

# **MULTINOMIAL ALZHEIMER'S CASE**

By Elka Segura Sanchez

**Anno Accademico 2022/2023**

## **ABSTRACT**

Tutti sanno che il fumo fa male al cuore e ai polmoni. Tuttavia, il danno che si verifica nel cervello riceve spesso meno attenzione di quanto dovrebbe, sia dai fumatori che dai medici. Sono stati condotti molti studi sull'influenza del fumo sullo sviluppo delle demenze e nel caso specifico dell'Alzheimer. La demenza è spesso il risultato dell'interazione di fattori genetici, fattori ambientali e fattori legati allo stile di vita. La demenza correlata all'Alzheimer di solito inizia in modo insidioso e progredisce lentamente e progressivamente per un periodo di diversi anni; la memoria episodica è una delle prime aree cognitive colpite. Attualmente non esiste un trattamento che possa curare la demenza o alterarne la progressione progressiva. Fino a quando non si troverà un trattamento efficace, le strategie per ridurre la frequenza della malattia, ritardarne l'insorgenza e ridurre l'onere devono concentrarsi sull'individuazione e l'azione sui fattori di rischio modificabili. I fattori di rischio modificabili per la demenza includono il fumo, uno stile di vita sedentario, l'ipertensione, l'obesità, il diabete, bassi livelli di attività cognitiva o un basso livello di istruzione. Questo lavoro consiste nel determinare se l'effetto del fumo influenza lo sviluppo dell'Alzheimer e di altre demenze. Come strumento statistico, per lo studio viene utilizzata la regressione multinomiale con el Software Rstudio.

## PROCEDURA

La regressione logistica multinomiale è un'estensione della regressione logistica binomiale, questa tecnica è utilizzata quando la variabile dipendente è categoriale (o nominale). Per la regressione logistica binaria il numero di variabili dipendenti è due, mentre il numero di variabili dipendenti per la regressione logistica multinomiale è superiore a due. La regressione logistica multinomiale funziona allo stesso modo di altri tipi di regressione: sta cercando una relazione tra le variabili indipendenti e dipendenti. In questo caso per simulare la regressione multinomiale si utilizza il set di dati Alzheimer dalla libreria "coin" del pacchetto R. Questo set di dati ha 538 osservazioni e tre variabili: "smoking", "disease" e "gender". Il riferimento bibliografico del caso studio è il seguente: Fonte: Salib, E. and Hillier, V. (1997). A case-control study of smoking and Alzheimer's disease. *International Journal of Geriatric Psychiatry* 12(3), 295–300. doi: 10.1002/(SICI)1099-1166(199703)12:3<295::AID-GPS476>3.0.CO;2-3.

## Modello Multinomial

Sia  $Y$  una variabile di risposta categoriale con  $J$  categorie e siano

$$\pi_1, \pi_2, \dots, \pi_J$$

le probabilità associate, tali che la somma di tutti i

$$\pi_j = 1$$

Questo modello è costruito prendendo come risposta base una delle categorie, ad esempio l'ultima,  $J$ , e rispetto ad essa viene definito un modello logit:

$$\frac{\log \pi_j}{\log \pi_J} = \beta_{0j} + \beta_{1j}x \text{ con } j = 1 \dots J - 1 \text{ categorie (1.1)}$$

Il modello ha  $J-1$  equazioni con i propri parametri. Ognuno di questi parametri esprime l'effetto rispetto alla categoria di riferimento. I parametri del modello sono stimati utilizzando il metodo della massima verosimiglianza come nel caso della regressione logistica. Dopo aver calcolato questi parametri, le probabilità previste di ciascuna categoria possono essere calcolate isolando

$$\pi_j$$

dall'equazione precedente:

$$\pi_j = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + \sum_{h=1}^{J-1} e^{(\beta_0 h + \beta_1 h x)}} \quad (1.2)$$

### Librerie da utilizzare:

```
library(coin)
library(nnet)
library(glmnet)
library(VGAM)
```

Il set di dati ha tre variabili categoriali. La variabile dipendente è “disease”, che ha tre categorie di malattie: “Alzheimer”, “Other dementias” e “Other diagnoses”. Le variabili predittive sono “Gender” e “Smoking”. Il genere è un fattore con livelli “Female” e “Male”. Smoking è un fattore con livelli “None”, “<10”, “10-20” e “>20” (sigarette al giorno). Questo studio arriva a studiare se le variabili predittive “Smoking” influenzano lo sviluppo dell’Alzheimer e Altre demenze.

Il dataset mostra n=538 individui, di cui 338 femmine e 200 maschi. Ci sono 198 individui con Alzheimer, 164 con altre demenze e 176 con altre diagnosi. Per quanto riguarda il fattore fumo, 309 individui non fumano, 28 fumano meno di 10 sigarette al giorno, 110 tra 10-20 sigarette al giorno e 91 individui fumano più di 20 sigarette al giorno.

```
data_alzhe<-data.frame(alzheimer)
summary(data_alzhe)

##   smoking      disease      gender
## None :309   Alzheimer   :198   Female:338
## <10  : 28   Other dementias:164   Male  :200
## 10-20:110   Other diagnoses:176
## >20  : 91
```

### Tabella di Frequenza

Questo output mostra la tabella delle frequenze, con le frequenze distribuite per ciascuna categoria di ciascun fattore. Le frequenze più alte sono persone con Alzheimer, donne e non fumatori, che rappresentano circa il 17%(90) del

totale; e le persone con “Altre diagnosi”, donne e non fumatori, che rappresentano circa il 15%(81) del totale.

```
knitr::kable(
  table(alzheimer), booktabs = TRUE,
  caption = '1.1 Alzheimer Case frequency table.'
)
```

*1.1 Alzheimer Case frequency table.*

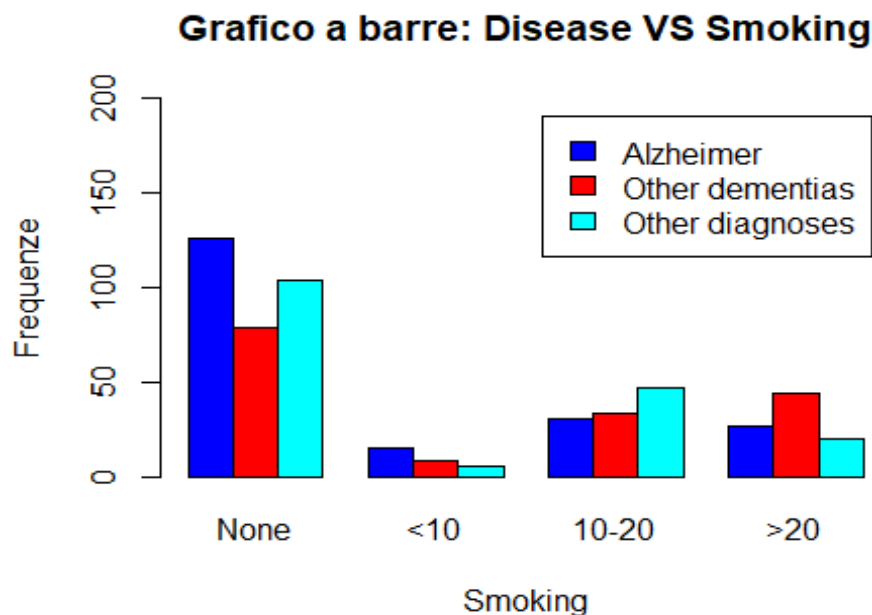
smoking	disease	gender	Freq
None	Alzheimer	Female	91
<10	Alzheimer	Female	7
10-20	Alzheimer	Female	15
>20	Alzheimer	Female	21
None	Other dementias	Female	55
<10	Other dementias	Female	7
10-20	Other dementias	Female	16
>20	Other dementias	Female	9
None	Other diagnoses	Female	80
<10	Other diagnoses	Female	3
10-20	Other diagnoses	Female	25
>20	Other diagnoses	Female	9
None	Alzheimer	Male	35
<10	Alzheimer	Male	8
10-20	Alzheimer	Male	15
>20	Alzheimer	Male	6
None	Other dementias	Male	24
<10	Other dementias	Male	1
10-20	Other dementias	Male	17
>20	Other dementias	Male	35
None	Other diagnoses	Male	24
<10	Other diagnoses	Male	2
10-20	Other diagnoses	Male	22
>20	Other diagnoses	Male	11

## Grafici

Per una migliore interpretazione del set di dati e delle sue caratteristiche, viene mostrato i seguenti grafici. Nel “Grafico a barre: Disease VS Smoking” le frequenze più alte per ciascuna delle malattie si concentrano nel segmento dei pazienti che non fumano, cioè c’è una maggiore incidenza di Alzheimer e demenza nelle persone che non fumano indipendentemente dal genere.

```
tab_smok <- table(alzheimer$disease, alzheimer$smoking)
```

```
barplot(tab_smok,  
  main="Grafico a barre: Disease VS Smoking",  
  xlab="Smoking",  
  ylab="Frequenze",  
  legend = rownames(tab_smok),  
  ylim = c(0, 200),  
  col=c("blue", "red", "cyan"),  
  beside=TRUE)
```



Nel “Grafico a barre: Disease VS Gender” mostra in generale, che la più alta incidenza di malattie è presente nel genere femminile, questo risultato influisce sul fatto che c’è una percentuale maggiore di donne rispetto agli uomini.

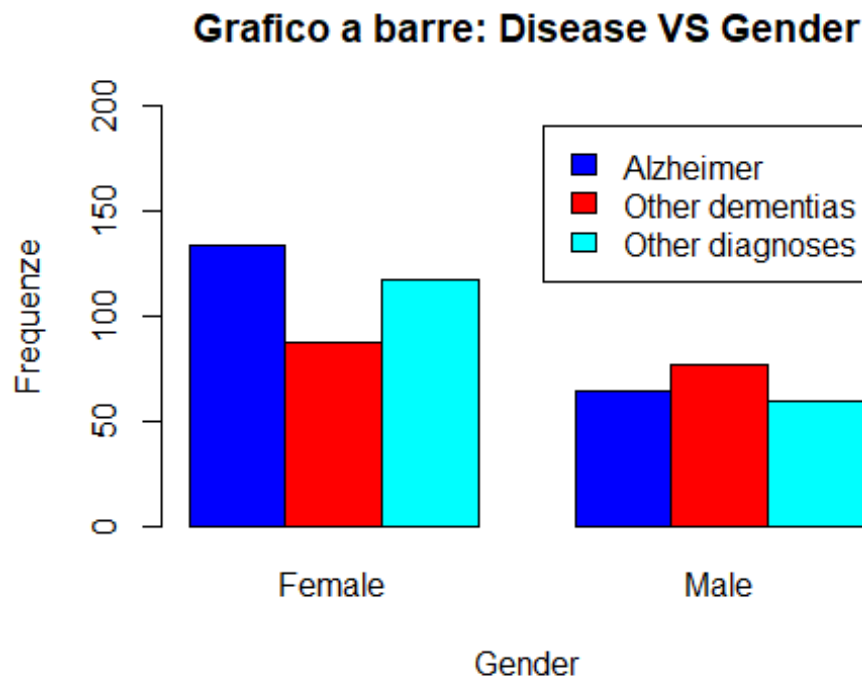
```
tab_smok1 <- table(alzheimer$disease, alzheimer$gender)
```

```
barplot(tab_smok1,
```

```

main="Grafico a barre: Disease VS Gender",
xlab="Gender",
ylab="Frequenze",
legend = rownames(tab_smok1),
ylim = c(0, 200),
col=c("blue", "red", "cyan"),
beside=TRUE)

```

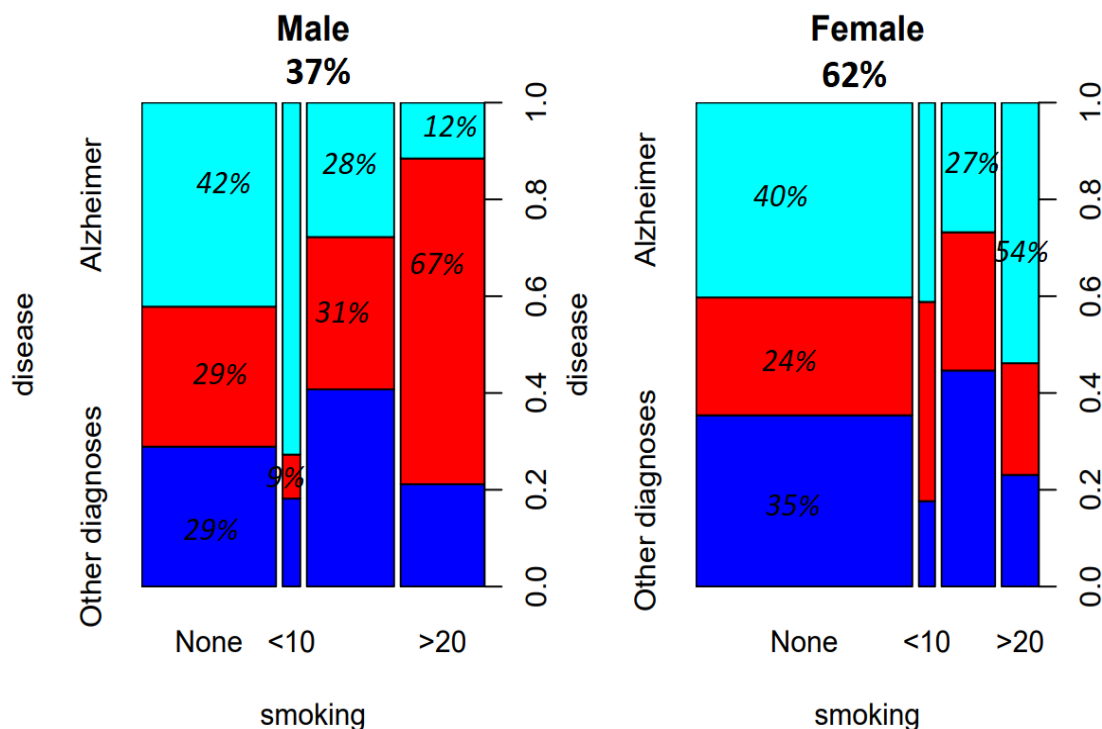


Il 62% delle persone nello studio sono donne, mentre il 37% sono uomini. Sia per gli uomini che per le donne, la percentuale più alta nella categoria dei non fumatori delle tre malattie è l'Alzheimer con rispettivamente il 42% e il 40%. Uno dei dettagli da evidenziare è che per gli uomini, quando i livelli di "Smoking" aumentano, la frequenza di "Alzheimer" diminuisce. In un altro caso e per la stessa fascia di genere, all'aumentare del consumo quotidiano di sigarette, aumenta la frequenza delle "Other dementias". Nel caso del segmento femminile, non esiste una distribuzione equi-ripartito, tra le categorie "Smoking", la maggioranza non fuma ed è qui che si riscontra la frequenza più alta, che rappresenta il 42% del totale, con una frequenza di 226 di  $n=538$ . Tutto ciò fornisce una prospettiva dei dati e della loro natura, nonché la loro distribuzione in ciascuna categoria.

```

op <- par(no.readonly = TRUE) # save current settings
layout(matrix(1:2, ncol = 2))
spineplot(disease ~ smoking, data = alzheimer,
          subset = gender == "Male", main = "Male", col = c("blue", "red", "cyan"))
spineplot(disease ~ smoking, data = alzheimer,
          subset = gender == "Female", main = "Female", col = c("blue", "red", "cyan"))

```



## Costruzione del modello

Le stime dei parametri logistici in funzione dell'inclusione o esclusione dei variabili "Gender" e "Smoking" sono ottenute con la funzione `multinom()` del pacchetto `nnet`. Nella prima opzione si stima il modello saturato, con il quale si ottiene un modello di maggiore overfit ai valori da cui è stato costruito. Nella seconda e terza opzione, vengono stimati i modelli che includono rispettivamente il "Gender" e "Smoking". Nella quarta e ultima opzione, le variabili sesso e smoking sono incluse contemporaneamente nel modello. "Other dementias", "None" e "Female" sono le categorie di riferimento delle tre variabili, e il comando in R che esegue questa azione è il seguente:



```

attach(data_alzhe)
data_alzhe$disease <- relevel(data_alzhe$disease, ref = "Other dementias")
options(contrasts=c("contr.treatment", "contr.poly"))
alzhe_modS<-multinom(disease~gender*smoking, data=data_alzhe) #Modello Saturo

## # weights:  27 (16 variable)
## initial  value 591.053411
## iter   10 value 562.535131
## final   value 561.224741
## converged

alzhe_mod0<-multinom(disease~1,data=data_alzhe) #Modelo null

## # weights:  6 (2 variable)
## initial  value 591.053411
## final   value 589.407756
## converged

alzhe_mod1<-multinom(disease~gender, data=data_alzhe) #Modello Gender

## # weights:  9 (4 variable)
## initial  value 591.053411
## final   value 584.617852
## converged

alzhe_mod2<-multinom(disease~smoking, data=data_alzhe) #Modello Smoking

## # weights:  15 (8 variable)
## initial  value 591.053411
## iter   10 value 575.814595
## final   value 575.783014
## converged

alzhe_mod3<-multinom(disease~gender+smoking, data=data_alzhe) #Modello Gender
+ Smoking

## # weights:  18 (10 variable)
## initial  value 591.053411
## iter   10 value 573.605715
## final   value 573.189760
## converged

```

I diversi modelli logit vengono aggiustati con la categoria di base “Other dementias” e confrontati in termini delle loro differenze di devianza, cioè attraverso le differenze nella verosimiglianza:

$$D = -2(\log(L(\mu_1, y)) - \log(L(\mu_2, y))) \quad (1.3)$$

```

deviance(alzhe_mod1)-deviance(alzhe_modS)

## [1] 46.78622

```

```

deviance(alzhe_mod2)-deviance(alzhe_modS)

## [1] 29.11655

deviance(alzhe_mod3)-deviance(alzhe_modS)

## [1] 23.93004

deviance(alzhe_mod0)-deviance(alzhe_modS)

## [1] 56.36603

summary(alzhe_mod3)

## Call:
## multinom(formula = disease ~ gender + smoking, data = data_alzhe)
##
## Coefficients:
##              (Intercept) genderMale smoking<10 smoking10-20 smoking>20
## Alzheimer              0.5988510 -0.4549438  0.2189701   -0.4602867  -0.8260256
## Other diagnoses        0.4079088 -0.4588215 -0.6872985    0.1813765  -0.9331809
##
## Std. Errors:
##              (Intercept) genderMale smoking<10 smoking10-20 smoking>20
## Alzheimer              0.1591086  0.2271575  0.4638437    0.2954492  0.2911728
## Other diagnoses        0.1647381  0.2338396  0.5916437    0.2780938  0.3156220
##
## Residual Deviance: 1146.38
## AIC: 1166.38

```

L'output di riepilogo contiene la tabella dei coefficienti e una tabella degli errori standard. Ogni riga della tabella dei coefficienti corrisponde alle J-1 equazioni del modello. La prima riga rappresenta i rapporti per la malattia "Alzheimer" rispetto alla categoria di riferimento "Other dementias", e la seconda riga rappresenta i rapporti per la malattia "Other diagnoses" rispetto alla categoria di riferimento. Questo rapporto tra la probabilità di avere "Alzheimer" e "Other diagnoses" rispetto alla linea di base "Other dementias" è noto come rischio relativo (spesso descritto come probabilità). Tuttavia, l'output del modello sono le probabilità logaritmiche. Per ottenere l'odds stimate dobbiamo elevare a potenza i coefficienti.

```

exp(coef(alzhe_mod3))

##              (Intercept) genderMale smoking<10 smoking10-20 smoking>20
## Alzheimer              1.820026  0.6344836  1.2447940    0.6311027  0.4377858
## Other diagnoses        1.503670  0.6320281  0.5029329    1.1988665  0.3933007

```

Log-likelihood del modello “alzhe\_mod3”:

Con

$$\pi_j = 1 - (\pi_1 + \dots + \pi_{j-1}) \text{ e } y_{ij} = 1 - (y_{i1} + \dots + y_{ij-1})$$

$$\log[\prod_{j=1}^J \pi_j(x_i)^{y_{ij}}] = \sum_{j=1}^{J-1} y_{ij} \log \pi_j(x_i) + \left(1 - \sum_{j=1}^{J-1} y_{ij}\right) \log \left(1 - \sum_{j=1}^{J-1} \pi_j(x_i)\right)$$

$$\log[\prod_{j=1}^J \pi_j(x_i)^{y_{ij}}] = \sum_{j=1}^{J-1} y_{ij} \log \pi_j(x_i) + y_{ij} \log \left(1 - \sum_{j=1}^{J-1} \pi_j(x_i)\right) \quad (1.4)$$

$$\begin{aligned} &= 55 * \log(0.2312835) + 80 * \log(0.3477741) + (91) * \log(1 - 0.2312835 - 0.3477741) \\ &\dots + 35 * \log(0.5321067) + 11 * \log(0.1988898) + (6) * \log(1 - 0.5321067 - 0.1988898) \\ &= -573.1897572 \end{aligned}$$

Smok	Gender	Alzheimer (Y <sub>ij</sub> )	Dementias (Y <sub>ij</sub> )	Diagnosis (Y <sub>ij</sub> )	nj	p_hat (Alzheimer)	p_hat (Dementias)	p_hat (Diagnoses)	Log-likelihood
none	F	91	55	80	226	0.4209	0.2313	0.3478	-243.76
<10	F	7	7	3	17	0.5633	0.2486	0.1880	-18.77
10-20	F	15	16	25	56	0.2907	0.2531	0.4562	-60.14
>20	F	21	9	9	39	0.3336	0.4187	0.2476	-43.45
none	M	35	24	24	83	0.3719	0.3220	0.3061	-90.23
<10	M	8	1	2	11	0.4931	0.3430	0.1639	-10.34
10-20	M	15	17	22	54	0.2541	0.3487	0.3972	-58.77
>20	M	6	35	11	52	0.2690	0.5321	0.1989	-47.73
<b>Log-likelihood (alzhe_mod3):</b>									-573.1898
<b>Residual Deviance -2(L):</b>									1146.38

Il criterio di informazione Akaike è altri criteri oltre ai test di significatività che possono aiutare a selezionare un buon modello in termini di stima. Valuta un modello in base a quanto i suoi valori adattati tendono ad essere vicini ai valori reali, in termini di un determinato valore atteso. Pertanto, il modello ottimale è quello che tende ad adattarsi maggiormente alla realtà. Akaike ha mostrato che questo criterio seleziona il modello che minimizza:

$$AIC = -2(\text{maximized log likelihood} - \text{number of parameters in model})$$

$$AIC(\text{alzhe\_mod3}) = -2(-573.1898 - 10) = 1166.38$$

Riassumendo gli output informativi dei modelli precedentemente descritti. In questo caso il parametro per scegliere il modello migliore è la Devianza e che corrisponde al valore più piccolo di AIC. La minima distanza tra il modello da scegliere e il modello saturato è quella che presenta il miglior adattamento ai dati con cui è stato costruito. Si risolve l'equazione (1.3), per trovare la differenza. Dati i risultati, viene scelto il modello D~G+S, che ha il miglior adattamento tra tutti i modelli, con gradi di libertà  $df=2[n-(1+\#parameter)]=2[538-(1+4)]=1066$ .

<b>Model Predictor</b>	<b>Log- likelihoo d</b>	<b>Residual Deviance</b>	<b>AIC</b>	<b>df</b>	<b>#parameters</b>	<b>Models Compare d</b>
D~G*SM <sub>S</sub>	-561.225	1122.449	1154.449	1060	16(2 intercept)	-2[L <sub>l</sub> - L <sub>s</sub> ]
D~G	-584.618	1169.236	1177.236	1072	4(2 intercept)	
M <sub>gender</sub> vSM <sub>S</sub>						46.78622 con df=12
D~S	-575.783	1151.566	1167.566	1068	8(2 intercept)	
M <sub>smok</sub> vSM <sub>S</sub>						29.11655 con df=8
D~G+S	-573.189	1146.38	1166.38	1066	10(2 intercept)	
M <sub>GenderSmok</sub> v <sub>S</sub>						23.93004 con df=6
D~1(Null)	-589.408	1178.815	1182.815	1074	2(2 intercept)	
M <sub>N</sub> vSM <sub>S</sub>						56.36603 con df=14

Altro codice per le stime dei parametri logistici sono ottenute con la funzione `vglm()` del pacchetto VGAM. Questo output contiene più informazioni di quelle del pacchetto `nmet()`. Contiene, come l'altro output, le stime dei coefficienti e il loro errore standard corrispondente. Contiene anche il z-value che si ottiene dividendo il coefficiente stimato per il suo errore standard e il p-value della stima.

```
alzhe<-vglm(disease~gender+smoking, multinomial(refLevel = 1),data=data_alzhe)
summary(alzhe)
```

```
##
## Call:
## vglm(formula = disease ~ gender + smoking, family = multinomial(refLevel =
1),
##      data = data_alzhe)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept):1    0.5989    0.1591   3.764 0.000167 ***
## (Intercept):2    0.4079    0.1647   2.476 0.013283 *
## genderMale:1    -0.4549    0.2272  -2.003 0.045202 *
## genderMale:2    -0.4588    0.2338  -1.962 0.049749 *
## smoking<10:1     0.2190    0.4638   0.472 0.636873
## smoking<10:2    -0.6873    0.5916  -1.162 0.245366
## smoking10-20:1  -0.4603    0.2954  -1.558 0.119252
## smoking10-20:2   0.1814    0.2781   0.652 0.514261
## smoking>20:1    -0.8260    0.2912  -2.837 0.004556 **
## smoking>20:2    -0.9332    0.3156  -2.957 0.003110 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1])
##
## Residual deviance: 1146.38 on 1066 degrees of freedom
##
## Log-likelihood: -573.1898 on 1066 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Reference group is level 1 of the response
```

Interpretando i risultati di questo modello stimato, “genderMale:1”, “genderMale:2”, “smoking>20:1” e “smoking>20:2” sono statisticamente significativi, anche “(Intercept):1” e “(Intercept):2” con p-value minore di 0.05. Tutti gli altri non sono statisticamente significativi, i loro p-values sono maggiori di 0,05. Nel caso di “genderMale:1” nella prima equazione, il p-value è 0,045202, essendo molto vicino a 0,05 e il suo significato nello spiegare la variabile dipendente cade in dubbio. Esponenziando il suo coefficiente stimato, l’odds stimate per una persona di sesso maschile sviluppi “Alzheimer” invece de “Other dementias” sono 0.63 (diminuisce in 37%) le probabilità stimate per il sesso femminile:

$$\frac{\pi_{alzheimer}(Gender = M)}{\pi_{dementias}(Gender = M)} = e^{(-0.454944204)} \frac{\pi_{alzheimer}(Gender = F)}{\pi_{dementias}(Gender = F)}$$

$$= 0.63 \frac{\pi_{alzheimer}(Gender = F)}{\pi_{dementias}(Gender = F)}$$

Dobbiamo calcolare il suo intervallo di confidenza per verificare se è davvero statisticamente significativo.

Intervallo de confidenza a 95% per i parametri stimati (genderMale:1):

$$\{\hat{\beta}_j - Z_{\alpha/2} * SE(\hat{\beta}_j); \hat{\beta}_j + Z_{\alpha/2} * SE(\hat{\beta}_j)\} \quad (1.5)$$

$$\{-0.454944 - 1.96(0.227157); -0.454944 + 1.96(0.227157)\}$$

$$\{e^{-0.90017172}; e^{-0.00971628}\}$$

$$\{0.41; 0.99\}$$

L'intervallo stimato è molto vicino a 1 e questo significa che il fattore "Gender" ha alcuna influenza sullo sviluppo della malattia, ma piccola, in alcuni casi la variabile verrebbe rimossa. Tuttavia, non dovremmo eliminare la variabile "Gender" dal modello perché stiamo rimuovendo le informazioni dal modello e questo viene verificato quando sono stati effettuati i confronti con il modello Saturo. C'è una differenza tra loro che è la minima di tutti, quindi non dovrebbe essere rimossa dal modello, perché il modello perde informazioni. Prima di fornire una migliore interpretazione del modello e calcolare le probabilità stimate per ciascuna delle categorie, è necessario verificare se il modello è veramente conforme ai dati. Ci si aspetta che questo test non respinga  $H_0$  (alto p-value), cioè che i dati ottenuti non siano contro il modello logistico. In altre parole, passando dal modello saturo al modello logistico, non si perdono informazioni statisticamente significative. Calcoliamo i gradi di libertà  $N(J-1)$  - #parametri, dove  $N$  è il numero di gruppi  $N$  gruppi  $\cdot (J-1) = 2(gender) \cdot 4(smoking) \cdot 2(logit)$  - # parametri per `alzhe_mod3` sarà  $16 - 10 = 6$  p-value del chi square

$$p - value = 1 - P(\chi_6^2 > 23.93004) = 0.0005379607$$

```
1-pchisq(deviance(alzhe_mod3)-deviance(alzhe_mod5), 6) # non si accetta il mo
dello alzhe_mod3
## [1] 0.0005379607
```

Dato il risultato del test precedente, il modello logistico non viene accettato, rifiutando l'ipotesi nulla con un p-value di 0.0005379606 minore de 0.05.

**UN'ALTRA SOLUZIONE:** trasformato il fattore "Smoking".

Dopo aver esplorato i dati e rifiutato il modello logistico, ci resta l'alternativa di modificare i dati iniziali. In questo caso modificheremo il fattore "Smoking" da quattro a due categorie. La risposta in questo caso sarebbe "Yes" o "No". Tutte le categorie originali ">10", "10-20" e ">20" verrebbero riassunte in un'unica categoria. Vengono mostrate le 6 prime e l'ultime osservazione, dopo aver effettuato la trasformazione.

```
smok<-ifelse(data_alzhe$smoking=="None","No","Yes")

alzheM<-cbind(data_alzhe, smok)
head(alzheM) #Visualizzazione delle prime 6 osservazioni

##   smoking   disease gender smok
## 1     None Alzheimer Female   No
## 2     None Alzheimer Female   No
## 3     None Alzheimer Female   No
## 4     None Alzheimer Female   No
## 5     None Alzheimer Female   No
## 6     None Alzheimer Female   No

tail(alzheM) #Visualizzazione delle ultime 6 osservazioni

##   smoking   disease gender smok
## 533    >20 Other diagnoses   Male  Yes
## 534    >20 Other diagnoses   Male  Yes
## 535    >20 Other diagnoses   Male  Yes
## 536    >20 Other diagnoses   Male  Yes
## 537    >20 Other diagnoses   Male  Yes
## 538    >20 Other diagnoses   Male  Yes
```

## Tabella delle frequenze

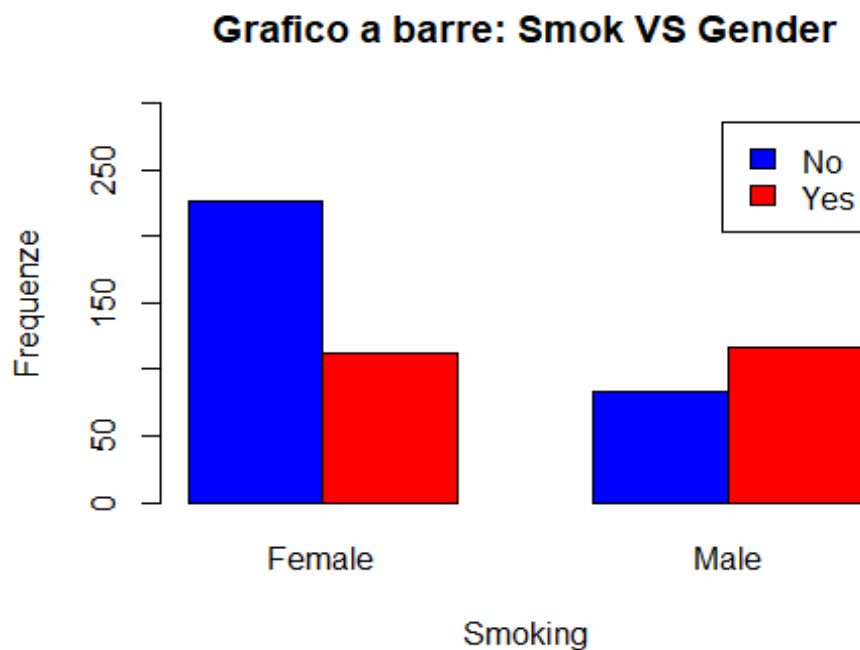
Questo output mostra la tabella delle frequenze, con le frequenze distribuite per ciascuna categoria di ciascun fattore, ma con il fattore "Smoking" trasformato nelle due categorie. Circa del 33,1% di tutti i pazienti di genere femminile ha il vizio del fumo, mentre il 58,5% di tutti i pazienti di sesso maschile ha il vizio del fumo.

```
table(alzheM$smok, alzheM$disease, alzheM$gender)

## , , = Female
##
##
##      Other dementias Alzheimer Other diagnoses
## No                55        91             80
## Yes               32        43             37
##
## , , = Male
##
##
##      Other dementias Alzheimer Other diagnoses
## No                24        35             24
## Yes               53        29             35
```

```
tab_smok2 <- table(alzheM$smok, alzheM$gender)

barplot(tab_smok2,
        main="Grafico a barre: Disease VS Smok",
        xlab="Smoking",
        ylab="Frequenze",
        legend = rownames(tab_smok2),
        ylim = c(0, 300),
        col=c("blue", "red", "cyan"),
        beside=TRUE)
```





## Costruzione del modello

Tutti i passi precedenti vengono rifatti per la costruzione del nuovo modello. Nuovi modelli vengono creati per essere confrontati con il modello saturo.

```
attach(alzheM)
alzheM$disease <- relevel(alzheM$disease, ref = "Other dementias")
options(contrasts=c("contr.treatment", "contr.poly"))
modS<-multinom(disease~gender*smok, data=alzheM) #Modello Saturo

## # weights:  15 (8 variable)
## initial  value 591.053411
## iter   10 value 580.298333
## final   value 580.260118
## converged

mod0<-multinom(disease~1,data=alzheM) #Modelo null

## # weights:  6 (2 variable)
## initial  value 591.053411
## final   value 589.407756
## converged

mod1<-multinom(disease~gender, data=alzheM) #Modello Gender

## # weights:  9 (4 variable)
## initial  value 591.053411
## final   value 584.617852
## converged

mod2<-multinom(disease~smok, data=alzheM) #Modello Smoking

## # weights:  9 (4 variable)
## initial  value 591.053411
## final   value 584.884961
## converged

mod3<-multinom(disease~gender+smok, data=alzheM) #Modello Gender + Smoking

## # weights:  12 (6 variable)
## initial  value 591.053411
## iter   10 value 581.861183
## final   value 581.861035
## converged
```

I diversi modelli logit vengono aggiustati con la categoria di base “Other dementias” e confrontati in termini delle loro differenze di devianza.

```
deviance(mod1)-deviance(modS)

## [1] 8.715468
```

```

deviance(mod2)-deviance(modS)

## [1] 9.249687

deviance(mod3)-deviance(modS)

## [1] 3.201834

deviance(mod0)-deviance(modS)

## [1] 18.29528

summary(mod3)

## Call:
## multinom(formula = disease ~ gender + smok, data = alzheM)
##
## Coefficients:
##              (Intercept) genderMale      smokYes
## Alzheimer              0.6098540 -0.4914707 -0.5172358
## Other diagnoses        0.4158041 -0.4825830 -0.3273280
##
## Std. Errors:
##              (Intercept) genderMale      smokYes
## Alzheimer              0.1590710  0.2252614  0.2219751
## Other diagnoses        0.1646062  0.2303975  0.2261444
##
## Residual Deviance: 1163.722
## AIC: 1175.722

```

Si risolve l'equazione (1.3), per trovare la differenza. Dati i risultati, viene scelto il modello  $D \sim G + S$ , che ha il miglior adattamento tra tutti i modelli, con gradi di libertà  $df = 2[n - (1 + \#parameter)] = 2[538 - (1 + 2)] = 1070$ .

Model Predictors	Log-likelihood	Residual Deviance	AIC	df	#parameters	Models Compared
$D \sim G^*S$ $M_S$	-580.2601	1160.52	1176.52	1068	8(2 intercept)	$-2[L_I - L_S]$
$D \sim G$ $M_{gender} \nu S M_S$	-584.6179	1169.236	1177.236	1072	4(2 intercept)	8.715468 con $df=4$
$D \sim S$ $M_{smok} \nu S M_S$	-584.885	1169.77	1177.77	1072	4(2 intercept)	9.249687 con $df=4$
$D \sim G + S$ $M_{GenderSmok}^1$	-581.861	1163.722	1175.722	1070	6(2 intercept)	3.201834 con $df=2$

Model Predictors	Log-likelihood	Residual Deviance	AIC	df	#parameters	Models Compared
D~1	-589.4078	1178.815	1182.816	1074	2(2intercept)	
$M_N vs M_S$						18.29528 con df=6

```
exp(coef(mod3))
```

```
##              (Intercept) genderMale  smokYes
## Alzheimer      1.840163   0.6117261 0.5961662
## Other diagnoses 1.515589   0.6171872 0.7208472
```

Questo output mostrato in precedenza, sono i rapporti di rischio relativo per il nuovo modello logit multinomiale stimato. Possono essere ottenuti esponenziando i coefficienti logit multinomiali con la funzione `exp(coef(alzhe_mod3))`. Il Rischio relativo di un coefficiente indica come cambia il rischio che l'esito ricada nel gruppo di confronto rispetto al rischio che l'esito ricada nel gruppo referente con la variabile in questione. Un rischio relativo (RR) > 1 indica che il rischio che l'esito ricada nel gruppo di confronto rispetto al rischio che l'esito ricada nel gruppo di riferimento aumenta all'aumentare della variabile. Un RR < 1 indica che il rischio che l'esito ricada nel gruppo di confronto rispetto al rischio che l'esito ricada nel gruppo di riferimento diminuisce all'aumentare della variabile. In generale, se RR < 1, è più probabile che il risultato appartenga al gruppo di riferimento.

```
alzhe<-vglm(disease~gender+smok, multinomial(refLevel = 1),data=alzheM)
summary(alzhe)

##
## Call:
## vglm(formula = disease ~ gender + smok, family = multinomial(refLevel = 1)
##      data = alzheM)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   0.6099    0.1591   3.834 0.000126 ***
## (Intercept):2   0.4158    0.1646   2.526 0.011535 *
## genderMale:1   -0.4915    0.2253  -2.182 0.029126 *
## genderMale:2   -0.4826    0.2304  -2.095 0.036210 *
## smokYes:1      -0.5172    0.2220  -2.330 0.019798 *
## smokYes:2      -0.3273    0.2261  -1.447 0.147780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1])
##
## Residual deviance: 1163.722 on 1070 degrees of freedom
##
## Log-likelihood: -581.861 on 1070 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Reference group is level 1 of the response
```

Riassumendo il calcolo del log-likelihood del nuovo modello “mod3”:

Smok	Gender	Dementia ( $Y_{ij}$ )	Alzheimer ( $Y_{ij}$ )	Diagnoses ( $Y_{ij}$ )	nj	p_hat (Dementia)	p_hat (Alzheimer)	p_hat (Diagnoses)	Log-likelihood
Yes	F	32	43	37	112	0.3135	0.3439	0.3425	-122.6504
Yes	M	53	29	35	117	0.4264	0.2861	0.2875	-125.0964
No	F	55	91	80	226	0.2296	0.4225	0.3480	-243.7973
No	M	24	35	24	83	0.3267	0.3677	0.3056	-90.3169
						Log-likelihood (mod3):			-581.861
						Residual Deviance		-2(L):	1163.722

Interpretando i risultati di questo nuovo modello stimato, “genderMale:1”, “genderMale:2” e “smokYes:1” sono statisticamente significativi, anche “(Intercept):1 e”(Intercept):2” con p-value minore di 0.05. “SmokYes:2” non sono statisticamente significativi, il p-value è maggiore di 0,05.

```
z1<-0.6098540/0.1590710
z2<-0.4158041/0.1646062
z3<-(-0.4914707/0.2252614)
z4<-(-0.4825830/0.2303975)
z5<-(-0.5172358/0.2219751)
z6<-(-0.3273280/0.2261444)

pval1<-(1-pnorm(abs(z1),0,1))*2
pval2<-(1-pnorm(abs(z2),0,1))*2
pval3<-(1-pnorm(abs(z3),0,1))*2
pval4<-(1-pnorm(abs(z4),0,1))*2
pval5<-(1-pnorm(abs(z5),0,1))*2
pval6<-(1-pnorm(abs(z6),0,1))*2

rbind(pval1, pval2, pval3, pval4, pval5, pval6)
```

```
##           [,1]
## pval1 0.0001261542
## pval2 0.0115351849
## pval3 0.0291258118
## pval4 0.0362095015
## pval5 0.0197981006
## pval6 0.1477767623
```

- Per l'Alzheimer relativo ad Altre demenze, la statistica del test z per il predittore genderMale  $(-0,4915/0,2253)$  è -2,182 con un p\_value associato di 0,029126. Con un livello alfa a 0,05 concludiamo che la differenza tra maschi e femmine è risultata statisticamente significativa per l'Alzheimer rispetto ad altre demenze dato che il fumo è nel modello.
- Per l'Alzheimer relativo ad Altre demenze, la statistica del test z per il predittore smokYes  $(-0,5172/0,2220)$  è -2,330 con un p\_value associato di 0,019789. Con un livello alfa a 0,05 concludiamo che la differenza tra fumatori e non fumatori è risultata statisticamente significativa per l'Alzheimer rispetto ad altre demenze dato che il gender è nel modello.
- Per Altre diagnoses relativo ad Altre demenze, la statistica del test z per il predittore genderMale  $(-0,4826/0,2304)$  è -2,095 con un p\_value associato di 0,036210. Con un livello alfa a 0,05 concludiamo che la differenza tra maschi e femmine è risultata statisticamente significativa per altre diagnoses rispetto ad altre demenze dato che il fumo è nel modello.
- Per Altre diagnoses relativo ad Altre demenze, la statistica del test z per il predittore smokYes  $(-0,3273/0,2261)$  è -1,447 con un p\_value associato di 0,147780. Con un livello alfa a 0,05 concludiamo che la differenza tra fumatori e non fumatori non è risultata statisticamente significativa per altre diagnoses rispetto ad altre demenze dato che il gender è nel modello.

### **Intervali de confidenza a 95% per i parametri stimati**

	Parame tri Stimati	Std. Error	Lower $\widehat{\beta}_j - Z_{\alpha/2} *$ $SE(\widehat{\beta}_j)$	Upper $\widehat{\beta}_j +$ $Z_{\alpha/2} *$ $SE(\widehat{\beta}_j)$	$e^{\widehat{\beta}_j}$	$e^L$	$e^U$
(Intercept):1	0.6099	0.1591	0.298064	0.921736	1.8402	1.3472	2.5137
(Intercept):2	0.4158	0.1646	0.093184	0.738416	1.5156	1.0977	2.0926
genderMale:1	-0.4915	0.2253	-0.933088	-0.049912	0.6117	0.3933	0.9513
genderMale:2	-0.4826	0.2304	-0.934184	-0.031016	0.6172	0.3929	0.9695
smokYes:1	-0.5172	0.2220	-0.95232	-0.08208	0.5967	0.3858	0.9212
smokYes:1	-0.3273	0.2261	-0.770456	0.115856	0.7208	0.4628	1.1228

### ***“Alzheimer correlate to Other dementias”***

genderMale:1- Questo è il rapporto di rischio relativo confrontando maschi e femmine affetti da “Alzheimer” rispetto ad “Other dementias”, dato che le altre variabili nel modello sono mantenute costanti. Per i maschi rispetto alle femmine, il rischio relativo di soffrire di “Alzheimer” rispetto a “Other dementias” dovrebbe diminuire di un fattore di 0.61 (disminuisce in 39%) dato che le altre variabili nel modello sono mantenute costanti. *I maschi hanno minore probabilità rispetto alle femmine di soffrire di “Alzheimer” rispetto a “Other dementias”.*

smokYes:1- Questo è il rapporto di rischio relativo confrontando fumatori e non fumatori affetti da “Alzheimer” rispetto a “Other dementias”, dato che le altre variabili nel modello sono mantenute costanti. Per i fumatori rispetto alle non fumatori, il rischio relativo di soffrire di “Alzheimer” rispetto a “Other dementias” dovrebbe diminuire di un fattore di 0.60 (disminuisce in 40%) dato che le altre variabili nel modello sono mantenute costanti. *I fumatori hanno minore probabilità rispetto alle non fumatori di soffrire di “Alzheimer” rispetto a “Other dementias”.*

### ***“Other diagnoses correlate to Other dementias”***

genderMale:2- Questo è il rapporto di rischio relativo confrontando maschi e femmine affetti da “Other diagnoses” rispetto a “Other dementias”, dato che le altre variabili nel modello sono mantenute costanti. Per i maschi rispetto alle femmine, il rischio relativo di soffrire di “Other diagnoses” rispetto a “Other dementias” dovrebbe diminuire di un fattore di 0.62 (disminuisce in 38%) dato

che le altre variabili nel modello sono mantenute costanti. *I maschi hanno minore probabilità rispetto alle femmine di soffrire di “Other diagnoses” rispetto a “Other dementias”.*

smokYes:2- Questo è il rapporto di rischio relativo confrontando fumatori e non fumatori affetti da “Other diagnoses” rispetto a “Other dementias”, dato che le altre variabili nel modello sono mantenute costanti. Per i fumatori rispetto alle non fumatori, il rischio relativo di soffrire di “Other diagnoses” rispetto a “Other dementias” dovrebbe diminuire di un fattore di 0.72 (diminuisce in 28%) dato che le altre variabili nel modello sono mantenute costanti. *I fumatori hanno minore probabilità rispetto alle non fumatori di soffrire di “Other diagnoses” rispetto a “Other dementias”.*

Affinché le interpretazioni precedenti siano valide, il test chi-quadrato viene eseguito nuovamente per vedere se il modello è accettato. Ci si aspetta che questo test non respinga  $H_0$  (high p-value).

$H_0$ : Modello Logistico

$H_1$ : Modello Saturo

Calcoliamo i gradi di libertà  $N(J-1) - \# \text{parametri}$ , dove  $N$  è il numero di gruppi  $N = \text{gruppi} * (J-1) = 2(\text{gender}) * 2(\text{smoking}) * 2(\text{logit}) - \# \text{parametri per alzhe\_mod3}$  sarà  $8 - 6 = 2$  p-value del chi square

$$p - \text{value} = 1 - P(\chi^2_2 > 3.201834) = 0.2017115$$

```
1-pchisq(deviance(mod3)-deviance(mod5), 2) # si accetta il modello alzhe_mod
## [1] 0.2017115
```

Dato il risultato del test precedente, il modello logistico viene accettato, con un p-value di 0.2017115, maggiore di 0.05.

## Predizioni

Sostituendo i valori nella formula 1.2, si ottengono le probabilità stimate per questo modello di regressione.

<i>G</i>	<i>Smok</i>	<i>Disease</i>	<i>P. Stimata</i>
----------	-------------	----------------	-------------------

F	Yes	Alzheimer	$\frac{e^{0.6099-0.5172}}{1 + e^{0.6099-0.5172} + e^{0.4158-0.3273}} = 0.3439490$
		Other dementias	$\frac{1}{1 + e^{0.6099-0.5172} + e^{0.4158-0.3273}} = 0.3135238$
		Other diagnoses	$\frac{e^{0.4158-0.3273}}{1 + e^{0.6099-0.5172} + e^{0.4158-0.3273}} = 0.3425273$
M	Yes	Alzheimer	$\frac{e^{0.6099-0.5172-0.4915}}{1 + e^{0.6099-0.5172-0.4915} + e^{0.4158-0.3273-0.4826}} = 0.2861336$
		Other dementias	$\frac{1}{1 + e^{0.6099-0.5172-0.4915} + e^{0.4158-0.3273-0.4826}} = 0.4263717$
		Other diagnoses	$\frac{e^{0.4158-0.3273-0.4826}}{1 + e^{0.6099-0.5172-0.4915} + e^{0.4158-0.3273-0.4826}} = 0.2874947$
F	No	Alzheimer	$\frac{e^{0.6099}}{1 + e^{0.6099} + e^{0.4158}} = 0.4224673$
		Other dementias	$\frac{1}{1 + e^{0.6099} + e^{0.4158}} = 0.2295815$
		Other diagnoses	$\frac{e^{0.4158}}{1 + e^{0.6099} + e^{0.4158}} = 0.3479512$
M	No	Alzheimer	$\frac{e^{0.6099-0.4915}}{1 + e^{0.6099-0.4915} + e^{0.4158-0.4826}} = 0.3677383$
		Other dementias	$\frac{1}{1 + e^{0.6099-0.4915} + e^{0.4158-0.4826}} = 0.3266824$
		Other diagnoses	$\frac{e^{0.4158-0.4826}}{1 + e^{0.6099-0.4915} + e^{0.4158-0.4826}} = 0.3055793$

```

smoking.labs<-factor(c("Yes", "No"),
levels=c("Yes", "No"))
gender.labs<-factor(c("Female", "Male"), levels=c("Female", "Male"))
disease.labs<-factor(c("Alzheimer", "Other dementias", "Other diagnoses"),
levels=c("Alzheimer", "Other dementias", "Other diagnoses"))

predictions1<-predict(mod3, type = "probs",
newdata = expand.grid(gender=gender.labs, smok=smoking.labs))
cbind(expand.grid(gender = gender.labs, smok = smoking.labs), predictions1)

##  gender smok Other dementias Alzheimer Other diagnoses
## 1 Female  Yes      0.3135238 0.3439490      0.3425273
## 2 Male   Yes      0.4263717 0.2861336      0.2874947
## 3 Female No       0.2295815 0.4224673      0.3479512
## 4 Male  No       0.3266824 0.3677383      0.3055793

```

L'output precedente mostra le probabilità stimate dal modello per ciascuna delle categorie della variabile dipendente e delle covariate. Le probabilità sono



distribuite quasi uniformemente, non ci sono piccole probabilità nei risultati che in qualche modo aiutano a distinguere una malattia da un'altra in modo inequivocabile.

- La probabilità stimata che una donna, con il vizio del fumo di soffrire di Alzheimer sono il 34.4%, per "Other dementias" sono il 31.4% e per "Other diagnoses" sono il 34.3%.
- La probabilità stimata che un uomo, con il vizio del fumo di soffrire di Alzheimer sono il 28.6%, per "Other dementias" sono il 42.6% e per "Other diagnoses" sono il 28.7%.
- La probabilità stimata che una donna, senza il vizio del fumo di soffrire di Alzheimer sono il 42.2%, per "Other dementias" sono il 22.9% e per "Other diagnoses" sono il 34.8%.
- La probabilità stimata che un uomo, senza il vizio del fumo di soffrire di Alzheimer sono il 36.8%, per "Other dementias" sono il 32.7% e per "Other diagnoses" sono il 30.5%.

In questo caso, la tabella di classificazione corretta verrebbe organizzata in base alle variabili dipendenti. L'accuratezza del modello è del 41,26%, che è la percentuale di previsioni corrette rispetto al totale. Pertanto, è il quoziente tra i casi ben classificati dal modello e la somma di tutte le frequenze. Il modello non presenta una buona percentuale di classificazioni corrette, ma non è trascurabile, potendo effettuare più ricerche di altre variabili che forniscono informazioni al modello. Un esempio di un'altra variabile che potrebbe contribuire al modello è l'età, poiché le demenze e l'Alzheimer sono malattie neurodegenerative con l'età, mentre una persona ha più anni, più si verificano questi tipi di malattie. In questo caso, non c'erano più variabili che potessero contribuire al modello.

```
predict_modalzhe<-predict(mod3, alzhem, type="class")
head(predict_modalzhe)

## [1] Alzheimer Alzheimer Alzheimer Alzheimer Alzheimer Alzheimer
## Levels: Other dementias Alzheimer Other diagnoses

t<-table(predict_modalzhe, disease)
```

```

t
##              disease
## predict_modalzhe Other dementias Alzheimer Other diagnoses
##   Other dementias          53          29          35
##   Alzheimer            111          169          141
##   Other diagnoses           0           0           0

sum(diag(t))/sum(t)
## [1] 0.4126394

```

## Tabella di frequenza teoriche

La tabella di frequenza teorica è costruita per ogni segmentazione, sulla base della frequenza osservata e delle probabilità stimate dal modello. Diventa un vettore la tabella collassata delle frequenze tra Fumo e Genere che è rappresentata dalla seguente per `freque.osse = [226 112 83 117]` ed è moltiplicata per le probabilità segmentate, stimate dal modello.

```

freque.osse <- tapply(alzheM$disease, list(factor(alzheM$smok, levels = c("No", "Yes")),
factor(alzheM$gender, levels = c("Female", "Male"))), length)
freque.osse

##      Female Male
## No      226   83
## Yes     112  117

# Le freq teoriche partendo dai conteggi in freque.osse e dalle prob previste
(predict)
row.names2 <- rev(expand.grid(dimnames(freque.osse)))
freque.teoriche<-round(as.vector(freque.osse)* fitted(mod3)[!duplicated(as.data.frame(fitted(mod3))),
],1)
structure(.Data = as.data.frame(freque.teoriche), row.names = apply(row.names2, 1,
paste, collapse = " "))

##              Other dementias Alzheimer Other diagnoses
## Female No          51.9          95.5          78.6
## Female Yes         35.1          38.5          38.4
## Male No            27.1          30.5          25.4
## Male Yes           49.9          33.5          33.6

```

<b>Segmento</b>	<b><math>n_{ij}</math></b>	<b>Other dementias</b>	<b>Alzheimer</b>	<b>Other diagnoses</b>
<b>Female No</b>	226	$226 * 0.2296 = 51.9$	$226 * 0.4225 = 95.5$	$226 * 0.3480 = 78.6$
<b>Female Yes</b>	112	$112 * 0.3135 = 35.1$	$112 * 0.3439 = 38.5$	$112 * 0.3425 = 38.4$
<b>Male No</b>	83	$83 * 0.3267 = 27.1$	$83 * 0.3677 = 30.5$	$83 * 0.3056 = 25.4$
<b>Male Yes</b>	117	$117 * 0.4264 = 49.9$	$117 * 0.2861 = 33.5$	$117 * 0.2875 = 33.6$

## CONCLUSIONE

Lo studio precedente è basato sulla revisione e formulazione di tutti i modelli da costruire per studiare la variabile dipendente “Disease” attraverso le variabili indipendenti “Gender” e “Smoking”. È stato stimato un primo modello, nel quale sono state riprese le variabili originarie, senza le loro modifiche per le stime di pertinenza. Il modello scelto era Disease~Genere+Smoking con la minima differenza di Devianza tra il modello saturo di 23,930004. Viene eseguito il test Chi-quadrato, in cui il modello logistico viene rifiutato con un valore p di 0,0005379607. Passando dal modello saturo al modello logistico, si perdono informazioni statisticamente significative. Viene stimato un secondo modello, in cui la variabile indipendente Smoking viene modificata da quattro categorie a due. Tutte le categorie originarie “>10”, “10-20” e “>20” saranno richiamate in due categoria. La categoria di riferimento per le stime delle equazioni J-1 è “Altre demenze”. I risultati sono stati significativi per i fattori “Smoking” e “Gender”, ovvero le variabili indipendenti spiegano la variabile dipendente “Disease”. Il modello scelto era Disease~Genere+Smok con la minima differenza di Devianza tra il modello saturo di 3.201834. Il risultato del test Chi-quadrato è che il modello logistico non viene rifiutato con un valore p di 0,2017115. Per quanto riguarda le probabilità stimate e in relazione all’abitudine al fumo, una donna con l’abitudine al fumo ha una probabilità stimata di soffrire di Alzheimer del 34,4% e di soffrire di altre demenze del 31.1%. Un uomo che fuma ha una probabilità stimata di soffrire di Alzheimer del 28.6% e di soffrire di altre demenze del 42.2%. Smettere di fumare potrebbe aiutare a ridurre il rischio di demenza. Incoraggiare e sostenere i consumatori di tabacco a smettere di fumare dovrebbe essere una priorità nella prevenzione della malattie di Alzheimer e di altre forme di demenza.

## RIFERIMENTI BIBLIOGRAFICI

Salib, E. and Hillier, V. (1997). A case-control study of smoking and Alzheimer's disease. *International Journal of Geriatric Psychiatry* 12(3), 295–300. doi: 10.1002/(SICI)1099-1166(199703)12:3<295::AID-GPS476>3.0.CO;2-3.

Agresti A. (2010). *Categorical Data Analysis*. Second Edition 2002, WILEY SERIES IN PROBABILITY AND STATISTICS

**Note:** tutte le equazioni di questo documento sono state riprese testualmente dal libro “Categorical Data Analysis” di Allan Agresti.