****

***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

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 indipende *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

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 [[30]](#footnote-31).

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 [[31]](#footnote-32). 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 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 *n o 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 da testare, 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[[33]](#footnote-34).

In realtà l’approssimazione della statistica test devianza alla variabile casuale si verifica se e solo se il numero dei gruppi di soggetti accomunati dalla stessa combinazione di valori delle variabili indipendenti (gruppi che chiamiamo *j*) è inferiore rispetto alla grandezza del campione [[34]](#footnote-35). Con molte variabili continue e non categoriche presenti nel dataset, questo assunto viene facilmente violato e il test non è più affidabile.

### 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-36), 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 [[36]](#footnote-37). 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-38).

Devo approfondire un po’ di più i residui, quelli di pearson e quelli di devianza e l’AIC. Troppo poco così.

# 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-39) , 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 [[39]](#footnote-40). 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 [[40]](#footnote-41). 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-42) e che tutte le variabili significative siano state incluse (la cosiddetta assunzione di specificazione del modello). Il primo è un problema di raccolta dei dati, 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.

### Analisi dei residui

<https://onlinecourses.science.psu.edu/stat504/node/161>

<http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00276.pdf>

<http://www.sociol.unimi.it/docenti/curini/Multivariate%20PHD%202015/Logit%20&%20Probit%20Diagnostic/PhD%202015%20Probit%20Logit%20Diagnostic.pdf>

ultima slide di rimini

### Multicollinearità

### Osservazioni influenti

Capitolo terzo

Titolo capitolo terzo

# Paragrafo 1

### Sottop

lkcjwndournornforuvnreouvneouvneouvnouvntouvntouvntouvntountouvntrouvnrtouvnrtouvntorun[[42]](#footnote-43)

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. klein 306 [↑](#footnote-ref-32)
32. hosmer 5.2.1 [↑](#footnote-ref-33)
33. <http://support.sas.com/resources/papers/proceedings14/1485-2014.pdf> [↑](#footnote-ref-34)
34. Klein 312-318, da cui si prende spunto per il paragrafo. [↑](#footnote-ref-35)
35. 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-36)
36. cita log reg using sas di allisan 3.6 [↑](#footnote-ref-37)
37. <http://statisticalhorizons.com/hosmer-lemeshow> [↑](#footnote-ref-38)
38. <http://www.readcube.com/articles/10.1186%2F1751-0473-3-17> [↑](#footnote-ref-39)
39. klein, p. 244 e seguenti. [↑](#footnote-ref-40)
40. <http://www.sociol.unimi.it/docenti/curini/Multivariate%20PHD%202015/Logit%20&%20Probit%20Diagnostic/PhD%202015%20Probit%20Logit%20Diagnostic.pdf> [↑](#footnote-ref-41)
41. cosa che, se non accade, può generare problemi di *overdispersion*, cfr. <https://onlinecourses.science.psu.edu/stat504/node/162> [↑](#footnote-ref-42)
42. [↑](#footnote-ref-43)