Il plebiscito cileno del 1988

ANALISI STATISTICA IN R DELLE INTENZIONI DI VOTO DEL POPOLO CILENO SUL PLEBISCITO RIGUARDANTE LA RIELEZIONE DI PINOCHET.

Gargiulo Sabato | Lab. analisi di mercato | 12 luglio 2017

SOMMARIO

INTRODUZIONE AL DATASET

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.

La FLASCO, nel bimestre Aprile-Maggio 1988 ha condotto un sondaggio sulle intenzioni di voto del popolo cileno.[1]

Carichiamo il file CSV in R e indaghiamo la struttura del dataset usando la funzione str().

```
>chile=read.csv("https://raw.github.com/vincentarelbundock/Rdatasets/master/
csv/car /Chile.csv")
```

```
> str(chile)

  'data.frame':      2700 obs. of  8 variables:
  $ region    : Factor w/ 5 levels "C","M","N","S",..: 3 3 3 3 3 3 3 3 3 3 .
..
  $ population: int  175000 175000 175000 175000 175000 175000 175000 175000
 175000    175000 ...
  $ sex       : Factor w/ 2 levels "F","M": 2 2 1 1 1 1 2 1 1 2 ...
  $ age       : int  65 29 38 49 23 28 26 24 41 41 ...
  $ education : Factor w/ 3 levels "P","PS","S": 1 2 1 1 3 1 2 3 1 1 ...
  $ income    : int  35000 7500 15000 35000 35000 7500 35000 15000 15000 150
00 ...
  $ statusquo : num  1.01 -1.3 1.23 -1.03 -1.1 ...
  $ vote      : Factor w/ 4 levels "A","N","U","Y": 4 2 4 2 2 2 2 2 3 2 ...
```

Il dataset è quindi composto da 2700 osservazioni di 8 variabili, di cui 4 quantitative e 4 qualitative.

Le variabili sono:

- **"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).

[1] https://raw.github.com/vincentarelbundock/Rdatasets/master/csv/car/Chile.csv

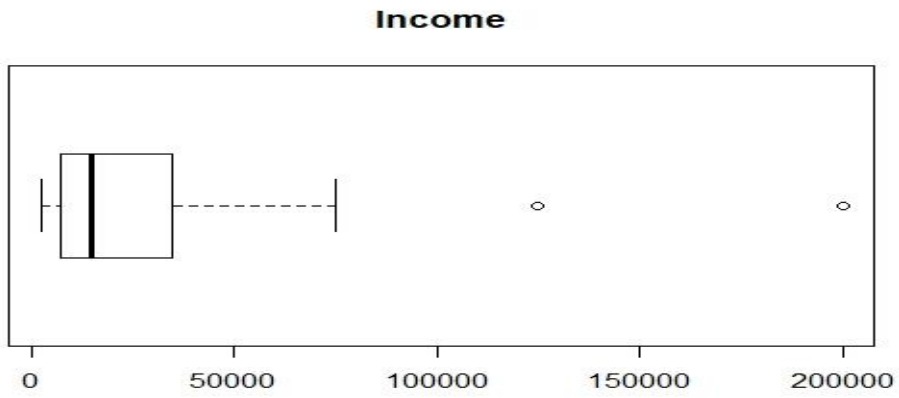
ANALISI DESCRITTIVA MULTIVARIATA

Applichiamo la funzione summary() per avere informazioni sommarie sul dataset.

```
> summary(chile)
 region    population     sex          age         education      income          statusquo          vote
 C :600   Min.   :  3750   F:1379   Min.   :18.00   P   :1107   Min.   :  2500   Min.   :-1.80301   A   :187
 M :100   1st Qu.: 25000   M:1321   1st Qu.:26.00   PS  : 462   1st Qu.:  7500   1st Qu.:-1.00223   N   :889
 N :322   Median :175000            Median :36.00   S   :1120   Median : 15000   Median :-0.04558   U   :588
 S :718   Mean   :152222            Mean   :38.55   NA's:  11   Mean   : 33876   Mean   : 0.00000   Y   :868
 SA:960   3rd Qu.:250000            3rd Qu.:49.00               3rd Qu.: 35000   3rd Qu.: 0.96857   NA's:168
          Max.   :250000            Max.   :70.00               Max.   :200000   Max.   : 2.04859
                                    NA's   :1                   NA's   :98       NA's   :17
```

Uno strumento utile per visualizzare graficamente i dati appena trovati è il BOXPLOT. Quest'ultimo ci consente di confrontare fenomeni differenti e sintetizzarli in una distribuzione di frequenza; Esso si basa su 5 valori noti (Quartili, min e max).

```
> boxplot(chile$income,horizontal=1, main="Income")
```

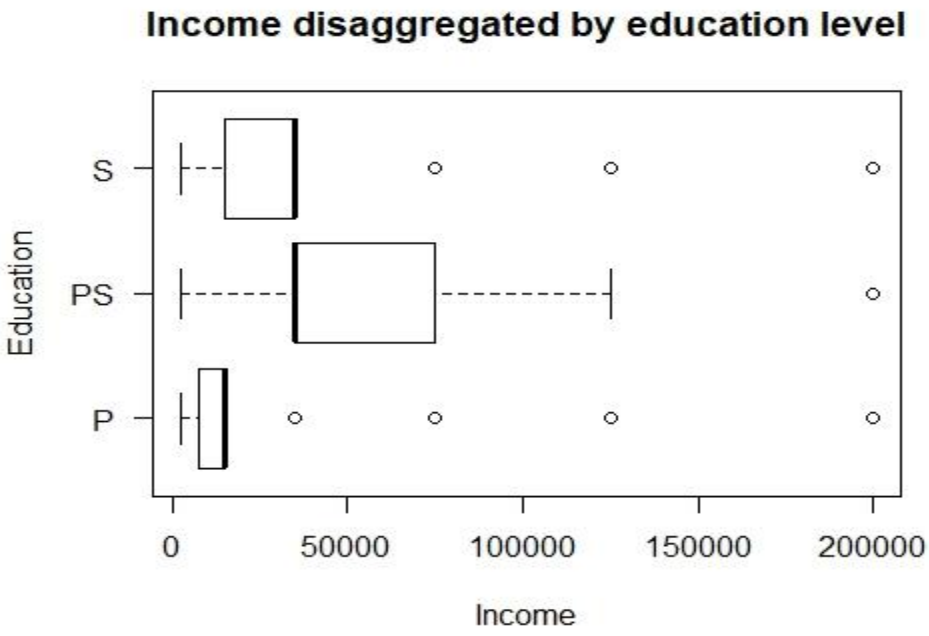


Per quanto riguarda la variabile "Income", il boxplot indica che la mediana (modalità che bipartisce la distribuzione di frequenza) è pari a 15000. La distribuzione presenta un' asimmetria positiva.

In particolare notiamo che la media è significativamente più grande della mediana. Osserviamo anche dei valori eccezionalmente grandi. (Si trovano oltre il valore cardine H_2)

Complichiamo un po' le cose. Ci interessa condizionare il boxplot precedente rispetto a fatt ori come età o livello d'istruzione.

```
> boxplot(chile$income~chile$education, horizontal=TRUE, ylab="Education", x
lab="Income", las=1, main="Income disaggregated by education level")
```



Abbiamo quindi condizionato il boxplot precedente rispetto al livello d'istruzione.
I risultati sono conformi alle aspettative.

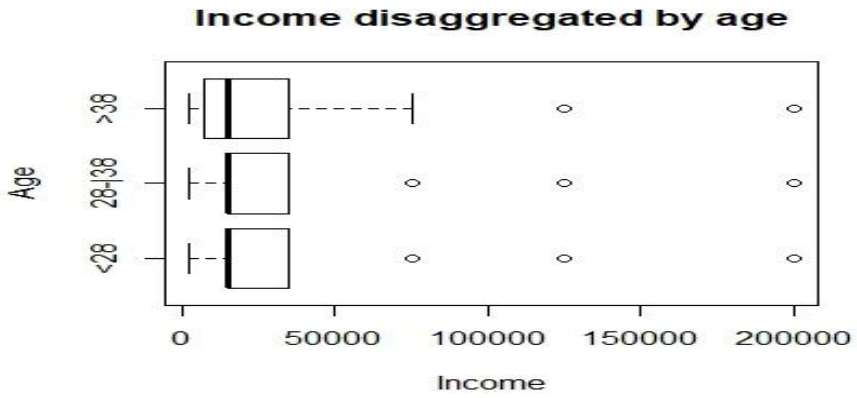
```
> aggregate(income~education,chile,mean)

 education   income
1         P 17534.93
2        PS 69027.45
3         S 35851.49
```

Coloro i quali hanno un titolo di studio "Post-secondario" hanno un reddito, IN MEDIA, molto più elevato rispetto a chi ha un titolo di studio inferiore.

La ricchezza è influenzata anche dall'età degli intervistati. Condizioniamo, quindi, il boxplot precedente utilizzato la variabile età.

Per evitare problemi di visualizzazione dividiamo in classi di modalità la variabile età.

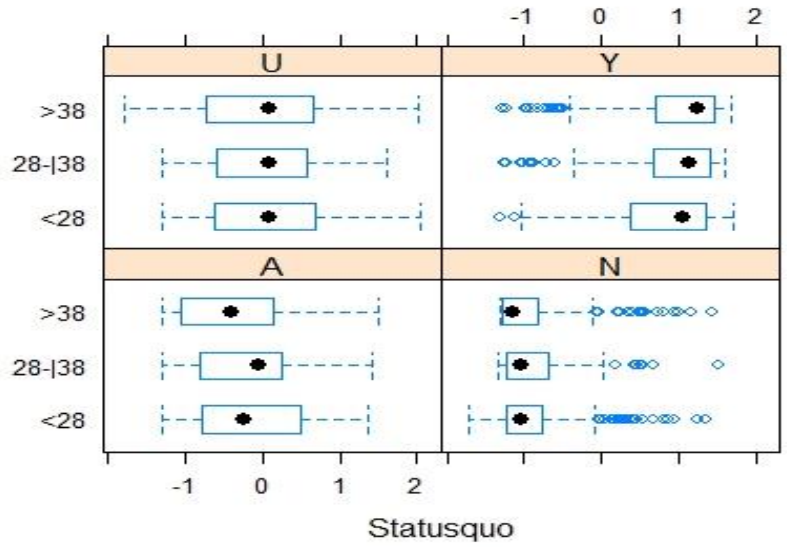
```
> chi=chile
> chi$age[chile$age<=28]="<28"                #DEFINIAMO LE CLASSI
> chi$age[chile$age>28 & chile$age<=38]="28-|38"
> chi$age[chile$age>38]=">38"
> chiage=factor(chi$age)                        #ORDINIAMO I FATTORI
> chiage=ordered(chiage,levels=c("<28","28-|38",">38"))
> boxplot(chile$income~chiage,horizontal=1,ylab="Age",
+         xlab="Income",main="Income disaggregated by age")
```



Carichiamo la libreria **"lattice"**.

```
> library(lattice)
> bwplot(chiage~statusquo|vote,data=chile,horizontal=1,xlab="Statusquo")
```

bwplot() è una versione migliorata di boxplot().



Osserviamo, sul fronte del SI, una distribuzione *asimmetrica negativa*.
I sostenitori di Pinochet,quindi,erano in gran parte soddisfatti dello statusquo.

E' presente anche un numero consistente di voti "controcorrente", a testimonianza del fatt o che c'erano persone poco soddisfatte dello statusquo che speravano in un cambiament o con un secondo mandato.

Sul fronte opposto la situazione si ribalta, infatti è presente una distribuzione *asimmetrica positiva.*
Gli astenuti e gli indecisi invece hanno una mediana prossima allo zero, eccezion fatta per gli astenuti con un'età maggiore di 36 anni; Possiamo azzardare un'ipotesi asserendo che quella fascia di intervistati aveva perso fiducia nel meccanismo rappresentativo e/o politic o.

Ordiniamo i fattori di $education e di $vote in modo "crescente", visto che R di default li dispone in ordine alfabetico.

```
> chile$education=factor(chile$education, levels=c("P","S","PS"))

> chile$vote=factor(chile$vote, levels=c("Y","U","A","N"))
```

Visualizziamo I dati In una tabella attraverso il comando table.

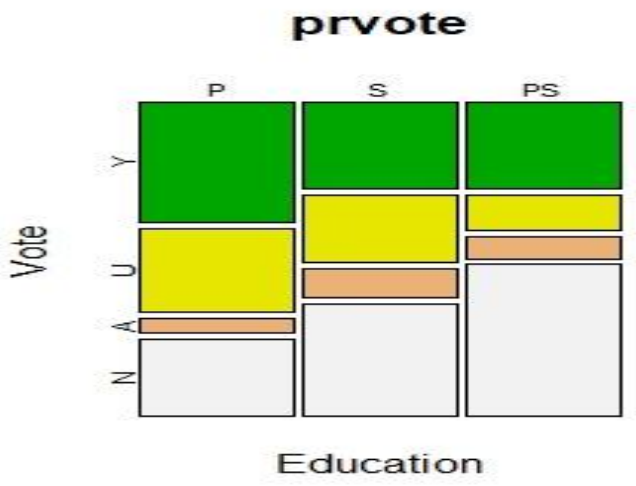
```
> votedu=table(chile$education,chile$vote)
> votedu

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

Gli intervistati con un grado d'istruzione basso tendono ad essere più favorevoli ad una rie lezione di Pinochet.
Proviamo a utilizzare il comando prop.table, con una lieve modifica. In questo caso la perc entuale visualizzata in tabella sarà calcolata sulla base dei livelli d'istruzione (row percenta ges).

```
> prvote=prop.table(tabvote,1)
> prvote

             Y          U          A          N
P  0.40733591 0.28571429 0.05019305 0.25675676
S  0.29675573 0.22614504 0.09828244 0.37881679
PS 0.29680365 0.11872146 0.07305936 0.51141553
```

Visualizziamo i risultati ottenuti col comando plot.

```
> plot(prvote,col=terrain.colors(4),xlab="Education",ylab="Vote")
```



Confrontando le distribuzioni condizionate notiamo che gli intervistati con un basso livello d'istruzione sono i più favorevoli a Pinochet (Y è il 40.7% rispetto al 29.7% degli altri). Invece, sul fronte del No, notiamo che le percentuali aumentano insieme al livello d'istruzione (25.7%, 37.9%,51.1% rispettivamente per "P","S","PS").

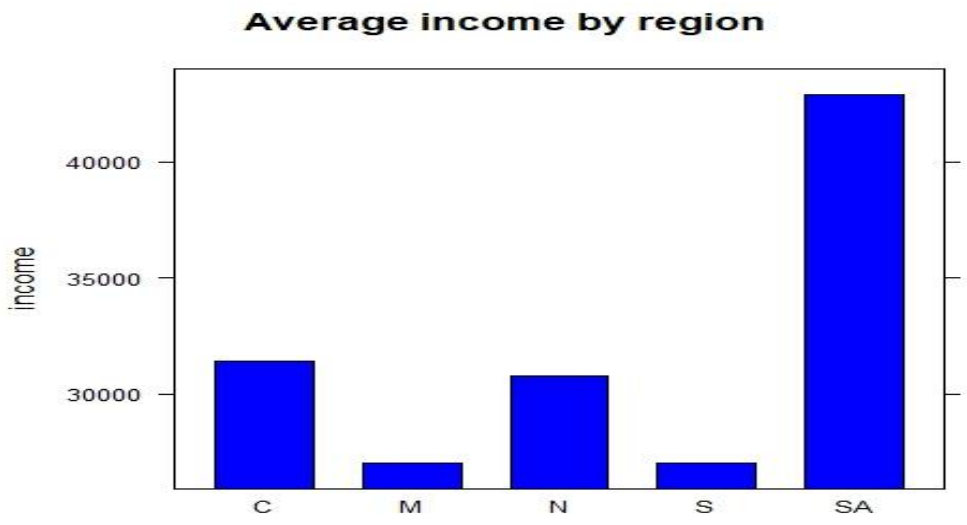
Utilizziamo la funzione aggregate() per analizzare anche la ricchezza media in ogni regione .

```
> aggregate(income~region,chile,mean)

  region   income
1      C 31395.99
2      M 27000.00
3      N 30770.49
4      S 27011.45
5     SA 42918.63
```

Notiamo che la città di Santiago è la zona più ricca, in media, del territorio cileno.

Possiamo facilitare la lettura di questi dati mediante la funzione barchart().

```
> incmean=aggregate(income~region,data=chile,mean)
> barchart(income~region,data=incmean,col="blue",
           main="Average income by region")
```

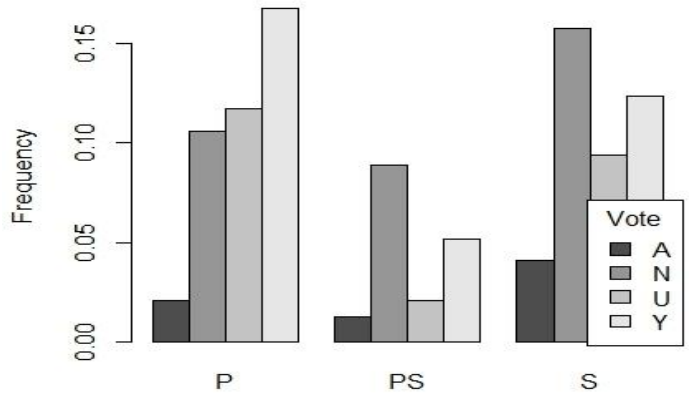


Procediamo nell'analisi bivariata su due mutabili: vote & education.
Utilizziamo dapprima il comando table().

```
> table(chile$vote,chile$education)

     P  PS   S
  A  52  32 103
  N 266 224 397
  U 296  52 237
  Y 422 130 311
> table.eduvot=prop.table(table(chile$vote,chile$education))
```

Attraverso la funzione barplot() costruiamo un grafico che ci permetta di visualizzare meglio i risultati.
La funzione legend() aggiunge al grafico una legenda.

```
> barplot(table.eduvot,beside=1)
> legend("bottomright",legend=levels(chile$vote),fill=gray.colors(4), title=
"Vote")
```



Con il test del chi quadrato di Pearson, si intende valutare se l'associazione tra due o più variabili sia statisticamente significativa.
Due variabili si dicono indipendenti se la distribuzione di probabilità di una non è influenza ta dalla presenza dell'altra.

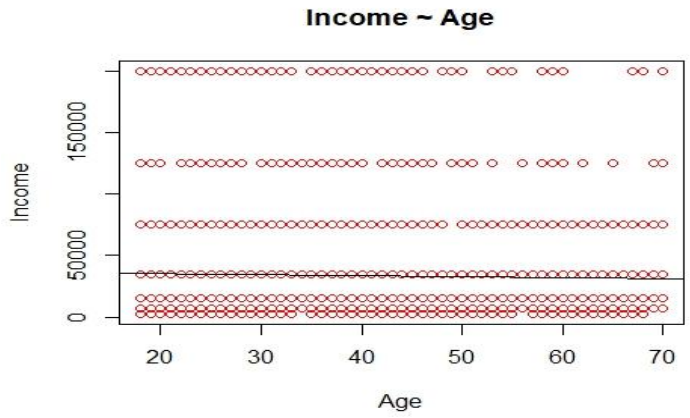
```
> chisq.test(chile$vote,chile$education)

        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 $2,2*10^{-16}$:la probabilità che l'associazione sia dovuta al caso è pari a 0.

Rifiutiamo quindi l'ipotesi nulla, ciò significa che la variabile voto non è indipendente dalla variabile education.

Proviamo a cercare una correlazione tra reddito mensile ed età. A tal fine utilizziamo la funzione plot() insieme al comando abline().
Un grafico a dispersione può rendere visibili vari tipi di correlazione tra variabili, con un certo intervallo di confidenza. Le correlazioni possono essere positive,negative o nulle.

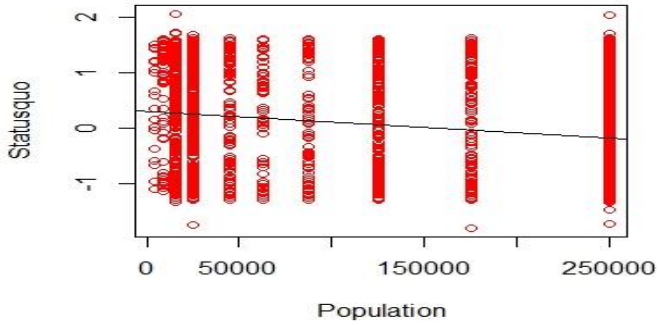
```
> plot(x=chile$age,y=chile$income,col="red",xlab="Age", ylab="Income",
      main="Income ~ Age")
> abline(lm(chile$income~chile$age))
```



Notiamo che all'aumentare della variabile età, il reddito resta pressoché costante. In questo caso si parla di assenza di correlazione.

Prendiamo in esame le variabili "population" & "statusquo".

```
> plot(chile$population,chile$statusquo,col="red",
       xlab="Population",ylab="Statusquo")

> abline(lm(chile$statusquo~chile$population))
```



Mediante la funzione cor.test() possiamo effettuare un test di Pearson sulla correlazione tra due variabili.

```
> cor.test(chile$population,chile$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
```

L'intervallo di confidenza è del 95%, il p value < $2.2*10^{-16}$.
C'è una correlazione negativa. Per quanto riguarda il valore p, data l'ipotesi nulla H_0 (non c'è correlazione),se quest'ultimo è minore del livello di significatività scelto, allora rifiuteremo l'ipotesi nulla.
Nel caso sopra riportato, visto che il p-value è minore del 5%, accettiamo l'ipotesi alternativa(H_1), cioè che è presente un grado di correlazione tra le due variabili.

Effettuiamo la stessa operazione sulle variabili statusquo & income.

```
> cor.test(chile$statusquo,chile$income)

        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
```

REGRESSIONE LOGISTICA

Il modello di regressione logistica semplice appartiene alla famiglia dei GLM (Generalized linear model).
Nel caso in cui la variabile risposta è di tipo dicotomico si parla, appunto, di regressione logistica.
Il modello utilizzato per legare la media dei valori della variabile dipendente dato il valore del predittore, indicata con E(Y|x), è il modello logistico:

$$\pi(x) \equiv E(Y|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Utilizzando la trasformazione logaritmica:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

si ottiene un modello lineare che lega g(x) a x.

Gli obiettivi che ci si pone sono da un lato stabilire quali delle caratteristiche rilevate abbiano un'influenza sulla probabilità di votare "Y", dall'altra di predire la probabilità in corrispondenza delle sopra citate caratteristiche.
In R utilizziamo la funzione glm() che permette di fittare un modello lineare generalizzato una volta specificata la funzione di link desiderata. (In questo caso la funzione è "logit").

Nella regressione logistica, quindi, la variabile dipendente definisce l'appartenenza a un gruppo o meno.
Ciò che ci interessa dunque non è il valore atteso , ma la probabilità che un dato soggetto appartenga o meno ad un gruppo.

Per esprimere la relazione tra variabile dipendente e variabili indipendenti possiamo partire da:

$$P(Y=1) = \alpha + \beta X$$

Questo modello però non è adeguato, poichè i valori della probabilità sono compresi tra 0 e 1, mentre il modello lineare può assumere valori che variano tra +\- infinito.
Per ovviare a questo problema applichiamo la trasformazione esponenziale alla funzione.
Successivamente utilizziamo la trasformazione logaritmica affinchè l'output sia compreso nell'intervallo (0,1).

$$P(Y=1) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

L'odds per Y=1 diventa:

$$odds_{Y=1} = \frac{\frac{e^{(\alpha+\beta X)}}{1+e^{(\alpha+\beta X)}}}{\frac{1}{1+e^{(\alpha+\beta X)}}} = e^{(\alpha+\beta X)}$$

Infine, per le propietà dei logaritmi, osserviamo che il logaritmo naturale dell'odds di Y=1 è funzione lineare della variabile x:

$$\ln(odds_{Y=1}) = \alpha + \beta X$$

Bisogna precisare che la probabilità, l'odds e il logit sono tre modi differenti di esprimere la stessa cosa. La trasformazione in logit serve solo a garantire la correttezza matematica dell'analisi.

Estraiamo dal dataframe originale solo le righe che contengono i voti "Y","N".

```
> s.chi=chile[chile$vote %in% c("Y","N"),]
> str(s.chi)
'data.frame': 1757 obs. of  9 variables:
 $ X         : int  1 2 3 4 5 6 7 8 10 11 ...
 $ region    : Factor w/ 5 levels "C","M","N","S",..: 3 3 3 3 3 3 3 3 3 3 ..
.
 $ population: int  175000 175000 175000 175000 175000 175000 175000 175000
175000 175000 ...
 $ sex       : Factor w/ 2 levels "F","M": 2 2 1 1 1 1 2 1 2 2 ...
 $ age       : int  65 29 38 49 23 28 26 24 41 64 ...
 $ education : Factor w/ 3 levels "P","PS","S": 1 2 1 1 3 1 2 3 1 1 ...
 $ income    : int  35000 7500 15000 35000 35000 7500 35000 15000 15000 1500
0 ...
 $ statusquo : num  1.01 -1.3 1.23 -1.03 -1.1 ...
 $ vote      : Factor w/ 4 levels "A","N","U","Y": 4 2 4 2 2 2 2 2 2 4 ...


> s.chi$vote=factor(s.chi$vote,exclude=1)    ##Escludiamo i valori "A" e "U"
> s.chi$vote=factor(s.chi$vote,exclude=3)
> table(s.chi$vote)

  N   Y
889 868

> s.chi=na.omit(s.chi)                  ##Omettiamo le righe contenenti NA

>num=as.numeric(s.chi$vote)

> str(num)
 num [1:1757] 2 1 2 1 1 1 1 1 1 2 ...
> num=recode(num,"c(1)='0';else='1'")             ##Ricodifichiamo I valori
   [1] 1 0 1 0 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 0 1 1 0 0 1
  [26] 1 1 0 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 0 1 0 0
```

```
> table(num)
Num                          ### Y=1 N=0
  0   1
889 868

> s.chi$vote=num
> str(s.chi)
'data.frame': 1703 obs. of  9 variables:
 $ X         : int  1 2 3 4 5 6 7 8 10 11 ...
 $ region    : Factor w/ 5 levels "C","M","N","S",..: 3 3 3 3 3 3 3 3 3 3 ..
.
 $ population: int  175000 175000 175000 175000 175000 175000 175000 175000
175000 175000 ...
 $ sex       : Factor w/ 2 levels "F","M": 2 2 1 1 1 1 2 1 2 2 ...
 $ age       : int  65 29 38 49 23 28 26 24 41 64 ...
 $ education : Factor w/ 3 levels "P","PS","S": 1 2 1 1 3 1 2 3 1 1 ...
 $ income    : int  35000 7500 15000 35000 35000 7500 35000 15000 15000 1500
0 ...
 $ statusquo : num  1.01 -1.3 1.23 -1.03 -1.1 ...
 $ vote      : num  1 0 1 0 0 0 0 0 0 1 ...
```

Per procedere nella nostra analisi è necessario chiarire il significato di ODDS e ODDS RATIO.
L'ODDS è un modo di esprimere una probabilità mediante un rapporto. Si ottiene calcolando il rapporto tra le frequenze osservate in un livello con le frequenze osservate nell'altro. Ad esempio, se ci sono 40 uomini e 25 donne (N=65) possiamo dire che la probabilità di essere uomini è .615.

Se vogliamo esprimere questa informazione mettendo in relazione le due categorie possiamo ricorrere all'odds. Mediante l'odds la relazione tra uomini e donne è pari a 1.6; Questo equivale a dire che per ogni donna ci sono 1.6 uomini.
L'ODDS RATIO è utilizzato per esprimere la relazione tra due categorie in funzione di un'altra variabile. Si ottiene calcolando il rapporto tra gli odds di una data variabile (ad esempio, la variabile Y) ottenuti per ciascun livello della seconda variabile (ad esempio, X).

Nel nostro caso Y=1 (Voto "Y"), Y=0 (Voto "N"). Se le due categorie sono equiprobabili le frequenze relative sono uguali a .5 , gli odds sono uguali a 1, mentre i logit sono uguali a 0.
Quando il numero di "Yes" è maggiore del numero di "No", gli odds assumono valori superiori a 1, mentre i logit valori superiori allo 0.
Nel caso opposto, gli odds assumono valori minori di 1, mentre i logit hanno valore negativo.
In sintesi, mentre le frequenze realtive hanno un range di variabilità che va da 0 a 1, gli odds hanno un range di variabilità che va da 0 a più infinito, mentre i logit possono variare da meno infinito a più infinito.

Fittiamo un modello in cui la VI è solo statusquo.

```
> summary(glm.statusquo)

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 della funzione ci mostra innanzitutto la distribuzione della devianza dei residui, che sono una misura della bontà di adattamento del modello.

R ci mostra due tipi di devianza. “null deviance” ci mostra quanto efficacemente la variabile dipendente viene prevista da un modello che include la sola intercetta.
Nel nostro caso la devianza passa da 2360 a 737, una riduzione significativa!
La parte successiva ci mostra i coefficienti, lo standard error, la statistica z (chiamata anche statistica “WALD”) e, infine, i rispettivi p-value. Questi ultimi sono tutti inferiori al livello di significatività fissato (5%), quindi le var. indipendenti sono tutte statisticamente significative.
I coefficienti della regressione logistica ci indicano la variazione nel log-odds (**LOGIT)** dell’outcome per una variazione unitaria del predittore.
Ad esempio, per ogni variazione positiva unitaria in “statusquo”, il log-odds di Y (rispetto a N) aumenta di 3.1819;

I risultati del modello ci mostrano che: ln(odds y=1)=0.22+3.1819(statusquo);
Ciò significa che quando la variabile statusquo assume valore 0, il logaritmo dell’odds è uguale a 0.22; La probabilità di votare “YES”, con statusquo=0, è pari a:

```
> exp(0.2237)
[1] 1.250696
```

Attraverso lo stesso procedimento possiamo calcolare la variazione del log-odds conseguente ad una variazione dello statusquo.

```
> exp(3.1819)
[1] 24.09249
```

Per un Δstatus=+1 la probabilità che l'intervistato voti favorevolmente è ΔProb.YES=+24.092%

Per ottenere l'intervallo di confidenza utilizziamo la funzione confint(), ricordandoci però di utilizzare anche exp().

```
> exp(confint(glm.statusquo))

Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept)  1.028233  1.52861
statusquo   18.393753 32.33075
```

Uno strumento utile per definire il "Goodness of fit" del modello è il **test di Hosmer-Lemeshow.**

Installiamo la libreria **"ResourceSelection"** e carichiamola.

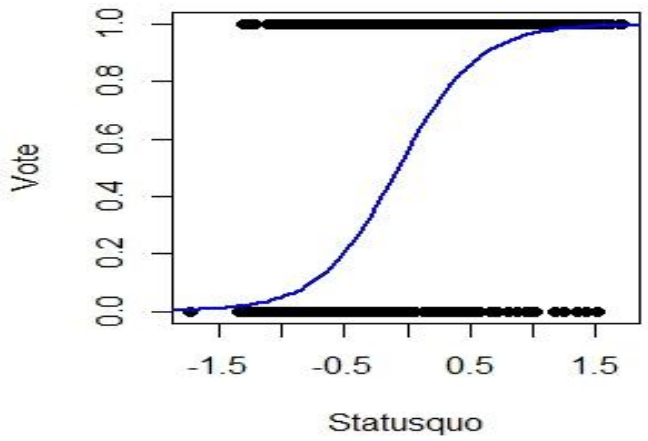
```
> install.packages("ResourceSelection")
> library(ResourceSelection)

> hoslem.test(s.chi$vote,fitted(glm.statusquo))

        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
```

Il nostro modello sembra fittare bene i dati perchè non è presente una differenza significativa tra il modello e i dati osservati.
Un alto valore del Chi-quadrato (p-value<0.05) indica uno scarso fitting, e viceversa.
Il risultato appena ottenuto è solo una parte del procedimento di verifica della bontà del modello.
Questo test non lavora bene per dataset con numerosità campionaria molto grande o molto piccola.

Procediamo adesso a rappresentare graficamente il modello logistico.

```
> range(s.chi$statusquo)

[1] -1.72594  1.71355            ##Creiamo un intervallo per l'asse delle x
> xstatus=seq(-2,2,0.05)

> ystat=predict(glm.statusquo,list(statusquo=xstatus),
                type="response")                    ##Stima della probabilità
```

```
> plot(s.chi$statusquo,s.chi$vote,pch=16,xlab="STATUSQUO",ylab="Vote") #Plot
> lines(xstatus,ystat,col="blue",lwd=2)      #Aggiungiamo la curva
```



Per ogni livello della variabile Statusquo, otteniamo la probabilità stimata che l'intervistato sia favorevole alla rielezione di Pinochet.
La probabilità del voto opposto si ottiene banalmente con (1-x).

Il fitting della regressione logistica avviene attraverso il metodo della massima verosimiglianza (i parametri stimati sono quelli che massimizzano la funzione di verosimiglianza dei dati osservati).
L'indice R^2 di McFadden[2] è definito come:

$$R^2_{\text{McFadden}} = 1 - \frac{log(L_c)}{log(L_{\text{null}})}$$

Dove L_c indica il valore della verosimiglianza del modello corrente, e L_{null} indica lo stesso valore calcolato sul modello nullo, cioè quel modello con la sola intercetta.
Possiamo calcolare lo pseudo R^2 per verificare il grado di adattamento del modello stimato :

```
> pseudoR2<-function(mod) {1-(deviance(mod)/mod$null.deviance)}
> pseudoR2(glm.statusquo)
[1] 0.687637
```

Un altro metodo per valutare la capacità predittiva del modello è la costruzione di una tabella di classificazione. In questa tabella vengono confrontati i dati iniziali con quelli previsti dal modello logistico.

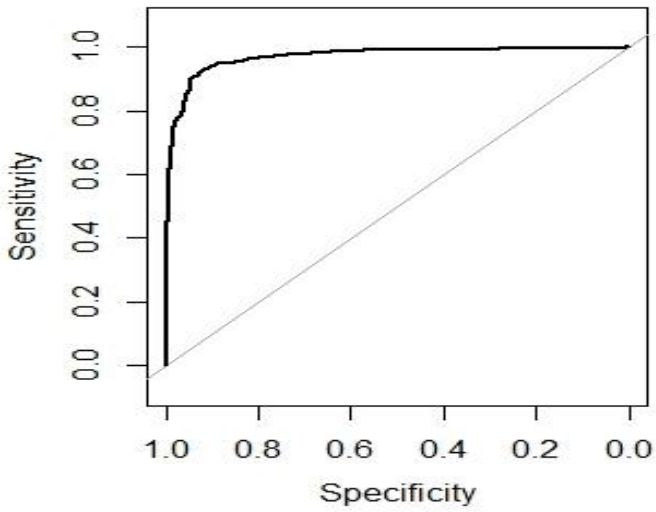
[2] McFadden,D. "Conditional logit analysis of qualitative chioce behavior" (1974)

```
> classDF<-data.frame(response=s.chi$vote,predicted=
                      round(fitted(glm.statusquo),0))

> xtabs(~predicted+response,data=classDF)
         response
predicted   0   1
        0 810  75
        1  57 761
```

Il modello ha previsto correttamente 810 "N" e 761 "Y".

Un ultimo strumento per il controllo di un modello logistico predittivo è la curva ROC. Utilizziamo la libreria **"pROC"** che rende molto semplice questo tipo di analisi.

```
> install.packages("pROC")
> library(pROC)
> g=roc(s.chi$vote~s.chi$statusquo)
> plot(g)
```



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 una curva ROC prossima alla retta a 45°.

```
> auc(g)
Area under the curve: 0.9699
```

REGRESSIONE LOGISTICA CON PIU' VARIABILI ESPLICATIVE

Definiamo un modello logistico che contiene tutte le variabili del dataset.

```
> glm.full=glm(formula = vote ~ ., family = binomial(link = logit), data = s
.chi)
```

Utilizziamo la funzione step() che ci permette di effettuare una stepwise regression.
Si tratta di un sistema per semplificare una regressione multipla. La regressione stepwise è un metodo di selezione delle variabili indipendenti allo scopo di selezionare un set di pre dittori che abbiano la migliore relazione con la variabile dipendente.
In questo caso utilizzeremo il metodo **BACKWARD,** che inizia con un modello che compre nde tutte le variabili e procede, step by step, ad eliminare le variabili partendo da quella con l'associazione meno significativa con la variabile dipendente.

L' AIC è l'**Akaike Information Criterion,** un indice che consente di valutare variazioni ne ll'adattamento di un modello rispetto a modifiche dello stesso, sebbene questo metodo sia molto criticato.
Esso serve solo a comparare due modelli; Verrà scelto il modello con un AIC minore rispetto agli altri.

```
> step(glm.full)[3]

vote ~ sex + education + statusquo       ###il modello scelto dall'algoritmo
                                              con AIC minore. AIC=718.2
            Df Deviance     AIC
<none>           708.24  718.24
- sex        1   716.37  724.37
- education  2   726.04  732.04
- statusquo  1  2264.31 2272.31

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

Coefficients:
(Intercept)         sexM  educationPS
     1.0153      -0.5742      -1.1074
 educationS    statusquo
    -0.6828       3.1689

Degrees of Freedom: 1702 Total (i.e. Null);  1698 Residual
Null Deviance:           2360
Residual Deviance: 708.2    AIC: 718.2

> glm.sta.edu.sex=glm(formula=vote~sex+education+statusquo,family=binomial,
                      data=s.chi)

> summary(glm.sta.edu.sex)

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

[3] Per i risultati si veda l'appendice.

```
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
```

Per quanto riguarda la valutazione degli OR, il predittore education è a 3 livelli.
In questo caso il livello "P" viene assunto come categoria di riferimento. Si ha:

$$\text{OR(educationPS="PS",educationP="P")} = \exp(-1.1074) = 0.3304$$

$$\text{OR(educationS="S",education="P")} = \exp(-0.6828) = 0.5052$$

Proviamo a confrontare questo modello con la sua versione semplificata introdotta in precedenza.
Attraverso la statistica G, che in questo caso si distribuisce $\sim\chi^2(4)$, operiamo un confronto tra i due modelli.
Utilizziamo, a tal fine, la funzione anova().

```
> anova(glm.statusquo,glm.sta.edu.sex,test="Chisq")
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 dell funzione, notiamo che la devianza residua è minore del modello "1", ma soprattutto che Pr è minore del livello di significatività fissato (5%). Ciò significa che aggiungendo altre variabili esplicative al modello, quest'ultimo è stato "migliorato".

Procediamo nuovamente col test di Hosmer-Lemeshow.

```
> library(ResourceSelection)
> hoslem.test(s.chi$vote,fitted(glm.sta.edu.sex))

        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
```

In questo caso la differenza tra i due p-value è risibile, anche se il modello precedente ha ottenuto un punteggio migliore.

Costruiamo una tabella di classificazione per valutare la capacità predittiva del modello.

```
> classDF1=data.frame(response=s.chi$vote,predicted=
                        round(fitted(glm.sta.edu.sex),0))
> xtabs(~predicted+response,data=classDF1)
          response
predicted   0   1
        0 809  64
        1  58 772
```

Anche in questo caso la differenza è risibile, sebbene il modello "2" abbia ottenuto uno score migliore.
Possiamo calcolare lo pseudo R^2 per verificare il grado di adattamento del modello stimato

```
> pseudoR2<-function(mod) {1-(deviance(mod)/mod$null.deviance)}
> pseudoR2(glm.sta.edu.sex)
[1] 0.6999373
```

Andrebbero effettuate analisi più approfondite per valutare con precisione la composizione del voto e poter fare delle previsioni. Sebbene, nel corso di quest'analisi, abbiamo visto come le variabili statusquo e education siano quelle che hanno avuto un maggior impatto nel determinare la scelta di voto degli intervistati cileni.

Riferimenti a fonti esterne

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

https://moodle2.units.it/pluginfile.php/58315/mod_resource/content/1/Laboratorio_Logit_1.pdf

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

https://www.medcalc.org/manual/logistic_regression.php

APPENDICE

```
> step(glm.full)


Start:  AIC=728.71
vote ~ X + region + population + sex + age + education + income +
    statusquo

             Df Deviance     AIC
- region      4   707.03  725.03
- age         1   702.72  726.72
- population  1   703.25  727.25
- income      1   703.40  727.40
- X           1   703.48  727.48
<none>            702.71  728.71
- sex         1   710.38  734.38
- education   2   713.03  735.03
- statusquo   1  2160.20 2184.20

Step:  AIC=725.03
vote ~ X + population + sex + age + education + income + statusquo

             Df Deviance     AIC
- age         1   707.05  723.05
- X           1   707.21  723.21
- population  1   707.49  723.49
- income      1   707.84  723.84
<none>            707.03  725.03
- sex         1   714.77  730.77
- education   2   717.29  731.29
```

```
- statusquo   1  2177.88 2193.88

Step:  AIC=723.05
vote ~ X + population + sex + education + income + statusquo

             Df Deviance     AIC
- X           1   707.22  721.22
- population  1   707.52  721.52
- income      1   707.84  721.84
<none>            707.05  723.05
- sex         1   714.83  728.83
- education   2   719.07  731.07
- statusquo   1  2193.90 2207.90

Step:  AIC=721.22
vote ~ population + sex + education + income + statusquo

             Df Deviance     AIC
- population  1   707.55  719.55
- income      1   708.08  720.08
<none>            707.22  721.22
- sex         1   714.97  726.97
- education   2   719.21  729.21
- statusquo   1  2195.43 2207.43

Step:  AIC=719.55
vote ~ sex + education + income + statusquo

            Df Deviance     AIC
- income     1   708.24  718.24
<none>           707.55  719.55
- sex        1   715.67  725.67
- education  2   719.28  727.28
- statusquo  1  2249.45 2259.45

Step:  AIC=718.24
vote ~ sex + education + statusquo

            Df Deviance     AIC
<none>           708.24  718.24
- sex        1   716.37  724.37
- education  2   726.04  732.04
- statusquo  1  2264.31 2272.31

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

Coefficients:
(Intercept)         sexM  educationPS
     1.0153      -0.5742      -1.1074
 educationS    statusquo
    -0.6828       3.1689

Degrees of Freedom: 1702 Total (i.e. Null);  1698 Residual
Null Deviance:     2360
Residual Deviance: 708.2      AIC: 718.2
```