****

***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

### sotto

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 *p* la probabilità che un evento accada,

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:

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

### Accenni di teoria della stima

### La stima tramite il metodo della massima verosimiglianza

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

### Test di ipotesi sui coefficienti del modello

In nota mettere bonferroni (vedi email)

Capitolo terzo

Titolo capitolo terzo

# Paragrafo 1

### Sottop

lkcjwndournornforuvnreouvneouvneouvnouvntouvntouvntouvntountouvntrouvnrtouvnrtouvntorun[[13]](#footnote-13)

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. [↑](#footnote-ref-13)