****

***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 su processi bancari**

Relatore:

Ch.mo Prof. **Pasquale Sarnacchiaro**

Candidato:

Americo Costantini

Anno Accademico 2016/2017

INTRODUZIONE

# Obiettivo dell’elaborato

# Notazione matematica

CAPITOLO PRIMO

TITOLO DEL CAPITOLO

wcnipwbifhvbfiebvihefbvihfbvfiehvbefihbvqeifhvbefihbveifhbvqiefbveihfbviefhbviefhqbviqfehvbihefbvihv

# Paragrafo

àldkwncdljwnvjrnvfjnvfjlnvjlfnvlajfnvfdjnvdljvnlfjvnfljvnfljvnfldjnjlvndfaljvndfljvndfljvndfljvndfljvnldfjnv

# Secondo paragrafo

### Terzo sotto

### Sottoparagrafo

### Sottoparagrafo

### Sottoparagrafo

# Paragrafi

### sottonmhmgmjg

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 *p* variabili indipendenti, l’uscita di una variabile dicotomica *Y* che può assumere soltanto due valori:

*Y*=

Un tipico approccio per la predizione è quello di stimare la probabilità che *Y* assuma il valore *1* in base ai valori del vettore delle variabili indipendenti, che denoteremo con ***X***. Come noto, la probabilità che un evento accada può assumere soltanto valori tra *0* e *1[[1]](#footnote-1)*, quindi a questo scopo la regressione lineare è inadatta, in quanto modello volto a predire il valore di una variabile quantitativa che non può essere sottoposta a una restrizione di intervallo di questo tipo[[2]](#footnote-2).

Inoltre una variabile casuale dicotomica ha una distribuzione di Bernoulli (o binomiale), che prevede una varianza pari a , dove è la probabilità che l’evento accada. Tale varianza dipende da un valore non costante (perché dipendente dal valore dei regressori), e quindi non è costante. Se ne conclude 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 delle variabili indipendenti[[3]](#footnote-3).

Insomma, l’obiettivo è quello di individuare una funzione che a partire dal vettore di valori delle variabili indipendenti ***X*** fornisca una probabilità che *Y = 1* come valore atteso di *Y*:

[[4]](#footnote-4)

### La funzione logistica

Una funzione adatta allo scopo di legare ***X*** a *Y* è la cosiddetta funzione logistica[[5]](#footnote-5):

dove *z* 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 *(0; 0,5)*, dove interseca l’asse delle ascisse. Tale forma assunta dalla funzione logistica si chiama sigmoide[[6]](#footnote-6). Una delle più importanti proprietà di questa funzione è che qualunque valore assuma *z*, il valore di *f(z)* sarà sempre compreso nell’intervallo [0; 1]. Se z tende a ∞ allora il valore di uscita tenderà a 0, in caso contrario a 1. Questo è esattamente quello che ci aspettiamo da una funzione la cui uscita deve essere una probabilità. *z* può quindi essere visto come una sintesi dei vari fattori che determinano l’accadere di un evento[[7]](#footnote-7).

### La trasformazione in relazione lineare[[8]](#footnote-8)

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

Sia la probabilità che un evento accada in dipendenza di una certa combinazione delle variabili indipendenti, allora:

ODDS =

L’ODDS è il rapporto tra probabilità complementari[[9]](#footnote-9) ed esprime quanto le probabilità a favore sono maggiori (o minori) di quelle contro.

L’ODDS è una misura che che si muove nell’intervallo (0; ∞). Tuttavia, se all’ODDS applichiamo il logaritmo naturale il risultato è una quantità che va da (-∞,∞), il cosiddetto LOGIT[[10]](#footnote-10). E, cosa molto più importante, è matematicamente dimostrabile[[11]](#footnote-11) questa equazione:

= *z*

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 *p* variabili indipendenti e una intercetta, *p + 1* parametri.

# La stima e l’interpretazione dei coefficienti

Ricorda la differenza tra coefficiente reale e stimato

### 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 vige 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 Y 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. Denoteremo con il coefficiente stimato della j-esima variabile, e con la predizione della variabile indipendente *Y* relativa alla i-esima osservazione, ottenuta con l’uso dei coefficienti stimati.

### La stima tramite il metodo della massima verosimiglianza

Esistono varie tecniche per stimare i coefficienti di un modello. Nella regressione lineare il metodo più usato è il famoso OLS, *ordinary least squares*, il metodo dei minimi quadrati. Nella regressione logistica, poiché non vige l’ipotesi 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 uscita. La massima verosimiglianza è una tecnica ampiamente usata, e invero la stima tramite OLS ne è un sottoinsieme[[14]](#footnote-14). I dettagli matematici della funzione di massima verosimiglianza sono aldilà dello scopo di questo elaborato, basti aggiungere che trasformandosi la sua massimizzazione in un set di q equazioni per p parametri incogniti, la si può risolvere solo con metodi iterativi (ad esempio Fisher[[15]](#footnote-15)). Asintoticamente, sotto condizioni non particolarmente restrittive, gli stimatori di massima verosimiglianza sono corretti, normodistribuiti ed efficienti[[16]](#footnote-16).

### 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[[17]](#footnote-17). 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[[18]](#footnote-18) , e quindi il modello logistico conterrà *p + 1* 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[[19]](#footnote-19).

# 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[[20]](#footnote-20). 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 rapporto di massima verosimiglianza confronta il valore assunto dalla funzione di massima verosimiglianza per due modelli: il modello cosiddetto *full*, che contiene più parametri, e il modello ridotto, che ne contiene di meno[[21]](#footnote-21). 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 L, *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:

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. Sapendo ciò, sarà immediato verificare, dato un certo livello di significatività del test (usualmente ), quanto probabile sia sotto l’ipotesi nulla osservare un valore di G diverso da zero. Se questa probabilità è inferiore allivello di significatività, l’ipotesi nulla viene rigettata e i parametri inclusi nel modello.

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[[22]](#footnote-23). Basterà quindi utilizzare la statistica test *Z*[[23]](#footnote-24)per rigettare o meno l’ipotesi nulla che il coefficiente sia diverso da zero.

Per completezza, andrebbe aperta una digressione sull’incremento della probabilità di errore 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 ad . La probabilità di non commettere nessun errore è quindi pari a 0.95. Questa probabilità è chiamata FWER (*family wise error rate*)[[24]](#footnote-25) e si può esprimere come Pr(rigettare almeno una ipotesi nulla | tutte le ipotesi nulle sono vere). Matematicamente si esprime come:

FWER = [[25]](#footnote-26)

dove T = 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 [[26]](#footnote-27). Il rischio è rendere troppo difficile il rigetto dell’ipotesi nulla [[27]](#footnote-28).

# 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[[28]](#footnote-29). Come misure del potere predittivo di una regressione logistica sono utilizzate, tra le altre, differenti versioni di R quadro e l’AUC (*area under the curve*) [[29]](#footnote-30). È fuori dallo scopo di questo elaborato approfondire le argomentazioni a favore di una certa versione dell’R quadro piuttosto che un’altra (ne sono state individuate almeno 12, di cui le più importanti qualla 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.

### La devianza

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 *n* 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.

Nella regressione lineare la bontà di adattamento 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 non è associato alla i-esima osservazione, ma al j-esimo gruppo di osservazioni accomunate dalla medesima combinazione di valori delle varabili indipendenti (*covariate pattern*) [[30]](#footnote-31).

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

Sia *j* il numero di *covariate pattern* delle variabili indipendenti osservate per le *n* unità statistiche; sia il numero di unità statistiche presenti in ogni j-esimo covariate pattern *k=1,.., j*; siano e rispettivamente il numero osservato e il numero stimato di unità statistiche appartenenti al k-esimo *covariate pattern* per cui *Y = 1*, 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 si distribuisce come una variabile casuale con *j -(p+1)* gradi di libertà (sotto l’ipotesi nulla che il modello stimato si adatta ai dati) [[31]](#footnote-32).

Per spiegare il residuo di deviance utilizzeremo invece un altro approccio [[32]](#footnote-33). 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 [[33]](#footnote-34) che la devianza è una statistica test risultante dalla differenza tra il valore di massima verosimiglianza del modello saturato (il modello che include tutte le variabili e i possibili effetti di interazione tra loro) 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 *j -(p+1)* 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 ci indurrebbe a rigettare l’ipotesi 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) possa performare ancora meglio in termini di residui[[34]](#footnote-35).

Ora, per le due statistiche sulla bontà di adattamento finora viste, la statistica chi-quadrato 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 inferiore rispetto alla grandezza del campione [[35]](#footnote-36). Con molte variabili continue e non categoriche 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 [[36]](#footnote-37), per ovviare ai limiti di statistiche test che si basato sul numero di *g*.

Hosmer e Lemeshow hanno proposto il raggruppamento delle osservazioni in base ai valori della variabile indipendente predetti dal modello di regressione logistica. In particolare, i valori predetti sono ordinati dal più piccolo al più grande, e poi separati in diversi 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 variabile casuale viene poi utilizzata per confrontare i conteggi osservati con conteggi attesi. I gradi di libertà è il numero di gruppi meno 2. P-valori inferiori al livello di significatività suggeriscono il rifiuto del modello [[37]](#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 [[38]](#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*[[39]](#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 predittività, bisogna avvalersi di misure del potere predittivo: in tal caso una *cross-validation* che seleziona la combinazione di variabili con l’AUC maggiore andrà più che bene. 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 uscita alla luce dell’influenza di alcuni fattori di rischio. In quel caso un approccio possibile è la costruzione di un modello gerarchico che includa la variabile di esposizione, alcune variabili di controllo e altre di interazione [[40]](#footnote-41). È aldilà 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 [[41]](#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 [[42]](#footnote-43) e che tutte le variabili significative siano state incluse (la cosiddetta assunzione di specificazione del modello). Assume inoltre che tra LOGIT e variabili dipendenti ci sia una relazione lineare. Il primo è un problema di raccolta dei dati (serie storiche piuttosto che campionamento a grappoli [[43]](#footnote-44)), il secondo spesso si affronta utilizzando un test per la funzione di link: si verifica se ulteriori predittori inseriti nel modello, quadrati dei predittori già inseriti, siano significativi. Se lo sono, il modello è mal specificato[[44]](#footnote-45).

La trattazione seguente è frutto di ricerche che hanno sofferto di na 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 [[45]](#footnote-46).

### 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, c’è un effetto sistematico che agisce e che può essere spiegato includendo nel modello qualche regressore non utilizzato, rimanendo così solo il residuo dovuto alla stima dei parametri e al fatto che nessuna funziona descrive perfettamente un fenomeno senza includere un termine di errore, di cui il residuo è realizzazione.

Nella regressione logistica il problema è che, innanzitutto, di residui si può parlare solo per dati raggruppati, dove *j < n*, 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 possa essere sospetto oppure no [[46]](#footnote-47). Chi scrive nota una certa carenza di letteratura sull’argomento, essendo anche i manuali più noti sprovvisti della trattazione di questo argomento [[47]](#footnote-48). 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 [[48]](#footnote-49). 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); 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 collinearità 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*) [[49]](#footnote-50). I due indici sono stati creati per il modello di regressione lineare, ed è poi stato adattato anche alla regressione logistica [[50]](#footnote-51). Spiegare come vengono calcolati i due indici richiederebbe tecniche di algebra matriciale [[51]](#footnote-52), che è aldilà 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 30, allora se almeno due valore della VDP legata al CNI sono maggiori di 0.5, allora c’è un problema di collinearità [[52]](#footnote-53). 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[[53]](#footnote-54)), 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 [[54]](#footnote-55). In un modello multivariato l’inspezione grafica è purtroppo insufficiente come strumento diagnostico, servono 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 [[55]](#footnote-56) tratteremo solo misure diagnostiche che non lo prevedono.

Tecnicamente una osservazione è influente 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 [[56]](#footnote-57).

I *dfbetas* o *delta-betas* sono una misura del cambiamento nel valore stimato d tutti i *p* coefficienti qualora venga rimossa una certa osservazione. Tale misura coincide con la differenza tra le due stime diviso l’errore standard dello stimatore. Ad ogni osservazione vengono quindi legati tutti i *p* 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 [[57]](#footnote-58). 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 specifizazione 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 *j* non è molto minore di *n* [[58]](#footnote-59).

### 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 citazione dal manuale di statistica. [↑](#footnote-ref-1)
2. Fai citazione dal pdf di rimini almeno [↑](#footnote-ref-2)
3. citare libro che parla di questo requisito [↑](#footnote-ref-3)
4. rifai citazione da rimini su non varianza anche dell’errore [↑](#footnote-ref-4)
5. citazione da p.5 di kleineibaum [↑](#footnote-ref-5)
6. fare una citazione sulla sigmoide. [↑](#footnote-ref-6)
7. In epidemiologia come fattori di rischio. Cita klein [↑](#footnote-ref-7)
8. di che questo approccio che parte dalla funzione logistica e non da quella lineare lo hai scelto da klein [↑](#footnote-ref-8)
9. cita epidemiologia di bottarelli [↑](#footnote-ref-9)
10. nota su sito di matematicamente sulle proprietà dei logaritmi [↑](#footnote-ref-10)
11. devo trovare un libro da citare al riguardo [↑](#footnote-ref-11)
12. fare citazione sul link dei GLM. [↑](#footnote-ref-12)
13. Cita rimini [↑](#footnote-ref-13)
14. Cita ISLR [↑](#footnote-ref-14)
15. trovare una citazione su safari [↑](#footnote-ref-15)
16. citazione da rimini [↑](#footnote-ref-16)
17. fai una citazione [↑](#footnote-ref-17)
18. cita da qualche parte [↑](#footnote-ref-18)
19. cita regmod di caffo Test di ipotesi sui coefficienti del modelloanti. deve essere interpretato numericamente come il valore che assume non troppo c [↑](#footnote-ref-19)
20. qua una citazione da qualche libro di statistica, magari quello di mine, sul test di ipotesi. [↑](#footnote-ref-20)
21. Cfr klein [↑](#footnote-ref-21)
22. klein [↑](#footnote-ref-23)
23. riferimento a manuale di statistica [↑](#footnote-ref-24)
24. cita non da klein, magari da caffo [↑](#footnote-ref-25)
25. klein 280 [↑](#footnote-ref-26)
26. che si possa usare anche nella regressione multipla cfr <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.490.7640&rep=rep1&type=pdf> [↑](#footnote-ref-27)
27. esistono anche altre tecniche, cita caffo. [↑](#footnote-ref-28)
28. <http://support.sas.com/resources/papers/proceedings14/1485-2014.pdf> di che buona parte del paragrafo si ispira a quello [↑](#footnote-ref-29)
29. rimanda al primo capitolo. [↑](#footnote-ref-30)
30. Hosmer 5.2.1 [↑](#footnote-ref-31)
31. cita rimini come grande fonte [↑](#footnote-ref-32)
32. klein 306 [↑](#footnote-ref-33)
33. hosmer 5.2.1 [↑](#footnote-ref-34)
34. <http://support.sas.com/resources/papers/proceedings14/1485-2014.pdf> [↑](#footnote-ref-35)
35. Klein 312-318, da cui si prende spunto per il paragrafo. [↑](#footnote-ref-36)
36. Hosmer D.W. and S. Lemeshow (1980) “A goodness-of-fit test for the multiple logistic regression model.” Communications in Statistics A10:1043-1069 [↑](#footnote-ref-37)
37. cita log reg using sas di allisan 3.6 [↑](#footnote-ref-38)
38. <http://statisticalhorizons.com/hosmer-lemeshow> [↑](#footnote-ref-39)
39. <http://www.readcube.com/articles/10.1186%2F1751-0473-3-17> [↑](#footnote-ref-40)
40. klein, p. 244 e seguenti. [↑](#footnote-ref-41)
41. <http://www.sociol.unimi.it/docenti/curini/Multivariate%20PHD%202015/Logit%20&%20Probit%20Diagnostic/PhD%202015%20Probit%20Logit%20Diagnostic.pdf> [↑](#footnote-ref-42)
42. cosa che, se non accade, può generare problemi di *overdispersion*, cfr. <https://onlinecourses.science.psu.edu/stat504/node/162> [↑](#footnote-ref-43)
43. <http://www.sociol.unimi.it/docenti/curini/Multivariate%20PHD%202015/Logit%20&%20Probit%20Diagnostic/PhD%202015%20Probit%20Logit%20Diagnostic.pdf> [↑](#footnote-ref-44)
44. <https://onlinecourses.science.psu.edu/stat504/node/161> [↑](#footnote-ref-45)
45. r in action 13.1 [↑](#footnote-ref-46)
46. <http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00276.pdf> pp 2-3. [↑](#footnote-ref-47)
47. Klein, logreg with sas e lemeshow. [↑](#footnote-ref-48)
48. Caffo pp 96 e 97 [↑](#footnote-ref-49)
49. klein 271 [↑](#footnote-ref-50)
50. <http://www.amstat.org/sections/srms/proceedings/y2002/files/JSM2002-000839.pdf> p1. [↑](#footnote-ref-51)
51. <http://www.statcan.gc.ca/pub/12-001-x/2012002/article/11757-eng.pdf> p 5 [↑](#footnote-ref-52)
52. klein 271 [↑](#footnote-ref-53)
53. klein 273 [↑](#footnote-ref-54)
54. <http://www.sociol.unimi.it/docenti/curini/Multivariate%20PHD%202015/Logit%20&%20Probit%20Diagnostic/PhD%202015%20Probit%20Logit%20Diagnostic.pdf> p 5 [↑](#footnote-ref-55)
55. il dataset utilizzato nel capitolo 3 contiene molte variabili quantitative [↑](#footnote-ref-56)
56. klein 275 [↑](#footnote-ref-57)
57. <http://www.statcan.gc.ca/pub/12-002-x/2015001/article/14147-eng.htm> . [↑](#footnote-ref-58)
58. <https://onlinecourses.science.psu.edu/stat504/node/161> [↑](#footnote-ref-59)