ampiamente la più usata e quella su cui c’è maggiore convergenza tra gli studiosi[[37]](#footnote-39).

# Selezione del miglior modello

I metodi di selezione delle variabili per la costruzione di un modello sono stati già descritti nel Capitolo Primo. Questi possono applicarsi anche nel caso della regressione logistica. Sono stati introdotti anche dei metodi specifici di selezione delle variabili per la regressione logistica come il *Purposeful selection**algorith*[[38]](#footnote-40) , che rappresenta comunque un ibrido dei modelli di selezione basati sul *p-value* dei coefficienti. Essenziale nella scelta del modello è lo scopo del medesimo: se il fine è la predizione, bisogna avvalersi di misure del potere predittivo: in tal caso una *cross-validation* che seleziona la combinazione di variabili con l’AUC maggiore sarà più adeguato. Se il fine è l’inferenza, specialmente nel caso in cui si voglia analizzare il rapporto causale tra una variabile di esposizione e una variabile di risposta alla luce dell’influenza di alcuni fattori di rischio, un approccio possibile è la costruzione di un modello gerarchico che includa la variabile di esposizione, alcune variabili di controllo e altre di interazione[[39]](#footnote-41). È al di là dello scopo dell’elaborato approfondire le tecniche di selezione di un modello gerarchico, in quanto nella analisi dei dati il nostro obiettivo sarà la predizione e tutte le variabili verranno trattate come possibili variabili esplicative.

# Diagnostica nella regressione logistica

### Assunzioni del modello

La regressione logistica, al contrario della regressione lineare, regge su molte meno assunzioni che, se violate, tolgono validità al modello. Non viene assunta una dipendenza lineare tra la variabile dipendente e quelle indipendenti; la regressione logistica può spiegare effetti non lineari in ragione del fatto che la funzione di link (il LOGIT) non è lineare. Inoltre non ci sono assunzioni sulla distribuzione delle variabili dipendenti, né sulla omogeneità della varianza della variabile dipendente per ogni livello (combinazione di valori) di quelle indipendenti. Da ultimo, non vengono fatte assunzioni sulla normalità della distribuzione del termine di errore casuale[[40]](#footnote-42). La diagnostica del modello non deve quindi orientarsi a verificare codeste assunzioni, come accade invece nel modello di regressione lineare.

Il modello di regressione logistica invece assume che le osservazioni siano indipendenti tra di loro[[41]](#footnote-43); questo può non verificarsi in ragione dei dati che si è raccolto (serie storiche piuttosto che campionamento a grappoli[[42]](#footnote-44)); altra assunzione è che tutte le variabili significative siano state incluse (la cosiddetta assunzione di specificazione del modello). La regressione logistica assume inoltre che tra LOGIT e variabili dipendenti ci sia una relazione lineare.

La trattazione seguente è frutto di ricerche che hanno sofferto di una difficoltà fondamentale: non c’è davvero comunanza di vedute tra gli studiosi su quali siano i migliori e più adatti e affidabili strumenti diagnostici nella regressione logistica[[43]](#footnote-45).

### Analisi dei residui

Nella regressione lineare l’analisi dei residui gioca un ruolo importante nella verifica della violazione delle assunzioni del modello; l’idea generale è che il residuo debba essere indipendente dai predittori, vale a dire che nella stima non ci si debba mai sbagliare in maniera sistematica, secondo un qualche trend. Se così fosse, ci sarebbe un effetto sistematico che agisce e che può essere spiegato includendo nel modello qualche regressore non utilizzato, rimanendo così solo il residuo dovuto alle limitazioni della stima campionaria e alla realizzazione del termine di errore casuale (il rumore) dei parametri (per il fatto che nessuna funziona descrive perfettamente un fenomeno).

Nella regressione logistica il problema è che, innanzitutto, di residui si può parlare solo per dati raggruppati, dove , e inoltre essendo il residuo k-esimo dipendente da , il residuo può dipendere dal regressore essendo comunque il modello corretto. Quindi non è chiaro che tipo di pattern rimanente possa essere sospetto oppure no[[44]](#footnote-46). Chi scrive nota una certa carenza di letteratura sull’argomento, essendo anche i manuali più noti sprovvisti della trattazione di questo argomento[[45]](#footnote-47). Questo non stupisce considerando la scarsità di assunzioni del modello di regressione logistica.

### Multicollinearità

È statisticamente dimostrabile che includere in un modello lineare (e anche in un modello lineare generalizzato) regressori linearmente correlati tra loro comporta un aumento dell’errore standard degli stimatori dei coefficienti dei regressori stessi, fenomeno noto come multicollinearità o inflazione della varianza[[46]](#footnote-48). A volte il ricercatore può decidere di includere delle variabili (magari di controllo o comunque variabili che se escluse creerebbero problemi di distorsione della stima) anche se incidono notevolmente sulla varianza dei coefficienti stimati (e quindi sui test di ipotesi e sulla stima intervallare della variabile di risposta); tuttavia una diagnostica su un modello deve garantire il governo di questo fenomeno e una scelta consapevole sull’inclusione o meno di variabili fortemente correlate ad altre.

Poiché la multicollinearità può riguardare più di due variabili, non basta una matrice degli scatterplot per individuare la correlazione; si pensi a = - .

Serve quindi uno strumento diagnostico adatto a modelli multivariati; nel nostro caso proponiamo il CNIs (*condition indices*) e il VDPs (*variance decomposition proportions*)[[47]](#footnote-49). I due indici sono stati creati per il modello di regressione lineare, e sono stati poi adattati anche alla regressione logistica[[48]](#footnote-50). Spiegare come vengono calcolati i due indici richiederebbe tecniche di algebra matriciale, che è al di là degli scopi di questo elaborato.

Come regola d’uso si può procedere come segue: data la tabella con CNIs e VDPs:



Sulla prima riga ci sono i vari CNIs. Se il CNI più grande è considerevolmente maggiore di , allora se almeno due valori della VDP legata a quel CNI sono maggiori di , allora c’è un problema di multicollinearità[[49]](#footnote-51). Esso si può affrontare togliendo la variabile con VDP più alto e riprocessando il modello per poi verificare in sede diagnostica se il problema di multicollinearità sussiste ancora.

Un metodo alternativo potrebbe essere quello di combinare due variabili correlate in un'unica variabile che contenga entrambe le informazioni (ad esempio combinare altezza e peso nell’indice di peso corporeo[[50]](#footnote-52)), mantenendo così l’informazione ma al tempo stesso non soffrendo di nessun problema di inflazione della varianza degli stimatori dei coefficienti.

### Osservazioni influenti

Nella regressione lineare vi sono misure atte a diagnosticare il comportamento anomalo di alcune osservazioni, per capire se alcune di esse sono molto lontane dal resto dei *datapoints* (gli *outlier*) o se alcune di esse esercitano una leva significativa sulla retta di regressione[[51]](#footnote-53). In un modello multivariato l’ispezione grafica è purtroppo insufficiente come strumento diagnostico, servono invece indici sintetici. Misure che si basano sull’analisi dei residui risentono dell’assunto di fondo che il numero di *covariate pattern* sia molto minore del numero di osservazioni. Poiché noi non vogliamo sottostare a questo assunto[[52]](#footnote-54) tratteremo solo misure diagnostiche che non lo prevedono.

Tecnicamente una osservazione è influente se e esercita alta leva se la sua rimozione dal modello già stimato comporta un cambiamento significativo del valore stimato di alcuni o tutti i coefficienti[[53]](#footnote-55).

I *dfbetas* o *delta-betas* sono una misura del cambiamento nel valore stimato di tutti i coefficienti qualora venga rimossa una certa osservazione. Tale misura coincide con la differenza tra le due stime diviso l’errore standard dello stimatore iniziale. Sia la stima di escludendo dal campione la i-esima osservazione, allora:

Ad ogni i-esima osservazione vengono quindi associati tutti i coefficienti e se ne può verificare il cambiamento nella stima qualora quella certa osservazione venga rimossa. La distribuzione dei *dfbetas* è ignota quindi non si può stabilire una volta per tutte quando sia davvero troppo grande per non considerare una osservazione influente; come regola d’suo si stabilisce che un valore maggiore di due sia altamente indicativo di una osservazione influente su almeno un coefficiente[[54]](#footnote-56). Individuare una osservazione come influente non determina necessariamente la decisione di rimuoverla dal dataset e quindi dal modello; i valori anomali che assume potrebbero dipendere da un errore di misurazione tanto quanto da caratteristiche proprie del fenomeno di cui bisogna comunque tenere conto.

### Errata funzione di link – mal specificazione del modello

Una assunzione fondamentale della regressione logistica è il rapporto lineare tra predittori e LOGIT, la cosiddetta funzione di link. Un modello può essere inadeguato perché manca questo assunto, o perché non sono state incluse variabili significative nel modello. Una tecnica per diagnosticare se la funzione di link è adeguata è quella di includere nel modello predittori quadrati dei predittori già inseriti; se il coefficiente è significativamente diverso da zero, allora la funzione di link è mal specificata. È importante che una tecnica del genere la si può applicare anche a dati non raggruppati, dove non è molto minore di [[55]](#footnote-57). Si potrebbe tentare anche la strada dell’ispezione grafica – sempre utile, che tuttavia soffrirebbe di due limitazioni: la prima è che in presenza di una analisi multivariata con l’inclusione di molti regressori, si dovrebbero creare e interpretare molti grafici e quindi verrebbe di molto compromesso il vantaggio dell’immediatezza visiva; il secondo è che un modello multivariato si basa sul concetto già menzionato dell’*adjustement*: il rapporto tra variabile esplicativa e variabile di risposta cambia a seconda di quali altre variabili si tengono in conto nello stesso modello, poiché ogni coefficiente stimato va interpretato come indicativo della relazione tra le due variabili mantenendo costanti i valori di tutti gli altri coefficienti.

### Riepilogo

In sintesi, una volta stimato un modello, le tecniche diagnostiche di cui bisogna avvalersi per sincerarsi che il modello sia adeguato e che si adatti ai dati sono le seguenti (alcuni punti sono stati trattati in paragrafi precedenti e a sé stanti per la loro rilevanza):

* Test per la bontà di adattamento;
* Analisi dei residui;
* Diagnosi di multicollinearità;
* Diagnosi di osservazioni particolarmente influenti;
* Correzione (eventuale) di Bonferroni per test multipli;
* Correttezza della funzione di link.

Capitolo terzo

CASO PRATICO DI REGRESSIONE LOGISTICA

BIBLIOGRAFIA

1. Fai una citazione qui [↑](#footnote-ref-1)
2. per i dettagli vedi capitolo terzo [↑](#footnote-ref-2)
3. grosso spunto da ISLR [↑](#footnote-ref-3)
4. BORRA S. - DI CIACCIO A., Statistica. Metodologia per le scienze economiche e sociali, Milano, 2014, p. 171. [↑](#footnote-ref-4)
5. LUBISCO A. - MIGNANI S. - PILLATI M., Analisi statistica multivariata. La regressione logistica (Internet), disponibile a **http://www2.stat.unibo.it/mignani/Didattica/analisideidati/logistica.pdf** (data di ultima consultazione: Marzo 2016), p.4. [↑](#footnote-ref-5)
6. *Ivi*, p.5. [↑](#footnote-ref-6)
7. *Ivi*, p.3 [↑](#footnote-ref-7)
8. Per la scelta di rappresentare la funzione logistica in questa forma si è preso spunto da KLEINBAUM D – KLEIN MITCHEL, *Logistic regression. A self learning text*, New York, 2010, p.5. [↑](#footnote-ref-8)
9. In epidemiologia la si potrebbe definire come la combinazione dei fattori di rischio sulla probabilità di insorgenza di una malattia, cfr. KLEINBAUM D – KLEIN MITCHEL, *Op. cit*, pp. 6-7. [↑](#footnote-ref-9)
10. BOTTARELLI E. – OSTANELLO F., *Epidemiologia. Teoria ed esempi di medicina veterinaria*, Milano, 2011, p. 65 ss. [↑](#footnote-ref-10)
11. KLEINBAUM D – KLEIN MITCHEL, *Op. cit*, p. 18. [↑](#footnote-ref-11)
12. La funzione di link è una funzione che trasforma in modo tale da legare la variabile dipendente a una combinazione lineare dei predittori. Cfr. AGRESTI A., *Foundations of linear and generalized linear models* (Internet), New York, 2015, par. 1.1. [↑](#footnote-ref-12)
13. LUBISCO A. - MIGNANI S. - PILLATI M., *Op. cit.*, p. 11. [↑](#footnote-ref-13)
14. HASTIE T – TIBSHIRANI R – FRIEDMANJ., *The elements of statistical learning*, New York, 2009, p. 31. [↑](#footnote-ref-14)
15. LUBISCO A. - MIGNANI S. - PILLATI M., *Op. cit.*, p. 15. Asintoticamente (cioè a dimensione del campione crescente) uno stimatore è corretto se il suo valore atteso (empiricamente, la media di varie stime basate su vari campioni) è uguale al parametro da stimare ed è efficiente se (essendo corretto) ha il minor errore standard possibile; Cfr. BORRA S. - DI CIACCIO A., *Op. cit*, p. 267 ss. [↑](#footnote-ref-15)
16. CAFFO B., *Regression models for data science in R*, Baltimore, 2015, p. 53 ss, dove ci sono anche curiosi esempi del *Simpon’s Paradox*, cioè il cambio di segno del coefficiente di un parametro tenendo conto di altre variabili oppure escludendole.Test di ipotesi sui coefficienti del modelloanti. deve essere interpretato numericamente come il valore che assume non troppo c [↑](#footnote-ref-16)
17. CAFFO B., *Statistical inference for data science*, Baltimore, 2015, p. 79 ss. [↑](#footnote-ref-17)
18. Cfr KLEINBAUM D – KLEIN MITCHEL, *Op. cit*, p. 135. [↑](#footnote-ref-18)
19. LUBISCO A. - MIGNANI S. - PILLATI M., *Op. cit.*, p. 23. [↑](#footnote-ref-20)
20. KLEINBAUM D – KLEIN MITCHEL, *Op. cit*, p. 139. [↑](#footnote-ref-21)
21. CAFFO B., Statistical inference for data science, cit., p. 81. [↑](#footnote-ref-22)
22. KLEINBAUM D – KLEIN MITCHEL, Op. cit, p. 280. [↑](#footnote-ref-23)
23. Ibidem. [↑](#footnote-ref-24)
24. Sul fatto che la correzione di Bonferroni si possa usare anche nella regressione multipla Cfr. MUNFROM D. – PERRETT J. – SCHAFFER J. – PICCONE A. – ROOZEBOOM M., Bonferroni adjustments in tests for regression coefficients (Internet), 2006, disponibile su **https://www.amstat.org/meetings/jsm/2008/onlineprogram/index.cfm?fuseaction=abstract\_details&abstractid=301702** (data di ultima consultazione: Marzo 2016), p.1 ss. [↑](#footnote-ref-25)
25. ESPA G. – MICCIOLO R. – CANAL L., Ricerca con R. Metodi di inferenza statistica, Milano, 2013, p. 145. [↑](#footnote-ref-26)
26. ALLISON P., *Measures of Fit for Logistic Regression* (Internet), 2014, disponibile su **http://statisticalhorizons.com/wp-content/uploads/GOFForLogisticRegression-Paper.pdf** (data di ultima consultazione: Marzo 2016), p.1 ss. Buona parte di questo paragrafo si ispira questo lavoro del Professor Allison della University of Pennsylvaina. [↑](#footnote-ref-27)
27. Vedi paragrafo . [↑](#footnote-ref-28)
28. LUBISCO A. - MIGNANI S. - PILLATI M., *Op. cit.*, p. 20. [↑](#footnote-ref-29)
29. HOSMER D. – LEMESHOW S., *Applied logistic regression* (Internet), New York, 2013, par. 5.2.1. [↑](#footnote-ref-30)
30. Lo sviluppo della formula per il residuo di *Pearson* deve molto a LUBISCO A. - MIGNANI S. - PILLATI M., *Op. cit.*, p. 20 ss. [↑](#footnote-ref-31)
31. Cfr. KLEINBAUM D – KLEIN MITCHEL, *Op. cit*, p. 305 ss. [↑](#footnote-ref-32)
32. HOSMER D. – LEMESHOW S., *Op. cit.*, par. 5.2.1. [↑](#footnote-ref-33)
33. ALLISON P., *Op. cit.*, p. 4. [↑](#footnote-ref-34)
34. Cfr. KLEINBAUM D – KLEIN MITCHEL, *Op. cit*, pp. 312-318, da cui si prende largamente spunto perla trattazione di questo sottoparagrafo. [↑](#footnote-ref-35)
35. HOSMER D. – LEMESHOW S., A goodness-of-fit test for the multiple logistic regression model, in Communications in Statistics, 1980, pp. 1043-1069. [↑](#footnote-ref-37)
36. ALLISON P., Logistic regression using SAS: theory and application (Internet), Cary, 2012, par. 3.6 [↑](#footnote-ref-38)
37. ALLISON P., *Why I don’t trust Lemeshow Test for Logistic Regression* (Internet), disponibile su **http://statisticalhorizons.com/hosmer-lemeshow,** (data di ultima consultazione: Marzo 2016).. [↑](#footnote-ref-39)
38. HOSMER D. – BURSAC Z. – GAUSS C. – WILLIAMS D., *Purposeful selection of variables in logistic regression* (Internet), Amherst, 2008, disponibile su **http://www.readcube.com/articles/10.1186%2F1751-0473-3-17** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-40)
39. KLEINBAUM D – KLEIN MITCHEL, *Op. cit*, p 244 ss. [↑](#footnote-ref-41)
40. Cfr. CURINI L., *Regression with a binary dipendent variable: logistic regression diagnostic* (Internet), Milano, 2014, disponibile su **http://www.sociol.unimi.it/docenti/curini/Multivariate%20PHD%202015/Logit%20&%20Probit%20Diagnostic/PhD%202015%20Probit%20Logit%20Diagnostic.pdf** (data di ultima consultazione: Marzo 2016), p. 1. [↑](#footnote-ref-42)
41. Cosa che, se non accade, può generare problemi di *overdispersion*, cfr. Penn State University, *Analysis of discrete data* (Internet), disponibile su **https://onlinecourses.science.psu.edu/stat504/node/162** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-43)
42. Cfr. CURINI L., *Op. cit.*, p. 4. [↑](#footnote-ref-44)
43. KABACOFF R., *R in action. Data analysis and graphics with R* (Internet), New York, 2015, par. 13.1 [↑](#footnote-ref-45)
44. PARDOE I., COOK D., *A Graphical Method for Assessing the Fit of a Logistic Regression Model* (Internet), 2002, disponibile su **http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00276.pdf** (data di ultima consultazione: Marzo 2016), pp. 2-3. [↑](#footnote-ref-46)
45. Come ALLISON P., *Logistic regression using SAS: theory and application,* cit., KLEINBAUM D – KLEIN MITCHEL, *Op. cit.*, HOSMER D. – LEMESHOW S., *Applied logistic regression*, cit. [↑](#footnote-ref-47)
46. CAFFO B., Regression models for data science in R, cit., pp. 96-97. [↑](#footnote-ref-48)
47. KLEINBAUM D – KLEIN MITCHEL, Op. cit., p. 271. [↑](#footnote-ref-49)
48. PARDOE I., COOK D., Op. cit., p1. [↑](#footnote-ref-50)
49. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.*, p. 271. [↑](#footnote-ref-51)
50. *Ivi*, p. 273. [↑](#footnote-ref-52)
51. CURINI L., *Op. cit.*, p. 5. [↑](#footnote-ref-53)
52. Il dataset utilizzato nel capitolo 3 contiene molte variabili quantitative. [↑](#footnote-ref-54)
53. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.*, p. 275. [↑](#footnote-ref-55)
54. KOVAL J. et alia, *Assessing the impact of potentially influential observations in weighted logistic regression* (Internet**),** disponibile su **http://www.statcan.gc.ca/pub/12-002-x/2015001/article/14147-eng.htm** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-56)
55. Penn State University, *Analysis of discrete data* (Internet), disponibile su **https://onlinecourses.science.psu.edu/stat504/node/161** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-57)