UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Scienze Economiche e Statistiche

Corso di Laurea Triennale in Economia e Commercio

Prova finale in

Metodi statistici per l'economia

Il plebiscito cileno del 1988: un'analisi statistica in R

Relatore:

Ch.mo Prof. Cira Perna

Candidato:

Sabato **********

matr. **********

Anno Accademico 2017/2018

Indice

1.Introduzione

Il referendum indetto il 5 Ottobre 1988 in Cile fu uno spartiacque per il popolo cileno. Quest'ultimo, come previsto dalla Costituzione cilena, fu chiamato a decidere se affidare ad Augusto Pinochet un'ulteriore mandato sessennale da Presidente della Repubblica. Nelle norme transitorie della Costituzione era infatti stabilito che fosse effettuato il plebiscito al termine del primo mandato presidenziale. Il "Si" avrebbe confermato Pinochet, il "No" avrebbe portato a nuove elzioni.

La Facultad Latinoamericana de Ciencias Sociales (FLACSO) nel bimestre Aprile-Maggio del 1988 ha somministrato un questionario a circa 2700 cileni, al fine di individuare le preferenze dei votanti, ma soprattutto i fattori più influenti sull'orientamento politico degli intervistati.

Il primo capitolo di questo lavoro metterà in evidenza le principali differenze tra il modello lineare e quello logistico e i diversi ambiti applicativi. Si procederà poi a illustrare brevemente il metodo di stima per il modello logistico e successivamente le statistiche più utilizzate in pratica per la verifica del modello e delle sue previsioni. Nella seconda parte dello stesso capitolo verrà affrontato brevemente il caso della regressione logistica multipla e delle relative peculiarità rispetto alla regressione semplice.
Dopo aver illustrato i pilastri teorici di tale metodo si procederà allo studio applicativo sui dati.

In primis lo studio analizzerà il quadro socio-economico nel quale si sarebbero svolte le votazioni grazie agli strumenti propri della statistica descrittiva. In particolare si utilizzeranno boxplot e istogrammi condizionati rispetto a vari fattori. Lo scopo di un'analisi preliminare di questo tipo è quello di poter selezionare le variabili più adatte alla costruzione di un modello e a sintetizzare chiaramente il contenuto dei dati utilizzati.

Nel corso dell'analisi descrittiva sono emerse numerose asimmetrie storicamente note, quali, ad esempio, la forte disuguaglianza reddituale e culturale tra le diverse fasce di popolazione, soprattutto nelle aree scarsamente urbanizzate.

Queste asimmetrie hanno sicuramente giocato un ruolo fondamentale nel processo decisionale e il fine principale di questo studio sarà quello di individuare, attraverso una rigorosa analisi statistica, quali tra questi fattori abbiano avuto un'influenza maggiore sulle intenzioni di voto del popolo cileno.

Sebbene lo studio non abbia pretese di esaustività riguardo le dinamiche sottostanti i processi decisionali collettivi, verrà presentata una metodologia semplice ed efficace per l'analisi di situazioni di questo tipo.

Sulla base delle evidenze riscontrate nell' analisi descrittiva effettuata in precedenza verrà selezionato un pool di variabili potenzialmente candidate a spiegare gran parte della variabilità delle intenzioni di voto.

Successivamente verrà stimato il modello di regressione logistica semplice. Si porrà particolare attenzione nel trattare con minuzia di particolari l'interpretazione delle stime fornite dal modello, soprattutto con riguardo ai coefficienti e ai relativi test affinchè i dati presentati possano essere quanto più fruibili anche per i non addetti ai lavori. Attraverso la costruzione di una tabella di contingenza e della relativa curva ROC verranno poi confrontati i valori stimati dal modello con quelli effettivamente osservati per testare la capacità predittiva del modello semplice.

Infine verrà modellata una regressione logistica multipla attraverso un particolare algoritmo che seleziona automaticamente il modello migliore confrontando tutti i possibili modelli e i relativi valori di indicatori quali l'AIC. Verranno poi effettuati i test applicati in precedenza per il modello semplice. Si presterà particolare attenzione anche in questo caso a rendere quanto più chiara possibile l'interpretazione dei dati e delle stime del modello.

Tutte le operazioni statistiche e i grafici che verranno presentati in questo lavoro sono ottenuti attraverso l'utilizzo del software open source R. Nell'appendice sarà possibile trovare i codici integralmente riportati dei comandi utilizzati.

2. Il modello di regressione logistica

I modelli di regressione sono diventati una componente integrante dell'analisi dei dati, soprattutto se l'obiettivo è descrivere le relazioni tra una variabile dipendente e una o più variabili indipendenti. Nelle scienze sociali è frequente che la variabile dipendente sia dicotomica, ad esempio occupato-disoccupato, celibe-nubile…

L'obiettivo di fondo dello statistico è lo stesso rispetto a qualsiasi altro approccio, e cioè trovare un modello che si adatti meglio ai dati e che possa descrivere la relazione tra una variabile dipendente e un set di variabili detti predittori. In questi specifici casi,però, il modello di regressione logistica si rileva più utile e corretto del classico modello di regressione lineare, e nel seguito verranno spiegati in dettaglio i motivi di questa particolare scelta.

Il modello, in forma generica, sarà così costituito:

$$p_i = \Pr(y_i = j) = F(x_i'\beta)$$

y_i è la variabile dipendente dicotomica, β è un vettore k-dimensionale di parametri incogniti, x_i è il vettore colonna di variabili esplicative e p_i rappresenta la probabilità associata all' iesimo individuo.

$$y_i = F(x_i'\beta) + \varepsilon_i$$

Nella regressione lineare semplice i dati sono composti da coppie di osservazioni. La variabile esplicativa (X) è numerica e verrà usata per stimare il valore della variabile dipendente (Y), anch'essa numerica. Il modello ha due componenti principali: una riguardante il valore medio di Y dato X e un'altra riguardante gli errori (ε) e la loro distribuzione probabilistica. Quest'ultima componente descrive lo scostamento dei valori osservati da quelli medi.

Di seguito indicheremo le ipotesi su cui si basa un modello di regressione lineare:

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- $E(\varepsilon_i) = 0$
- $Var(\varepsilon_i) = \sigma^2$
- $Cov(\varepsilon_i, \varepsilon_j) = 0$
- $\varepsilon_i \sim$ Normale

La prima ipotesi esprime y come funzione lineare di x, con l'aggiunta del termine di errore ε. Dato $E(y_i|x_i) = F(\mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$ l'equazione del modello diventerà:

$$y_i = x_i'\beta + \varepsilon_i$$

Tutte le altre ipotesi si riferiscono alla distribuzione di ε . La seconda ipotesi è molto importante poiché il valore atteso degli errori non varia al variare di x. Ciò implica che ε e x sono incorrelati.

La terza ipotesi è l' *omoschedasticità* degli errori. Ciò significa che la varianza di ε è costante per tutte le osservazioni.

La quarta ipotesi presuppone l'incorrelazione tra gli errori. Se tutte e cinque le ipotesi sono verificate, la stima OLS dei parametri è *non distorta* e ha varianza minima.

Cominciamo col dire che sia il modello di regressione lineare che quello logistico appartengono alla più grande famiglia dei *generalized linear model* (GLM).

Tutti i modelli GLM hanno tre componenti principali:

RANDOM COMPONENT

La componente random di un GLM indentifica la variabile di risposta Y e seleziona per quest'ultima una distribuzione di probabilità. Date n osservazioni su Y, indicate con $(Y_1, Y_2, \ldots, Y_n)$, la teoria dei GLM assume che tali osservazioni siano indipendenti. Nel caso di regressione logistica Y è binaria. Più in generale, ogni Y_i potrebbe essere un numero di successi su un certo numero di prove. In questo caso si assume che Y segua una distribuzione binomiale.

SYSTEMATIC COMPONENT

La componente sistematica di un GLM specifica le variabili esplicative. Quindi la combinazione lineare delle variabili esplicative e dei paramentri sarà definita *predittore lineare.*

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

LINK FUNCTION

Posto μ=E(Y), la funzione link è una funzione g(°) che lega μ al predittore lineare descritto in precedenza, da qui il nome link function. Essa infatti collega la componente sistematica e quella random del modello.

La link function più semplice è g(μ)= μ, cosiddetta funzione di link *identità*. Altri tipi di link function consentono a μ di essere legato ai predittori non solo linearmente. Infatti nel modello logistico la link function sarà proprio la funzione logit. Un modello GLM che utilizza come link function la funzione logit è detto, appunto, modello di regressione logistica.

$$g(\mu) = \frac{\mu}{(1-\mu)}$$

Nel caso di una variabile Y dicotomica, la sua distribuzione sarà composta dalla probabilità di successo P(Y=1)=π e dalla probabilità opposta, cioè l'insuccesso P(Y=0)=(1-π). Il suo valore atteso sarà E(Y)= π. Nel caso di n osservazioni indipendenti, il numero di successi avrà distribuzione binomiale B(n, π).

Qualora utilizzassimo il metodo lineare per stimare un modello con variabile dicotomica Y, otterremo la seguente funzione:

$$\pi(x) = \alpha + \beta x$$

Il modello così ottenuto sarà chiamato di probabilità lineare, poiché la probabilità di successo varia linearmente al variare di x.

Il parametro β rappresenta la variazione della probabilità per ogni variazione unitaria in x. Questo modello GLM ha una componente random binomiale ma una funzione link identità.

Sebbene il modello sia semplice, esso presenta dei difetti strutturali. Infatti, mentre la probabilità di un evento deve necessariamente essere compresa nell'intervallo [0,1], la funzione lineare può assumere anche valori esterni a questo intervallo e questo inconveniente conduce invitabilmente a delle problematiche.

Data la natura binaria della variabile Y l'assunzione di varianza costante degli errori non sarebbe però valida, e con lei anche quella di normalità degli errori verrebbe rigettata.

$$y_i = 1 \rightarrow \varepsilon_i = 1 - x_i'\beta \rightarrow \Pr(\varepsilon_i = 1 - x_i'\beta) = p_i$$

$$y_i = 0 \rightarrow \varepsilon_i = -x_i'\beta \rightarrow \Pr(\varepsilon_i = -x_i'\beta) = 1 - p_i$$

Poiché ε_i può assumere solo due valori, risulta impossibile che possa avere una distribuzione normale (che invece è continua e non ha limiti superiori ne inferiori). Premesso ciò, bisognerà rigettare l'ipotesi di normalità della distribuzione di ε .

E' facile dimostrare che la $\mathrm{Var}(\varepsilon_i)=p_i(1-p_i)$. E' possibile notare che l'espressione ottenuta per la varianza è quella di una variabile casuale bernoulliana con probabilità p. Abbiamo appena dimostrato che una variabile dipendente dicotomica in una regressione lineare viola necessariamente l'ipotesi di omoschedasticità e quella di distribuzione normale degli errori.

Innanzitutto bisogna precisare che la violazione delle suddette ipotesi non comporta necessariamente la distorsione degli stimatori, infatti qualora le prime due ipotesi siano verificate il metodo OLS fornirà stimatori non distorti. In secondo luogo, l'assunzione di normalità degli errori può mancare quando si dispone di campioni molto numerosi;

Infatti il teorema del limite centrale ci assicura che qualsiasi sia la distribuzione di ε essa converge ad una normale per campioni sufficientemente numerosi.

La violazione dell'ipotesi di omoschedasticità però crea due ordini di problemi. In mancanza di quest'ipotesi, infatti, gli stimatori non sono più *efficienti*. Cio significa che ci saranno metodi di stima alternativi che presentano uno standard error minore.

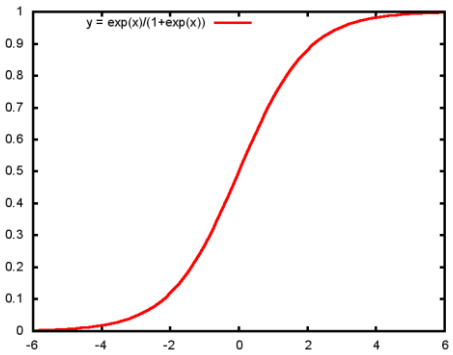
In secondo luogo, gli stimatori dello standard error non saranno più *consistenti* e questo potrebbe condurre a dei problemi riguardanti le statistiche test.

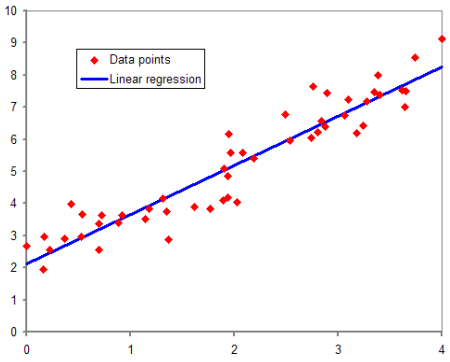
Alcuni studiosi come Huber (1967)[1] hanno ovviato a questi problemi mediante dei particolari stimatori , come il HCC(*heteroschedasticity consistent covariance estimator)* la cui analisi esula dall'obiettivo di questa tesi;

Le problematiche esposte sino a questo punto hanno condotto gli statistici verso approcci alternativi all'analisi di regressione su variabili dicotomiche. La relazione tra π(x) e x spesso è non lineare. Una variazione di x potrebbe avere meno impatto quando π è vicino a 0 o ad 1 rispetto a quando π è nel mezzo di questi valori.

Consideriamo il caso in cui un imprenditore voglia decidere se acquistare un macchinario nuovo o uno usato per la sua azienda. Indichiamo con π(x) la probabilità di acquistarlo nuovo quando il reddito della società è pari ad x.

Un aumento di € 10 000 del reddito annuo avrà meno effetto quando x= €1 000 000 (punto in cui π è vicino ad 1) rispetto a quando il reddito della società è di soli €40 000. L'immagine seguente chiarirà meglio le idee. Come si vede la curva assumerà una forma ad S, e questa relazione è sicuramente più vicina alla realtà rispetto ad una meramente lineare.





[1] (Huber,J.Peter,1967)

Sebbene un modello possa avere più variabili esplicative, per semplicità di trattazione, considereremo una sola variabile esplicativa. Successivamente proveremo a introdurre il caso multivariato.

La funzione matematica S-shaped più utilizzata nella pratica è la seguente:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

Innanzitutto questa funzione è limitata tra [0,1], questo elimina la possibilità di ottenere valori per la probabilità che siano esterni all'intervallo [0,1]. In secondo luogo, mediante la seguente trasformazione, è possibile notare un modello lineare "nascosto" nella funzione precedente.

Applichiamo la trasformazione *logit* al modello precedente.

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \alpha + \beta x$$

Il modello logistico potrà essere definito come un caso specifico di GLM. La componente random avrà distribuzione binomiale e la funzione link sarà la funzione logit $\ln[\pi/(1-\pi)]$, indicata con $\text{logit}(\pi)$. I modelli di regressione logistica vengono spesso chiamati *logit models* .

E' facile notare che sebbene π appartenga al solo intervallo [0,1], la funzione logit può assumere qualsiasi valore reale. Questo è di particolare importanza poiché i numeri reali sono anche il range di un predittore lineare che è componente fondamentale di un GLM. Difatti un modello logit non soffre dei problemi che riguardano il modello di probabilità lineare descritto in precedenza.

La regressione logit ha una forma simile ad una funzione di ripartizione F(x), in particolare si assume che la funzione di ripartizione scelta sia quella di una distribuzione logistica con due parametri.

La funzione di densità di una variabile casuale logistica sarà:

$$\lambda(\eta) = \Lambda(\eta)[1 - \Lambda(\eta)]$$

La v.c. logistica ha distribuzione molto simile alla normale standardizzata ed è leptocurtica, cioè comprende una massa di probabilità maggiore nelle code rispetto alla v.c. normale.

E' possibile notare un'altra interessante caratteristica e cioè che il log-odds ratio è lineare nelle variabili e nei parametri, infatti:

$$\ln[\Lambda(\eta)] = \ln\left[\frac{\exp(\eta)}{1 + \exp(\eta)}\right] = \eta - \ln(1 + e^{\eta})$$

$$\ln\left[1 - \Lambda(\eta) = \ln\left[\frac{1}{1 + \exp(-\eta)}\right] = -\ln(1 + e^{-\eta})\right)$$

Visto che il log-odds ratio è esprimibile come il logaritmo del rapporto tra la probabilità che y=1 e y=0 avremo:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \ln\left[\frac{\Lambda(\eta)}{1 - \Lambda(\eta)}\right] = \ln[\Lambda(\eta)] - \ln[1 - \Lambda(\eta)] = \eta \quad \eta = x_i'\beta$$

In sintesi, abbiamo visto come in un'analisi di regressione quando la variabile outcome è dicotomica:

- La media condizionata dell'equazione di regressione deve essere matematicamente coerente con la variabile outcome, e quindi deve essere definita tra zero e uno. Si può facilmente verificare che l'equazione proposta in precedenza per la regressione logistica soddisfa questo requisito.
- Gli errori si distribuiscono con distribuzione binomiale, a differenza della regressione lineare. Ed è proprio su questo assunto che si baseranno le nostre analisi.

I principi cardine che guidano l'analisi di regressione sono gli stessi sia per quella lineare che per quella logistica. (Vitale 2005).

ODDS RATIO

Siamo soliti quantificare le chance di un evento attraverso la definizione classica di probabilità, e cioè p=(casi fav./casi possibili). Ovviamente questo rapporto potrà oscillare nell'intervallo [0,1]. Un'alternativa a questa definizione è l'*odds*.
L'odds esprime il rapporto tra la probabilità di un evento e la probabilità che l'evento non accada. Data la probabilita p di un evento, l'odds O sarà definito come:

$$O = \frac{p}{1-p}$$

Prendendo in considerazione la tabella sottostante procediamo ad alcuni esempi sul concetto di odds.

In questo caso la probabilità di essere uomini, espressa come frequenza, è 30/42=0.714.

Volendo esprimere questa informazione mettendo in relazione le due categorie possiamo ricorrere all'odds, che in questo caso sarà (30/12)=2.5 ; Ciò significa che per ogni donna ci sono 2.5 uomini.

Invece l'*odds ratio* è la misura dell'associazione tra due fattori e si ottiene calcolando il rapporto di due odds. Ad esempio, nel caso sotto esposto, l'*odds ratio* sarà così calcolato:

Distribuzione congiunta di una rilevazione fittizia

Lavoro/Sesso	Uomini	Donne	Totale
Ingegneri	18	2	20
Insegnanti	12	10	22
Totale	30	12	42

$$OR = \frac{\frac{P(ing|uomini)}{1-P(ing|uomini)}}{\frac{P(ins|uomini)}{1-P(ins|uomini)}} = \frac{18}{2} * \frac{10}{12} = 7.5$$

E' possibile notare che un O<1 corrisponde ad un P<0.5 , viceversa se O>1 allora P>0.5 .

Similmente alla probabilità, che è definita nell'intervallo [0,1], l'Odds è definito nel'intervallo [0,+∞[. L'Odds condivide quindi con P il limite inferiore, ma a differenza di quest'ultimo può assumere valori potenzialmente infiniti

2.1 Metodi di stima dei parametri

Supponiamo di avere un campione di n osservazioni indipendenti della coppia (x_i,y_i) $i=1,2,\ldots,n$ dove y_i indica il valore di una variabile dicotomica e x_i è il valore della variabile indipendente x per l' i-esimo soggetto.

Inoltre, assumiamo che la variabile outcome possa avere due possibili output, codificati con 0 e 1, che rappresentano, rispettivamente, l'assenza o la presenza di una determinata caratteristica.

Per stimare il modello di regressione su un set di dati abbiamo bisogno dei valori β_0 e β_1, definiti parametri incogniti. Nella regressione lineare il metodo usato per la stima dei parametri è quello dei minimi quadrati (OLS). Tramite questo metodo vengono scelti i valori β_0 e β_1 tali che la somma dei quadrati degli scarti dei valori osservati di Y dai valori stimati dal modello, sia minimizzata.

Sotto le assunzioni tipiche del modello di regressione, ad esempio la normalità in distribuzione degli ε, il metodo OLS per la stima dei parametri risulta ottimale. Purtroppo, il metodo OLS può risultare non corretto quando la variabile dipendente è binaria per i motivi esposti in precedenza.

Il metodo utilizzato per il fitting di modelli GLM è quello della massima verosimiglianza (maximum likelihood, o ML).

Questo metodo rappresenta le fondamenta del nostro approccio alla stima parametrica di una regressione logistica. In generale, il metodo ML stima i valori dei parametri incogniti che massimizzano la probabilità di ottenere il set osservato di dati.

Il primo passo per applicare questo modello è quello di definire la funzione di massima verosimiglianza. Questa relazione esprime la probabilità di ottenere il campione osservato come funzione dei parametri incogniti. Gli stimatori ML per i parametri saranno quindi quelli che massimizzano la suddetta funzione.

Passiamo ora all'applicazione di questo metodo alla regressione logistica.

Se la variabile Y è dicotomica allora l'espressione per $\pi(x)$ fornirà, per il vettore dei paramentri $\boldsymbol{\beta}=(\beta_0,\beta_1)$, la probabilità condizionata di Y=1 dato x. La suddetta relazione si può esprimere come P(Y=1|x). Segue che la quantità 1- $\pi(x)$ è la probabilità che Y assuma valore 0 dato x, e cioè P(Y=0|x).

Quindi per le coppie (x_i,y_i), dove $y_i=1$ il contributo alla funzione ML sarà $\pi(x_i)$, cioè il valore della funzione $\pi(x)$ calcolato in x_i.

Un metodo utile per esprimere il contributo della coppia i-esima (x_i,y_i) è questo:

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

La funzione likelihood per un modello binario sarà la seguente:

$$l(\beta)=\prod\prod \pi(x_i)^{y_i}\,[1-\pi(x_i)]^{1-y_i}$$

Spesso alla funzione L viene applicata una trasformazione logaritmica al fine di semplificare la notazione e le procedure di calcolo. La funzione ottenuta viene chiamata *log-likelihood.*

$$L(\beta)=\ln\,[l(\beta)]=\sum\{y_i\,\ln\,[\pi(x_i)]+(1-y_i)\ln\,[1-\pi(x_i)\}$$

Il principio del metodo ML è che verranno usati come valori di **β** quelli che massimizzano la funzione di log-verosimiglianza. A tal fine deriviamo la funzione rispetto a β_0 e β_1 e poniamo entrambe le derivate uguali a zero.

$$\sum [y_i - \pi(x_i)] = 0 \qquad \sum x_i [y_i - \pi(x_i)] = 0$$

Le espressioni non sono lineari nei paramentri e perciò richiedono metodi particolari per la loro risoluzione. Tali metodi sono spesso iterativi e implementati in software per il calcolo statistico. Per una più approfondita analisi di questi metodi si rimanda al lavoro di McCullagh e Nelder (1989)[2], in particolare essi hanno dimostrato che un metodo per trovare le soluzioni di queste equazioni è quello dei minimi quadrati pesati iterativo.

I valori di **β** ottenuti dalla soluzione delle precedenti equazioni sono chiamati *stime di massima verosimiglianza* e saranno indicate con B^ . Ad esempio, $\hat{\pi}$ (x_i) è la stima ML di $\pi(x_i)$. Questa quantità indica una stima della probabilità condizionata di Y=1, dato $x=x_i$. Un' interessante conseguenza della relazione precedente è la seguente:

$$\sum y_i = \sum \hat{\pi}(x_i)$$

[2] (McCullagh, P. and Nelder, J.A.,1989)

Questa espressione è ci dice che la somma dei valori osservati di y è uguale alla somma dei valori stimati. Questa proprietà sarà utile quando discuteremo i metodi per valutare un modello nel suo complesso.

2.2 Verifica del modello

Dopo aver stimato i coefficienti, un primo sguardo al modello serve a assicurarsi che le variabili ottenute siano significative. Questo, in genere, implica l'applicazione di test sulle ipotesi per determinare quali variabili sono "significative" rispetto alla variabile outcome.

Cominceremo a discutere l'approccio generale per un singolo predittore. Dopo aver stimato i coefficienti abbiamo bisogno di valutare la significatività dei dati ottenuti e la loro correttezza. Questo processo si esplica attraverso il test delle ipotesi per determinare se le variabili indipendenti sono *significative* rispetto alla variabile dipendente. Nella regressione logistica i coefficienti esprimono la variazione del logit per ogni variazione unitaria del predittore. Innanzitutto analizziamo in modo generale un test su una proporzione binomiale. Supponiamo di avere una popolazione finita P composta da n unità con un carattere d'interesse dicotomico. Consideriamo l'ipotesi nulla $H_0 : \pi=\pi_0$.e cioè che il parametro sia uguale a un certo valore fissato .

Se estraiamo con reimmissione un campione da P di ampiezza n otterremo $(x_1,x_2,\ldots,x_n)$. Al variare del campione ciascuna x_i descriverà una variabile casuale con la stessa struttura della popolazione. Ogni variabile casuale così estratta sarà una Bernoulli indipendente e la stima ottimale p sarà semplicemente la frequenza relativa, che avrà una distribuzione identica a quella di una binomiale frequenza e cioè:

$$E(p) = \pi \qquad \sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

In questo caso p è lo stimatore, mentre π è il parametro incognito. La statistica test sarà allora:

$$z = \frac{p - \pi_o}{\sqrt{\frac{\pi_0(1-\pi_o)}{n}}}$$

Se H_0 è vera, z descrive una variabile casuale con media zero e varianza 1. Il teorema del limite centrale ci assicura che per n moderatamente grande la variabile casuale Z si distribuisce come una normale standardizzata.

Un test di significatività indica se un particolare valore di un paramento è plausibile. Possiamo ottenere più informazioni attraverso la costruzione di un intervallo di confidenza per quel parametro. Se definiamo SE l'errore standard di p allora un intervallo di confidenza al 100(1-α)% per π è dato dalla seguente formula.

$$p \pm z_{\alpha/2}(SE) \quad con \quad SE = \sqrt{\frac{p(1-p)}{n}}$$

Dove z indica il percentile di una normale standardizzata che ha probabilità α/2 ;

Un'approccio iniziale potrebbe essere quello di chiedersi se il modello che include il singolo predittore ci dà più informazioni rispetto al modello che non include tale predittore. La risposta a questa domanda si ottiene confrontando i valori osservati con quelli ottenuti dal modello con e da quello senza il predittore. Bisogna precisare che si sta procedendo ad un' analisi relativa alla singola variabile e non alla bontà dell'intero modello, problematica che verrà trattata in seguito.

Il metodo generale per valutare la significatività di una variabile è facilmente illustrato nel caso di regressione lineare. Un confronto tra i due modelli ci permetterà di cogliere meglio le similitudini e le differenze tra questi approcci.

Nella regressione lineare il confronto tra valori predicted e osservazioni si basa sul quadrato della distanza tra valore osservato e valore predicted (cioè il punto sulla retta di regressione). Se y_i indica i valori osservati e $\hat{y}_i$ la stima di y per l'i-esimo individuo, allora la statistica utilizzata per effettuare il confronto è la seguente (dove SSE sta per *sum of squared errors of prediction*):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Nel modello che non contiene la variabile indipendente in questione l'unico parametro è β_0, quindi lo stimatore sarà uguale a $\bar{y}$, cioè la media di y. Quando includiamo nel modello la variabile indipendente ogni decremento di SSE sarà dato dal fatto che il coefficiente angolare (β) per la variabile indipendente è diverso da zero.

La variazione di SSE sarà quindi guidata dalla fonte di variabilità del modello e cioè da SSR (*sum of squared residuals*).

$$SSR = \left[\sum_{i=1}^{n} (y_i - \bar{y}_i)^2\right] - \left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]$$

Nella regressione lineare il focus è sulla quantità SSR. Un valore abbastanza grande ci indica che la variabile indipendente è importante.

I principi guida per la regressione logistica sono gli stessi. Ci interessa confrontare i valori osservati e quelli stimati dai modelli con e senza la variabile indipendente. In questo caso il confronto verrà operato mediante la funzione di log-verosimiglianza esposta in precedenza. Innanzitutto è bene chiarire che il valore osservato può essere visto come un valore stimato dal modello *saturo*. Un modello saturo è un modello che contiene tanti parametri quante sono le coppie di valori assunti dalle variabili. (Ad esempio una regressione lineare con due sole coppie di osservazioni).

2.2.1 Devianza e statistica G

Nel modello logistico viene operato un confronto tra i valori osservati e quelli stimati attraverso la seguente statistica:

$$D = -2 \ln \left[\frac{(likelihood\ fitted\ model)}{(likelihood\ saturated\ model)} \right]$$

La quantità all'interno delle parentesi è chiamata *likelihood ratio.* L'uso del termine (-2) conduce ad una più semplice trattazione matematica, ma soprattutto ci permette di ottenere una quantità la cui distribuzione è nota e ciò ci permetterà di testare le ipotesi in secondo momento. Il test a cui si è fatto riferimento è conosciuto come *likelihood ratio test.* Nel caso di un variabile dipendente binaria dove i valori assunti sono uno o zero, il valore della funzione likelihood per il modello saturo è:

$$l(saturated\ model) = \prod {y_i}^{y_i} (1 - y_i)^{(1-y_i)} = 1$$

Sostituendo nell'equazione di D la funzione di log-verosimiglianza del modello logistico otterremo:

$$D = -2 \ln(likelihood\ fitted\ model)$$

$$D = -2 \sum \left[y_i \ln \left(\frac{\hat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}_i}{1 - y_i} \right) \right]$$

La statistica D è chiamata *devianza* e gioca un ruolo fondamentale nel determinare l'adeguatezza del modello (*goodness of fit*). A questo punto è bene precisare che bisogna guardare alla devianza negli stessi termini con cui si guarda a SSR nella regressione lineare.

Al fine di valutare la significatività delle variabili indipendenti confrontiamo il valore di D con e senza la variabile indipendente. La variazione in D dopo l'inclusione della variabile indipendente nel modello sarà così ottenuta:

$$G = D(modello\ senza\ variabile) - D(modello\ con\ variabile)$$

Quest'ultima relazione può essere espressa come:

$$G = -2\ln\left[\frac{(likelihood\ senza\ variabile)}{(likelihood\ con\ variabile)}\right]$$

Nello specifico caso di una singola variabile indipendente è facile dimostrare che quando la variabile non è presente nel modello, la stima ML di β_0 e il relativo valore di G sono:

$$\hat{\beta}_o = \frac{\sum y_i}{\sum(1-y_i)} = \frac{n_1}{n_0} \qquad G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod \hat{\pi}_i^{\,y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right]$$

Sotto l'ipotesi che $\beta_1=0$, la statistica G si distribuisce come una chi-quadrato con un grado di libertà. Operativamente, una volta calcolato in valore di G diremo, con un livello di significatività α, che la variabile indipendente è significativa se $P[\chi^2_{(1)}>G]<\alpha$ .Il calcolo del suddetto test è integrato nella maggior parte dei software per l'elaborazione statistica dei dati, compreso quello di cui vi avvarremo nel corso dell'analisi, cioè R.

2.2.2 Il test di Wald (W)

Sia β un parametro. Consideriamo un test con ipotesi $H_0 : \beta=\beta_0$. La statistica test più semplice sfrutta l'ipotesi di distribuzione normale per grandi campioni dello stimatore ML per β. Sia SE la deviazione standard della stima, calcolata sostituendo la stima ML nell'espressione classica della deviazione standard.

$$SE(\hat{\beta}) = \sqrt{p(1-p)/n} \qquad\qquad z = \frac{(\hat{\beta} - \beta_0)}{SE}$$

Quando l'ipotesi H_0 è verificata la statistica test z avrà distribuzione approssimativamente normale. Allo stesso modo la statistica z^2 avrà distribuzione χ^2 con un grado di libertà. Questo tipo di test è chiamato *test di Wald*.

La statistica di Wald è usata per testare la significatività di ogni coefficiente preso singolarmente ed è calcolata dividendo ogni coefficiente per il suo standard error. L'idea di fondo è testare l'ipotesi che il coefficiente di una variabile indipendente nel modello è significativamente diverso da zero. Se il test non riesce a rigettare l'ipotesi nulla ciò significa che rimuovere la variabile dal modello non comprometterà la correttezza di quest'ultimo.

La statistica Wald è semplice da calcolare ma in letteratura è stata molto discussa la sua validità per campioni numericamente piccoli. Menard(1995)[3] ha evidenziato che per grandi coefficienti, lo standard error è eccessivamente alto e questo porta ad un abbassamento non giustificato del valore della statistica di Wald. Anche Hauck e Donner[4] (1977) hanno esaminato le performance del test di Wald dimostrando che esso spesso non riesce a rigettare l'ipotesi nulla, essi infatti ritengono più affidabile per la regressione logit l'uso del likelihood ratio test.

Un importante ausilio alle tecniche precedenti è il calcolo e l'interpretazione di appropiati intervalli di confidenza per i parametri di interesse. Le basi per la costruzione di un intervallo di confidenza sono le nozioni teoriche introdotte per formulare i test di significatività precedenti. In particolare gli intervalli di confidenza per l'intercetta e per β si basano sui rispettivi test di Wald.

Gli estremi di un intervallo di confidenza 100(1-α)% per β saranno quindi:

$$\widehat{\beta_1} \pm z_{1-\alpha/2} SE\left(\widehat{\beta_1}\right)$$

Il logit, come spiegato in precedenza, è la parte lineare di un modello logistico. La sua stima sarà quindi data da:

$$\hat{g}(x) = \widehat{\beta_0} + \widehat{\beta_1} x$$

[3] (Menard,S.,2000)

[4] (W.W.Hauck Jr. and A. Donner,1977)

Lo stimatore della varianza dello stimatore del logit è la varianza di una somma. In generale tale valore è dato dalla somma delle varianze di ogni termine e dalle rispettive covarianze di ogni coppia di termini. Potremmo quindi scrivere gli estremi di un intervallo di confidenza Wald per il logit come:

$$\hat{g}(x) \pm z_{1-\alpha/2}\widehat{SE}[\hat{g}(x)]$$

dove SE è la radice positiva dello stimatore per la varianza.

2.3 Cenni alla regressione logistica multipla

Nel capitolo precedente è stata fatta un'introduzione alla regressione logistica in forma univariata. Nel seguito verrà fornita una trattazione generale, ma sintetica, del modello logit multivariato. Consideriamo una serie di p variabili indipendenti $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. Possiamo scrivere la probabilità di successo condizionata come $P(Y=1|\mathbf{x})=\pi(\mathbf{x})$. Il logit e l'equazione del modello saranno così definite:

$$g(x) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Se alcune variabili esplicative sono discrete, mentre altre non lo sono, è utile e necessario codificare le variabili come sesso,livello di istruzione … mediante l'utilizzo di variabili dummy. Nel caso pratico che esamineremo successivamente sono presenti delle rilevazioni quali ad esempio il livello di istruzione degli intervistati. Nel caso specifico i livelli sono "P","S","PS". In questo caso sono necessarie due variabili dummy. La strategia usata è questa: quando il livello di istruzione rilevato è "P", le due variabili dummy D_1 e D_2 assumono entrambe valore zero. Quando invece si osserva "S", D_1 sarà pari a uno e D_2 sarà pari a zero. Infine quando si osserva "PS" si avrà $D_1=0$ e $D_2=1$.

	D_1	D_2
P	0	0
S	1	0
PS	0	1

Più in generale, se una variabile può assumere k valori, allora serviranno k-1 variabili dummy per inserire quella variabile correttamente nel modello. Si supponga che la j-esima variabile x_j abbia k_j livelli. Le k_j-1 variabili dummy saranno indicate con D_{jl} e i coefficienti di queste variabili saranno β_{jl},l=1,2,…,k_j-1. Premesso ciò potremmo scrivere il logit per un modello con p variabili x e le dummy necessarie per le variabili x discrete come:

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p x_p$$

2.3.1 Fitting del modello logistico multiplo

Consideriamo un campione di n osservazioni indipendenti ($\mathbf{x}_i$,y_i), i=1,2,…,n . Come nel caso univariato il fitting del modello mira ad ottenere la stima del vettore $\boldsymbol{\beta}'$=(β_0,β_1,…,β_p). Il metodo utilizzato sarà sempre quello della massima verosimiglianza. La funzione ML differisce dalla precedente solo per il termine $\pi(\mathbf{x})$. Differenziando la funzione log likelihood rispetto ai p+1 coefficienti si otterranno p+1 equazioni likelihood. Tali equazioni saranno:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \qquad \sum_{i=1}^{n} x_{ij} [y_{ij} - \pi(x_i)] = 0$$

Come nel caso univariato la soluzione di queste equazioni necessita di complessi algoritmi di calcolo fortunatamente già integrati all'interno dei sfotware statistici più comuni. Definiremo con $\hat{\pi}(\mathbf{x}_i)$ i valori calcolati usando le stime di β e i valori x_i.

Nel caso multivariato un importante contributo riguardo i metodi di stima di varianza e covarianza dei coefficienti è quello di Rao[5] (1973). Rao dimostra che gli stimatori per la varianza e la covarianza sono ottenuti dalla matrice delle derivate seconde parziali della funzione di log likelihood. La forma generale delle derivate parziali sarà la seguente per l'elemento j-esimo, con l=0,1,2,…,p .

$$\frac{\partial^2 L(\beta)}{\partial {\beta_j}^2} = -\sum {x_{ij}}^2 \pi_i (1-\pi_i)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum x_{ij} x_{il} \pi_i (1-\pi_i)$$

Definiamo la matrice **I(β)** di ampiezza (p+1) x (p+1) contenente gli opposti delle derivate precedenti. Questa matrice è detta *informazione di Fisher osservata.* Var. e Covar. sono ottenuti dall'inverso della matrice di Fisher, in particolare: Var(**β**)= $\mathbf{I}^{-1}$(**β**). La notazione Var(β_j) indicherà l'elemento j-esimo sulla diagonale della matrice, cioè la varianza di βj.

Gli stimatori della varianza saranno quindi ottenuti calcolando il valore di Var in β^. Una semplice conseguenza di quanto detto sarà che la deviazione standard dei coefficienti stimati sarà:

$$\widehat{SE}(\hat{\beta}_j) = \left[\widehat{Var}(\hat{\beta}_j)\right]^{1/2}$$

[5] C.R.Rao (1973)

E' possibile riformulare la matrice di informazione in una forma più adatta e utile per valutare il fitting del modello. Tale forma è

$$\hat{I}(\hat{\beta}) = X'VX$$

Dove X è una matrice n x (p+1) contenente i valori delle osservazioni per ogni soggetto, V è una matrice diagonale n x n contenente gli elementi generici $\pi_i(1-\pi_i)$

$$X = \begin{bmatrix} 1 & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{np} \end{bmatrix} \qquad V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & 0 \\ 0 & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

Una volta fittato il modello cominciamo il processo di controllo e testing. Il test likelihood ratio esposto in precedenza ha la stessa formulazione anche per il caso multivariato. Il test si basa sulla statistica G. L'unica differenza sono i valori di π^ essendo questi ultimi calcolati sulla base del vettore β^. Sotto l'ipotesi nulla di coefficienti pari a 0, la distribuzione di G sarà ugualmente una χ^2 con p gradi di libertà.

Il test di Wald invece è dato dalla seguente statistica:

$$W = \hat{\beta}'\left[\widehat{Var}(\hat{\beta})\right]^{-1}\hat{\beta} = \hat{\beta}'(X'VX)\hat{\beta}$$

Che avrà distribuzione χ^2 con p+1 gradi di libertà sotto l'ipotesi che ognuno dei p+1 coefficienti sia uguale a 0.

2.4 La bontà di adattamento e le metodologie di costruzione

Dopo aver introdotto i concetti basilari del modello di regressione logistica, ora esamineremo le metodologie di costruzione di un modello e quelle di valutazione della bontà di adattamento del modello stesso. In primo luogo ci interessa applicare metodi efficaci per la selezione delle variabili esplicative da inserire nel modello. Il processo di selezione diventa più complesso all'aumentare delle variabili esplicative poiché aumentano anche gli effetti e le possibili interazioni tra queste ultime. Le metodologie che verranno brevemente introdotte nel seguito sono però solo complementari ai fini perseguiti, infatti, oltre agli approcci meramente statistici è opportuno dotarsi di buon senso quando si prova a costruire un modello.

I criteri per includere o meno una variabile in un modello possono variare molto rispetto al problema da analizzare. L'approccio classico prevede la ricerca di un modello corretto, valido e parsimonioso. Un modello parsimonioso sul numero di variabili utilizzate sarà numericamente stabile e di più semplice generalizzazione. Più variabili sono incluse nel modello, viceversa, più grandi diventano gli SE delle stime e più il modello sarà dipendente dai dati osservati. In ambito epidemiologico è prassi inserire nel modello qualsiasi variabile si ritenga intuitivamente significativa sebbene quest'ultima sia statisticamente poco significativa. Il principale problema di questo approccio è che il modello possa presentare *overfitting*, producendo stime numericamente non valide. L'overfitting è infatti spesso caratterizzato da stime dei coefficienti o degli SE irrealisticamente grandi.

Come nella regressione lineare, possiamo avvalerci di algoritmi che selezionano le il possibile set di variabili esplicative automaticamente. In particolare nell'applicazione pratica alla fine di questo lavoro verrà utilizzato il metodo *stepwise backward.* Ogni procedura stepwise per la selezione di variabili si basa su algoritmi statistici che controllano l'"importanza" delle variabili e procedono alla loro inclusione o eliminazione sulla base di una legge ben definita. Per "importanza" di una variabile statistica si intende un termine di misura per la significatività statistica del coefficiente della variabile.

La statistica utilizzata al fine di valutare l'importanza di una variabile varia in base alla assunzioni del modello preso in considerazione. Nella regressione stepwise lineare viene

utilizzato il test F poiché si ipotizza una distribuzione normale degli errori. Nella regressione logistica abbiamo visto che gli errori hanno una distribuzione binomiale e la significatività viene valutata attraverso il likelihood ratio test. Premesso ciò, la variabile più significativa ad ogni step sarà quella che produce la maggior variazione rispetto al log likelihood relativo ad un modello che non contiene la variabile. (In pratica quella per cui il valore della statistica G descritta in precedenza è il più alto). Un aspetto cruciale per l'uso della stepwwise regression è la scelta del livello di significatività del test. Lee and Koval (1997), in particolare, hanno studiato la scelta ottimale del livello di significatività nel caso di stepwise logistic regression. I risultati delle loro ricerche raccomandano di utilizzare un livello di significatività compreso tra 0.15 e 0.2 . Un variabile sarà quindi considerata importante per il modello se il p value associato alla statistica test G sarà minore del livello di significatività scelto.

Un altro criterio per quantificare la bontà di un modello è il cosiddetto AIC, che sta per *Akaike Information Criterion*. L'AIC giudica un modello in base a quanto i valori stimati tendono ad essere vicini ai valori osservati, sintetizzando questa distanza secondo un criterio ben preciso.

$$AIC = -2(\log likelihood - numero\ di\ parametri)$$

Spesso è utile sintetizzare un modello logistico binario attraverso una tabella di classificazione. Una tabella di questo genere classifica la variabile binaria y e i suoi valori stimati. Per le stime ipotizzeremo che quando la probabilità stimata è maggiore di una generica probabilità π_0 allora y stimato sarà uguale ad uno e viceversa. Due interessati definizioni utili a sintetizzare il potere predittivo del modello sono la sensibilità e la specificità.

$$sensibilità = P(\hat{y} = 1|y = 1) \qquad specificità = P(\hat{y} = 0|y = 0)$$

La tavola di classificazione così definita ha però dei limiti , infatti essa fissando un valore di π_0 trasforma la probabilità stimata da un intervallo potenzialmente continuo ad uno binario. Una metodologia per ovviare a questo problema è la costruzione della curva ROC. Una curva ROC è costruita disegnando la sensitività sopra definita come funzione di (1-specificità) per i possibili valori di π_0. La curva ROC è più informativa della tabella di classificazione poiché sintetizza il potere predittivo del modello per ogni possibile π_0. Quando π_0 tende a zero, otterremo quasi tutte y^=1 e quindi la sensibilità sarà vicina ad uno, la specificità sarà vicina allo zero e il punto (1-specificità, sensibilità) avrà coordinate (1,1).

La curva ROC, infatti, assume spesso una forma concava che si estende tra i punti (0,0) e (1,1). Dato un certo livello di specificità otterremo un potere predittivo maggiore per valori tanto più alti della sensibilità. Quindi più la curva ROC sarà alta più il modello avrà potere predittivo. L'area di piano sottostante la ROC è essa stessa una misura del potere predittivo del modello, ed è definita come concordanza. L'indice di concordanza c stima la probabilità che i valori osservati e quelli stimati siano concordanti, ciò significa che le osservazioni aventi y maggiore avranno anche un π^ maggiore. Ad esempio, un valore di c pari a 0.5 indica che le stime non hanno nulla di meglio rispetto ad una scelta casuale. Questo caso è quello di un modello che ha sola intercetta. Per questo modello infatti la curva ROC sarà una retta passante per i punti (0,0) e (1,1).

3. Applicazione del modello logistico: Il caso cileno

Il plebiscito cileno del 1988 fu un plebiscito nazionale, previsto nella costituzione cilena del 1980, indetto in Cile il 5 Ottobre 1988 per determinare se il popolo volesse conferire ad Augusto Pinochet un ulteriore mandato di 8 anni come Presidente della Repubblica.

Nella precedente parte del lavoro sono state introdotte le metodologie statistiche utilizzate nel caso in cui si voglia costruire un modello volto a stimare relazioni esistenti tra tipologie eterogenee di variabili. In particolare nel seguito verrà costruito un modello logistico per studiare le variabili che più hanno influito sulle scelte del popolo cileno. Innanzitutto analizzeremo descrittivamente il dataset oggetto di studio per visualizzare le relazioni esistenti tra le variabili e inquadrare la situazione socio-economica degli intervistati. Il dataset oggetto di studio è composto da 2700 intervistati, su cui sono state rilevate 8 variabili, descritte come segue:

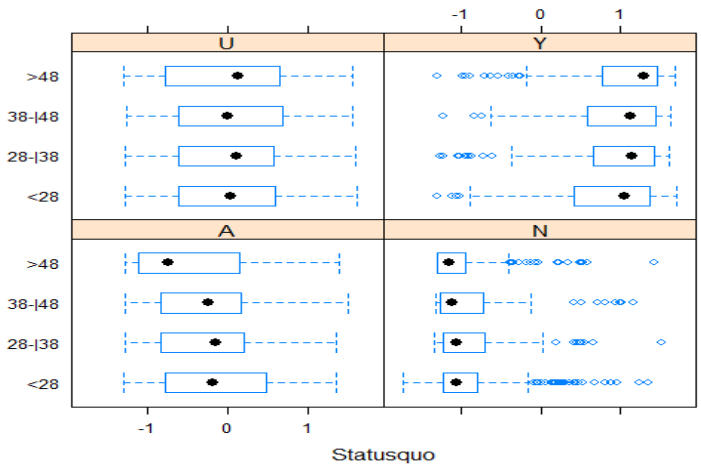
- **"Region"** indica la zona di residenza degli intervistati (**C**=Zona Centrale, **M**=Area metropolitana di Santiago, **N**=Nord, **S**=Sud, **SA**=Città di Santiago)

- **"Population"** è composta da numeri interi ed indica il numero di residenti per ogni regione.
- **"Sex"** e **"Age"** indicano, rispettivamente, il sesso e l'età degli intervistati.
- **"Education"** ci fornisce informazioni sul grado d'istruzione degli intervistati (**P**=Primaria, **S**=Secondaria, **PS**=Post-secondaria).
- **"Income"** indica il reddito mensile degli intervistati (espresso in Pesos).
- **"Status quo"** rispecchia la preferenza o meno per lo status-quo su una scala di valori.
- **"Vote"** è l'intenzione di voto (**A**=Astenuto, **N**=No, **Y**=Si, **U**=Indeciso).

Nel seguito si analizzerà graficamente l'impatto dello statusquo sull'orientamento dei votanti, attraverso l'uso di boxplot condizionati. L'esigenza di questo tipo di analisi nasce dalla necessità di comprendere le relazioni esistenti tra variabili per costruire un modello quanto più conforme e adatto allo scopo perseguito. Il software R fornisce ottimi pacchetti grafici come "lattice" e "ggplot2". Nell'appendice sarà inserito il codice utilizzato per grafici e modelli.

Nel grafico sottostante sono stati costruiti i boxplot condizionati alle varie fasce d'età e alll'intenzione di voto, rispetto alla variabile statusquo, che rappresenta la loro soddisfazione o insoddisfazione rispetto alla situazione complessiva del paese.

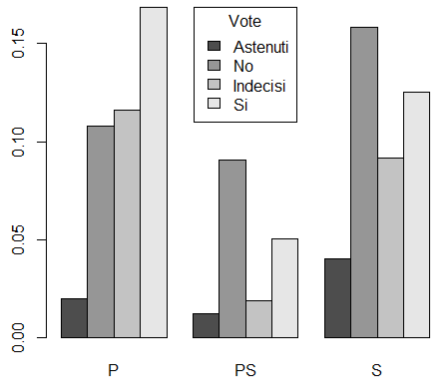
I riscontri ottenuti sono abbastanza prevedibili dal buon senso. In particolare si nota come tra i sostenitori di Pinochet ci fosse un ampia soddisfazione rispetto a qualsiasi fascia d'età Infatti il punto nero al centro del boxplot è la mediana, in questo caso stabilmente vicina al valore 1 per tutte le fasce d'età considerate. Sul fronte politico opposto, cioè quello dei potenziali votanti a sfavore di Pinochet la situazione si ribalta completamente, con una mediana stavolta vicina a -1. Le distribuzioni rispettivamente di Y e di N si definiscono asimmetrica negativa e positiva. Sebbene la densità maggiore di persone si trovi vicina alla mediana, si osseperrvano anche valori estremi, rappresentati dai punti blu al di fuori del "baffo" del boxplot. singolo punto al di fuori del box Y, ad esempio, potremmo pensare che quella persona sebbene avesse una percezione negativa dello statusquo era intenzionata a votare a favore di un nuovo mandato per il dittatore Cileno.



Passiamo ora a scomporre il voto rispetto al livello d'istruzione conseguito. Attraverso R costruiamo una tabella che ci aiuta a tale scopo.

```
   P  PS   S
A  52  32 103
N 266 224 397
U 296  52 237
Y 422 130 311
```

Provando a visualizare, in termini percentuali, le intenzioni di voto rispetto al grado d'istruzione conseguito otterremo l'istogramma sottostante. Ogni barra rappresenta un' informazione contenuta nella tabella precedente. Ad esempio gli intervistati con grado d'istruzione primario e favorevoli ad una rielezione di Pinochet sono 422, che espressi in percentuali rispetto all'intero numero di intervistati è pari ad oltre il 15%.



Possiamo valutare la dipendenza tra due o più variabili attraverso il test Chi quadrato di Pearson. Il software R, nel caso in questione, ha fornito un output di questo tipo:

```
        Pearson's Chi-squared test

data:  chile$vote and chile$education
X-squared = 135.85, df = 6, p-value < 2.2e-16
```

Il valore p è pari a $2{,}2*10^{-16}$, ciò significa ce la probabilità che l'associazione tra le due variabili sia dovuta al caso è pari a 0. Un valore p così basso, infatti, ci porta a rifiutare l'ipotesi nulla, in più potremmo dire che la variabile voto non è indipendente dalla variabile education. Passiamo ora a valutare il grado di correlazione presente tra le variabili oggetto di studio. La metodologia utilizzata sarà quella del test di Pearson per la correlazione. Si procederà prima a confrontare le variabili popolazione e statusquo e successivamente invece il livello di reddito e lo statusquo.

```
        Pearson's product-moment correlation

data:  chile$population and chile$statusquo
t = -10.487, df = 2681, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2345769 -0.1618710
sample estimates:
      cor
-0.198497

        Pearson's product-moment correlation

data:  chile$statusquo and chile$income
t = 1.9732, df = 2589, p-value = 0.04858
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.0002430528 0.0771436085
sample estimates:
       cor
0.03875071
```

Al fine di interpretare gli output di R bisogna sempre tenere a mente che il valore di p ci indica se rifiutare o meno l'ipotesi nulla rispetto al livello di significatività scelto. Nel primo caso p è molto piccolo,quindi saremo propensi a rifiutare l'ipotesi H_0, e la correlazione tra statusquo e popolazione residente è negativa di circa lo 0,19%.

Nel secondo caso invece il p value è poco più piccolo di 0.05, ciò significa che rigetteremo l'ipotesi nulla ma con un grado di significatività minore.

Nella seconda parte di questo capitolo verrà applicato il modello di regressione logistica attraverso R e ne verrà data un'interpretazione statistica sulla base dei concetti teorici esposti nel capitolo precedente. Innanzitutto bisogna precisare che il dataset oggetto di studio presentava valori mancanti e quindi sono state rimosse le relative righe. Una seconda operazione di raffinatura è stata quella di ri-codificare le modalità della variabile voto, da lettere ("Y","N") a numeri (1,0).

Cominciamo col costruire un modello logistico tra le variabili voto e statusquo, ottenendo il seguente output:

```
Call:
glm(formula = vote ~ statusquo, family = binomial(link = logit),
    data = s.chi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1761  -0.2852  -0.2001   0.1886   2.8084


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2237     0.1010   2.215   0.0267 *
statusquo     3.1819     0.1437  22.148   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2360.29  on 1702  degrees of freedom
Residual deviance:  737.27  on 1701  degrees of freedom
AIC: 741.27

Number of Fisher Scoring iterations: 6
```

L'output ci mostra innanzitutto la devianza dei residui, che verrà utilizzata per valutare la b ontà del modello. E' possibile notare che R ci mostra due tipi di devianza, la *null devianc e* che è la devianza di un modello contenente la sola intercetta, mentre quella definita *resid ual deviance* è la devianza classica. Nel caso sopra riportato si ha una riduzione significati va della devianza rispetto al modello nullo.

La parte dei *coefficients* ci mostra la stima del paramentro sotto la voce estimare, il relativo standard error, lo z value che come visto in precedenza rappresenta il valore della statistica di Wald e i rispettivi p value. E' facile notare che i p value sono minori del livello del 5%.

La statistica di Wald si distribuisce come una chi quadrato con un solo grado di libertà, qui ndi se prendiamo il valore z=22.148 e lo confrontiamo con la tavola della chi quadrato, acc oglieremo l'ipotesi nulla $\beta = \beta_0$. Come ci si aspettava, la variabile statusquo è molto import ante nello spiegare la variabilità delle intenzioni di voto del popolo cileno. Nel modello lin eare i parametri **β** rappresentano gli effetti marginali che le variabili x_i hanno sulla variabil e dipendente, infatti

$$\frac{\delta E\left(y_i | x_i \right)}{\delta x_i} = \beta$$

Quest'effetto marginale è quindi indipendente da i=1,2…,N .

Nel modello logit questa condizione non è vera poiché il vettore dei paramentri è inserito in una funzione non lineare, infatti esso dipende necessariamente dalle variabili esplicative x_i. La formula del modello indica che il *logit* aumenta di β per ogni variazione unitaria di x. Il paramentro β indica il tasso di crescita della curva. Quando β>0 la curva sarà crescente e viceversa. Quando il | β | cresce, allora la curva avrà un più alto tasso di crescita. Quando β=0 la curva degenera in una retta orizzontale. In quest'ultimo caso diremo che la variabile Y è indipendente da X.

$$\frac{\delta E\left(y_i | x_i \right)}{\delta x_i} = \frac{\delta \Lambda(\eta)}{\delta x_i} = \lambda(\eta)\,\beta$$

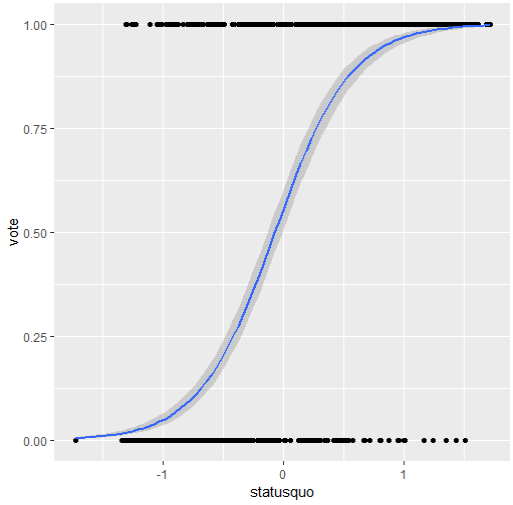
I coefficienti della regressione, prima indicati con β indicano la variazione nel logit dell'outcome per una variazione unitaria della variabile statusquo. Ad esempio per ogni variazione unitaria di statusquo, il logit di Y rispetto ad N aumenta di 3,1819 .

Un ‘ importante interpretazione del modello logistico si basa sul concetto di odds ratio più volte trattato in precedenza. L’odds di un “Si” sono date da:

$$\frac{\pi(x)}{1-\pi(x)} = exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}$$

Questa relazione esponenziale fornisce un’interpretazione per β: L’odds si moltiplica di e^{β} per ogni aumento unitario di x. Detto questo, l’odds al livello (x+1) sarà pari all’odds del livello x moltiplicato per e^{β}. Quando β=0, e^{β}=1, e l’odds non varia al variare di x.

I risultati del modello ci mostrano infatti che logit[π^(x)] = 0,22 + 3,1819*(statusquo). Ciò significa che quando la variabile statusquo assume valore 0, il logit è uguale a 0,22 ; La probabilità di votare “Si”, per statusquo=0, è pari a 0,5547 . Se invece lo statusquo è pari a 0.5 , le probabilità di un foto favorevole passano allo 0.8594. Mediante i parametri ottenuti è quindi possibile conoscere la probabilità di osservare intenzioni di voto favorevoli a Pinochet per ogni specifico livello della variabile statusquo. Nel seguito è stata inserita una rappresentazione grafica della caratteristica curva S-shaped del modello in questione.



L’indice R quadro di McFadden è un’ ulteriore misura della varianza spiegata dal modello rispetto a quella totale. Esso assume significato omologo all’indice R^2 utilizzato nella regressione lineare ed è dato da 1- [$\ln(L_c)/\ln(L_0)$] , dove L rappresenta la verosimiglianza del modello e quella del modello nullo. Per il modello sopra menzionato l’indice di McFadden è pari a 0.687637 .

Un ulteriore test applicabile è quello di Hosmer-Lemeshaw, che in R riporterà il seguente output:

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  s.chi$vote, fitted(glm.statusquo)
X-squared = 7.5647, df = 8, p-value = 0.4771
```

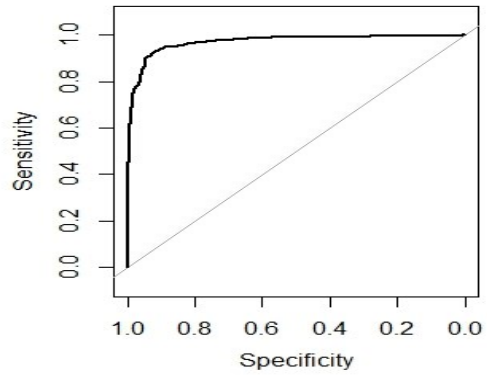
Dato il valore del p-value possiamo dire che il modello sembra fittare bene i dati poiché no n è presente una differenza significativa tra il modello e i valori osservati. Un p-value mino re di 0.05 indica scarso fitting e viceversa.

Procediamo ora alla costruzione della curva ROC costruendo innanzitutto la corrispondente tabella di classificazione.

```
          response
predicted   0   1
        0 810  75
        1  57 761
```

Questa tabella serve a confrontare i valori previsti dal modello e quelli effettivamente osservati. Nel seguito è riportata la curva ROC per il modello. Essa è costruita a partire dalla specificità e dalla sensitività del modello. Un modello con un’ottima capacità discriminatoria ha una curva ROC che si avvicina alla parte superiore sinistra del grafico, mentre un modello senza questa caratteristica avrà curva ROC prossima alla retta a 45° .

Volendo esprimere quantitativamente la tendenza della curva a spostarsi all’aumentare della capacità discriminatoria si misura l’area sottostante la curva. Nel caso specifico l’area è pari a 0.9699 .



Per selezionare le variabili più significative da inserire nel modello logistico multiplo utilizzeremo il metodo stepwise backward esposto in precedenza. Esso parte dal modello completo di tutte le variabili esplicative e procede step by step ad eliminare le variabili partendo da quella con l'associazione meno significativa con la variabile dipendente. Il modello scelto sarà quello che presenta un AIC minore rispetto agli altri. L'output del comando in R sarà il seguente:

```
Call:
glm(formula = vote ~ statusquo + education + sex, family = binomial,
    data = s.chi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2553  -0.2845  -0.1297   0.2009   2.9614

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0153     0.1890   5.373 7.75e-08 ***
statusquo     3.1689     0.1448  21.886  < 2e-16 ***
educationPS  -1.1074     0.2914  -3.800 0.000145 ***
educationS   -0.6828     0.2217  -3.079 0.002077 **
sexM         -0.5742     0.2022  -2.840 0.004518 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2360.29  on 1702  degrees of freedom
Residual deviance:  708.24  on 1698  degrees of freedom
AIC: 718.24

Number of Fisher Scoring iterations: 6
```

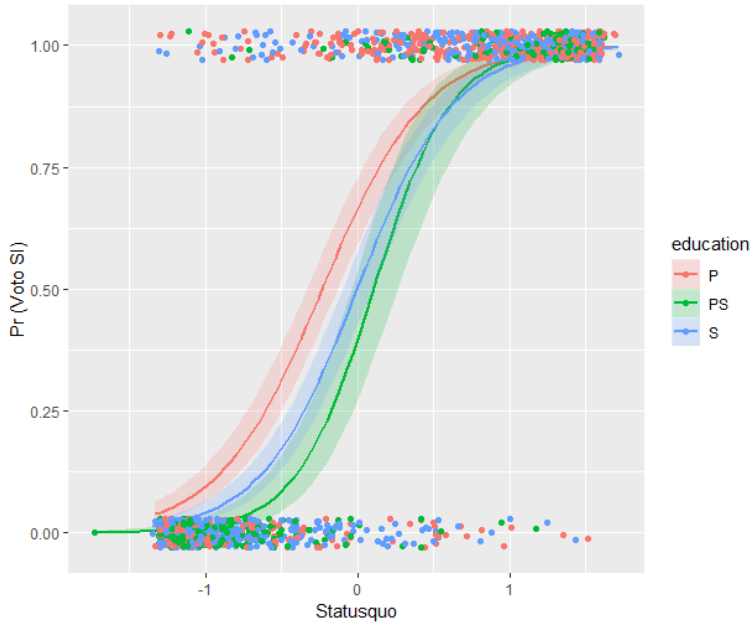
Sebbene questa volta si tratti di una regressione multipla l'output di R non è cambiato molto rispetto al caso univariato. Una differenza da sottolineare è però quella relativa alla variabile education. Quest'ultima infatti presentava 3 diversi livelli, P S PS . Il software ha creato quindi due variabili dummy per codificare e inserire nel modello anche quella variabile. In tal caso la categoria "P" è stata scelta come riferimento e quindi il significato dei coefficienti sarà il seguente:

$$OR(educationPS=”PS”,educationP=”P”) = exp(-1.1074) = 0.3304$$
$$OR(educationS=”S”,education=”P”) = exp(-0.6828) = 0.5052$$

In questo caso a parità di condizioni,ad esempio,per una donna con statusquo pari a 0 e un livello educativo Secondary, il logit stimato dal modello sarebbe stato pari a -0.09 che corrisponde ad una probabilità di votare a favore, per un individuo donna con statusquo nullo e livello d'istruzione secondario pari a 0,4775.

Il grafico sottostante mostra il modello multiplo, senza la variabile età, diviso per le singole componenti della variabile education. I punti presenti alle estremità sono i campioni osservati.

E' possibile notare come la pendenza delle curve sia differente rispetto ai vari livelli d'istruzione. In particolare la curva in verde presenta una più marcata pendenza, nei punti vicini allo zero di statusuqo, rispetto alle altre due. Ciò significa che in quei punti una variazione di statusquo di pari entità avrebbe un impatto maggiore sugli individui con livello d'istruzione elevato rispetto a quelli meno istruiti.



Proviamo a confrontare questo modello con la sua versione semplificata introdotta in precedenza. E' possibile utilizzare la statistica G, che in questo caso ha distribuzione chi quadrato con quattro gradi di libertà. A tal fine è possibile utilizzare la funzione *anova* già presente in R di default. L'output ottenuto sarà il seguente:

```
Analysis of Deviance Table

Model 1: vote ~ statusquo
Model 2: vote ~ statusquo + education + sex
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1701     737.27
2      1698     708.24  3   29.032 2.205e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analizzando l'output della funzione è possibile notare che la devianza residui è minore nel modello 1 ma soprattutto che Pr(>Chi) è minore del livello di significatività fissato.

Si potrebbe asserire quindi che aggiungendo al modello semplice altre variabili si è ottenuto un miglioramento statisticamente significativo in quest'ultimo. Se proviamo ad utilizzare nuovamente il test di Hosmer-Lemeshaw otteniamo il seguente output

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  s.chi$vote, fitted(glm.sta.edu.sex)
X-squared = 7.6996, df = 8, p-value = 0.4633
```

Rispetto allo stesso test sul modello precedente questo presenta una risibile differenza nel p value, sebbene il modello precedente abbia ottenuto un punteggio migliore. Costruiamo ancora una tabella di classificazione per il modello al fine di testare la capacità predittiva.

```
          response
predicted   0   1
        0 809  64
        1  58 772
```

4. Conclusioni

Dopo il golpe del settembre 1973 il Cile ha vissuto in vera e propria dittatura militare. E' doveroso precisare il contesto storico entro il quale si svolse il plebiscito. L'agenda di Pinochet era chiara fin da subito, repressione del dissenso e politiche fortemente anticomuniste e liberiste. L'economia cilena era ridotta male nei mesi successivi al golpe. Pinochet nominò un gruppo di economisti, i *Chicago Boys*, che indirizzarono la politica economica in senso neoliberista, in netto contrasto con l'agenda del governo precedente che parlava di nazionalizzazioni ed economia pianificata. Fu proposta una ricetta di austerità, la spesa pubblica venne drasticamente ridotta. Tali politiche economiche inizialmente causarono ingenti danni ai settori più poveri della società cilena, gli stipendi medi ad esempio scesero dell'8% e i risparmi delle famiglie erano il 28% di ciò che erano stati nel 1970. I budget per l'istruzione, la salute e gli alloggi erano scesi di oltre il 20% in media. Tutto ciò, ed è bene ricordarlo, in un clima di repressione militare feroce e disumano che vessava ancor di più i ceti meno abbienti della popolazione. Non sorprende che nell'analisi descrittiva dei dati siano emerse forti disparità sociali, dato il contesto politico e socio economico dell'epoca. Lo scopo principale di quella parte del lavoro è stato in primis fornire strumenti grafici che potessero sintetizzare al meglio i dati raccolti, e in secundis di analizzare preliminarmente i dati nell'ottica di costruzione di un modello logistico appropriato. Ed è proprio col modello semplice che si è avuta la conferma della forte influenza della variabile statusquo sull'intenzione di voto, peculiarità già vista in precedenza con il box plot condizionato. Un indice R^2 di MacFadden alto e i risultati positivi ottenuti nei vari test sul modello confermano ancora questo legame. Sebbene usando il buon senso possa apparire evidente che il livello di gradimento influisce inevitabilmente sulle intenzioni di voto, si è voluto procedere inizialmente a confermare questa opinione utilizzando metodi statistici efficaci e rigorosi. Nella seconda fase invece si è cercato di inserire altre variabili che potessero migliorare il modello e quindi le stime ottenute con esso. Il modello migliore ottenuto tramite il metodo stepwise è stato quello che comprendeva le variabili statusquo, livello d'istruzione e sesso.

E’ stata riscontrata una probabilità maggiore di votare “Si” in una fascia trasversale di popolazione ma con un valore di statusquo vicino ad 1. Viceversa per valori bassi di statusquo la probabilità di essere favorevoli ad una rielezione di Pinochet scendono drasticamente. Un’interessante informazione che si evince dal grafico e dai risultati delle stime è che le persone con un titolo di studio post secondario, in corrispondenza di valori neutri di statusquo, sono quelli più propensi a cambiare opinione rapidamente rispetto a variazioni dello statsquo. I risultati di un qualsiasi processo elettorale dipendono spesso da quanto una fazione politica riesca a convincere gli indecisi o gli astenuti a votare in proprio favore, soprattutto nel caso in cui il divario in termini percentuali tra le due fazioni sia ridotto. In quegli anni in Cile i mezzi di informazione e i cittadini i gruppi organizzati riuscirono a mettere in piedi una campagna referendaria imponente, che unita al forte malcontento sociale finì col determinare la fine di un’epoca di dittatura militare.

E’ necessario precisare che nei modelli analizzati in questo lavoro sono stati opportunamente eliminati dal dataset tutti gli intervistati che si dicevano indecisi o si sarebbero astenuti. Questa scelta è stata dettata dall’eccessiva complessità di un siffatto modello e dalla mancanza soggettiva dell’autore di alcune nozioni complesse che si tratteranno in corsi di studio successivi. Un modello multinomiale del genere ci avrebbe fornito una visione d’insieme anche rispetto alle categorie degli astenuti e degli indecisi, magari arricchendo quest’ultimo anche con dati non compresi nel dataset trattato. A valle, infatti, sono stati gli elettori marginali a decidere le sorti di quel voto essendo quest’ultimo conclusosi con una percentuale di voti “No” dell’ 56% .

5. Appendice

Nel seguito verrà inserito il codice script di R utilizzato nel corso dell'analisi:

```
chi=chile
chi$age[chile$age<=28]="<28"
chi$age[chile$age>28 & chile$age<=38]="28-|38"
chi$age[chile$age>38 & chile$age<=48]="38-|48"
chi$age[chile$age>48]=">48"
chiage=factor(chi$age)
chiage=ordered(chiage, levels=c("<28","28-|38","38-|48",">48"))
bwplot(chiage~statusquo|vote,data=chile,horizontal=1,xlab="Statusquo")
bwplot(chiage~income|vote,data=chile,horizontal=1,xlab="Income")
bwplot(income~statusquo|vote,data=chile,horizontal=1,xlab="Statusquo")
table(chile$vote,chile$education)
table.eduvot=prop.table(table(chile$vote,chile$education))
barplot(table.eduvot,beside=1)
legend("top",legend=c("Astenuti","No","Indecisi","Si"),fill=terrain.colors(4),
title="Vote")
##REGRESSIONE LOGISTICA####
s.chi=chile[chile$vote %in% c("Y","N"),]
str(s.chi)
s.chi$vote=factor(s.chi$vote,exclude=1)
s.chi$vote=factor(s.chi$vote,exclude=3)
table(s.chi$vote)
num=as.numeric(s.chi$vote)
str(num)
num=recode(num,"c(1)='0';else='1'")
table(num)      ###Y=1 N=0###
s.chi$vote=num
str(s.chi)
glm.statusquo=glm(formula=vote~statusquo,family=binomial(link=logit),data=s.chi)
summary(glm.statusquo)
exp(confint(glm.statusquo))
ggplot(s.chi, aes(x=s.chi$vote,y=s.chi$statusquo,size=depth))+geom_point(alpha=0.2)
pseudoR2<-function(mod) {1-(deviance(mod)/mod$null.deviance)}
pseudoR2(glm.statusquo)
classDF<-data.frame(response=s.chi$vote,predicted=
                    round(fitted(glm.statusquo),0))
```

RIFERIMENTI BIBLIOGRAFICI

A.Agresti (2018) "*An introduction to categorical data analysis*", "Wiley series in probability and statistics", 2018, Wiley;

Allison Paul D. (2012) "*Logistic Regression Using SAS:theory and application,Second edition*" , 2012, NC : SAS Institute Inc.;

C.R.Rao (1973), *Linear statistical inference and its applications*, Ed.2, Wiley series in probability and statistics, John Wiley & Sons .

Chambless, L. E., and Boyle, K. E. (1985) Maximum likelihood methods for compelx sample data: Logistic regression and discrete proportional hazards models, *Communications in Statistics: Theory and Methods,*14, 1377-1392

Christensen, R. (1977) "*Log-Linear Models and Logistic Regression"* ,NewYork: Springer.;

Cox, D. R. and Snell, E. J. (1989), *Analysis of Binary Data*, Second Edition. Chapman & Hall, London.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models,* Berlin: Springer.;

Grizzle, J. E., C. F. Starmer, and G. G. Koch (1969*). Analysis of categorical data by linear models. Biometri*cs, **25**, 489-504

Hosmer David W. and Lemeshow Stanley (2000), Applied logistic regression, *Wiley series in probability and statistics*, Wiley interscience publication;

Hosmer, D. W., and Lemehow, S. (1980), A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, **A10**,1043-1069

Huber,J.Peter (1967) Maximum likelihood estimates under nonstandard conditions, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley and Los Angeles, University of California Press,1967,Vol. 1, pag.221-233;

McCullagh, P. , Nelder, J.A. , Generalized linear models , *Monographs on Statistics and Applied Probability 37, Chapman & Hall, Ed.2,* 1989;

Menard,S. (2000) Coefficients of Determination for Multiple Logistic Regression Analysis, *The American Statistician*, 54:1,17-24,2000,Taylor & Francis;

W.W.Hauck Jr. & A. Donner, (1977) Wald's Test as Applied to Hypotheses in Logit Analysis, *Journal of the American Statistical Association* , Vol 72, Issue 360°, Taylor & Francis;

https://stats.idre.ucla.edu/r/dae/logit-regression/

https://cran.r-project.org/doc/contrib/Ricci-regression-it.pdf

https://cran.r-project.org/doc/contrib/DellOmodarme-esercitazioni-R.pdf

https://cran.r-project.org/doc/contrib/Frascati-FormularioStatisticaR.pdf

http://utenti.dises.univpm.it/palomba/Mat/LogitProbit.pdf