

NMC - Foundations of Statistical Modelling

Lorenzo Baiardi

28 Febbraio 2023

Indice

1	Introduzione	3
2	Visualizzazione del dataset	3
2.1	Variabili	3
2.2	Tabella delle Frequenze	5
3	Regressioni Logistiche Semplici	7
3.1	Age	7
3.2	Sex	10
3.3	BMI	12
3.4	Fitness	14
3.5	PA	18
3.6	Smoke	18
3.7	Alchol	22
3.8	Commento	22
4	Regressioni Logistiche Multiple	23
4.1	Modello Completo	23
4.2	Modello Significativo	24
4.3	Commento	27
4.4	Dati di esempio	28
5	Considerazioni sul Modello	30
5.1	Maschio e Femmina	30
5.2	Attività Fisica	33
6	Interazioni fra le variabili	37
6.1	Smoke e Alchol	37
6.2	Smoke e BMI	38
6.3	Alchol e BMI	39
6.4	Sex e Smoke	40
6.5	Sex e Age	41
6.6	PA e Age	42
6.7	PA e Fitness	43
6.8	Modello con interazioni	44
6.9	Commento	45
7	Selezione del Modello	46
7.1	Backward	46
7.1.1	AIC	46
7.1.2	BIC	47
7.2	Forward	50
7.2.1	AIC	50
7.2.2	BIC	52
7.3	Both	54

7.3.1	AIC	54
7.3.2	BIC	57
7.4	Commento	59
8	Test sul Modello	60
8.1	Age	60
8.2	Sex	60
8.3	Smoke	61
9	Visualizzazione del modello	63
10	Conclusioni	64

1 Introduzione

In questo elaborato andremo a studiare l'effetto delle attività personali di un individuo per la prevenzione di problemi cardiovascolari. Andremo a ipotizzare modelli specifici, differenze che si possono verificare tra le diverse categorie di persone e quanto queste categorie possono influire sulla presenza o meno di un problema cardiovascolare.

2 Visualizzazione del dataset

Per lo studio di questo fenomeno utilizzeremo il dataset fornito: *Sjolander et al.(2009)*

Il dataset fornisce un campione di numerosità: $n = 33327$ osservazioni.

```
load("nmc.RData")
#bmi dicotomizzata
nmc$bmi = as.numeric(nmc$bmi>=30)
str(nmc)

## 'data.frame': 33327 obs. of 8 variables:
## $ sex : chr "Male" "Female" "Male" "Female" ...
## $ age : int 94 93 92 92 91 90 89 89 89 89 ...
## $ bmi : num 0 0 0 0 0 0 0 0 1 0 ...
## $ cvd : int 0 0 0 1 0 0 0 1 0 1 ...
## $ fitness: chr "Just as good" "Much Worse" "A bit better" "Just as good" ...
## $ pa : int 0 1 1 0 0 0 0 0 0 0 ...
## $ smoke : chr "NO" "NO" "Former" "Former" ...
## $ alc : chr "Medium" "Low" "Never" "Never" ...
```

2.1 Variabili

- CVD: variabile d'interesse.
 0. Nessun problema cardiovascolare
 1. Uno o più problemi cardiovascolari
- Sex: rappresenta il genere dell'individuo.
 - Male
 - Female
- Age: età dell'individuo.
- BMI: Body Mass Index, valore dicotomizzato.
 0. $BMI < 30$
 1. $BMI \geq 30$

- Fitness: statico fisico dell'individuo.
 1. Much Worse
 2. Little Worse
 3. Just as good
 4. A bit better
 5. Much better
- PA: Personal Activities.
 0. high-level exerciser
 1. low-level exerciser
- Smoke: tipologia di fumatore.
 - NO
 - Former
 - Current
- Alchol: frequenza dell'uso di alchol dell'individuo.
 1. Never
 2. Low
 3. Medium
 4. High

Per una maggiore comprensione del problema, convertiremo le variabili di tipo categoriale Fitness e Alchol, fornite dal dataset, in variabili di tipo ordinale facilitando la valutazione di quest'ultime durante l'analisi.

Di seguito mostreremo la legenda utilizzata.

```
#LEGENDA:
#Fitness: 1-MUCH WORSE, 2-LITTLE WORSE, 3-JUST AS GOOD,
#         4-A BIT BETTER, 5-MUCH BETTER
#Alchol: 1-NEVER, 2-LOW, 3-MEDIUM, 4-HIGH

c.fit = c('Much Worse', 'Little Worse', 'Just as good',
          'A bit better', 'Much better')
c.alc = c('Never', 'Low', 'Medium', 'High')

#Variabili ordinali
nmc$fitness = as.numeric(ordered(nmc$fitness, c.fit))
nmc$alc = as.numeric(ordered(nmc$alc, c.alc))

str(nmc)
```

```
## 'data.frame': 33327 obs. of 8 variables:
## $ sex : chr "Male" "Female" "Male" "Female" ...
## $ age : int 94 93 92 92 91 90 89 89 89 89 ...
## $ bmi : num 0 0 0 0 0 0 0 0 1 0 ...
## $ cvd : int 0 0 0 1 0 0 0 1 0 1 ...
## $ fitness: num 3 1 4 3 4 4 4 4 4 4 ...
## $ pa : int 0 1 1 0 0 0 0 0 0 0 ...
## $ smoke : chr "NO" "NO" "Former" "Former" ...
## $ alc : num 3 2 1 1 3 2 1 3 3 1 ...
```

2.2 Tabella delle Frequenze

Visualizziamo la frequenza delle varie categorie all'interno del dataset.

```
#Tabella
ftable(sex+bmi+pa ~ smoke+alc+fitness, nmc)
```

			sex Female			Male				
			bmi							
			pa	0	1	0	1	0	1	
smoke	alc	fitness								
Current	1	1	4	1	1	0	2	1	0	1
		2	11	6	6	0	3	2	2	0
		3	31	1	3	0	8	2	1	0
		4	6	1	0	0	3	0	0	0
		5	5	0	0	0	4	0	0	0
	2	1	26	8	12	3	6	3	4	2
		2	166	44	25	9	26	6	2	1
		3	449	37	21	1	94	10	5	1
		4	193	4	8	1	52	1	1	0
		5	55	0	2	0	18	1	0	0
	3	1	11	8	6	2	10	6	2	1
		2	74	21	16	3	49	15	2	2
		3	287	32	14	3	156	24	11	0
		4	200	8	5	0	129	4	3	0
		5	49	0	2	0	32	0	0	0
	4	1	0	0	0	2	0	2	0	0
		2	4	1	0	0	5	0	1	0
		3	12	3	1	0	17	0	1	0
		4	9	1	1	0	7	0	1	0
		5	8	0	0	0	6	1	0	0
Former	1	2	4	6	3	2	2	0	1	
	2	16	5	10	2	7	4	0	2	
	3	83	14	17	1	42	1	6	0	
	4	63	0	8	0	32	5	3	0	

##		5	35	0	0	0	15	0	0	0
##	2	1	35	18	17	12	9	8	8	4
##		2	185	41	70	12	39	16	12	3
##		3	1005	70	124	8	279	40	27	5
##		4	799	24	33	0	348	14	16	0
##		5	290	1	3	1	140	0	4	1
##	3	1	11	7	9	8	9	5	6	5
##		2	132	26	27	8	94	34	30	14
##		3	846	81	57	6	563	57	47	11
##		4	829	15	22	2	740	24	34	2
##		5	281	2	3	0	308	2	6	0
##	4	1	3	0	0	0	2	1	0	0
##		2	4	0	0	0	3	2	0	0
##		3	39	2	3	0	45	7	6	2
##		4	28	2	1	0	60	2	5	0
##		5	14	0	0	0	24	0	0	0
## NO	1	1	30	13	21	10	10	3	1	1
##		2	209	38	49	10	68	22	10	1
##		3	1024	83	111	9	272	30	15	1
##		4	704	19	44	0	386	21	6	1
##		5	255	1	9	0	194	4	2	0
##	2	1	81	28	38	13	29	8	9	6
##		2	624	132	139	20	192	55	26	14
##		3	3169	258	246	18	805	92	48	4
##		4	2549	49	80	3	1059	36	19	1
##		5	876	6	21	0	628	12	3	1
##	3	1	24	9	14	4	21	10	7	3
##		2	212	42	41	9	124	48	23	5
##		3	1288	109	82	11	715	101	40	9
##		4	1326	29	36	2	1093	35	29	0
##		5	441	4	4	0	577	6	6	0
##	4	1	3	0	0	0	3	2	0	0
##		2	1	2	2	0	9	5	4	0
##		3	35	5	2	0	45	5	5	0
##		4	48	0	1	0	53	6	4	0
##		5	14	0	0	0	47	0	1	0

3 Regressioni Logistiche Semplici

Dato che stiamo analizzando un problema che presenta come variabile di risposta una variabile binaria (CVD), utilizzeremo la regressione logistica, implementata in R tramite la funzione `glm()`.

Per prima cosa analizzeremo le regressioni logistiche semplici delle singole variabili presenti nel dataset, visualizzandone il loro comportamento verso la nostra variabile di risposta.

3.1 Age

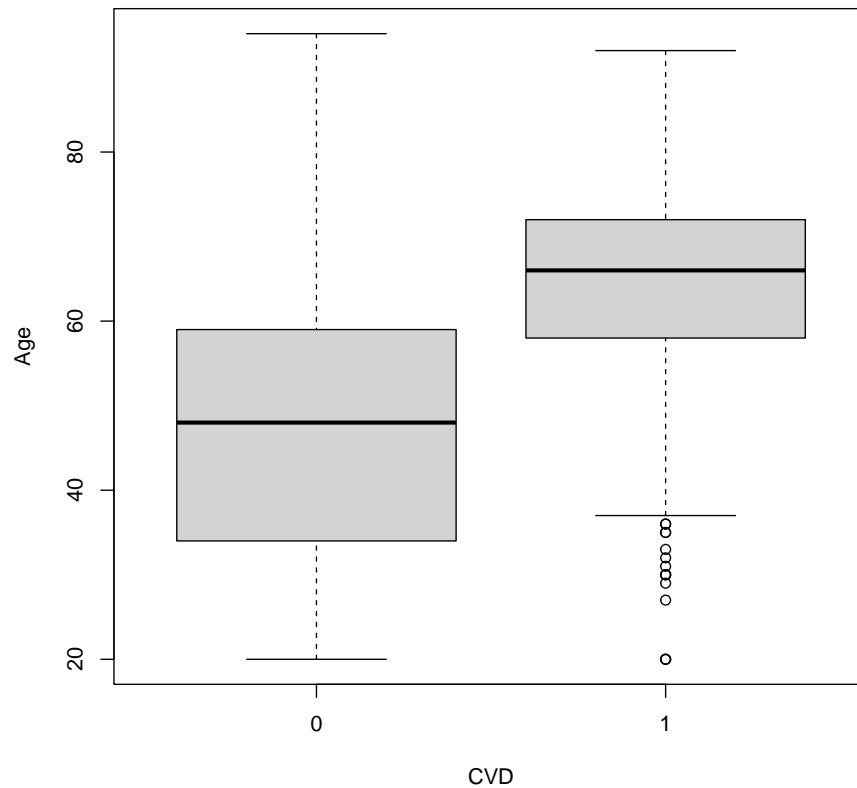
```
#Age
fit.age <- glm(nmc$cvd ~ nmc$age, family=binomial)
summary(fit.age)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$age, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3868  -0.3530  -0.2052  -0.0986   3.5606
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.179700   0.151662  -53.93  <2e-16 ***
## nmc$age      0.092122   0.002345   39.28  <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: 13400  on 33326  degrees of freedom
## Residual deviance: 11150  on 33325  degrees of freedom
## AIC: 11154
##
## Number of Fisher Scoring iterations: 7
```

- L'età influenza positivamente l'insorgenza di un problema cardiovascolare, con valore stimato: $\text{Age} \sim 0.092$.
- La variabile Age è significativa secondo il *p-value*.

Stampiamo ora il boxplot per valutare l'età delle persone che presentano o meno un problema cardiovascolare.

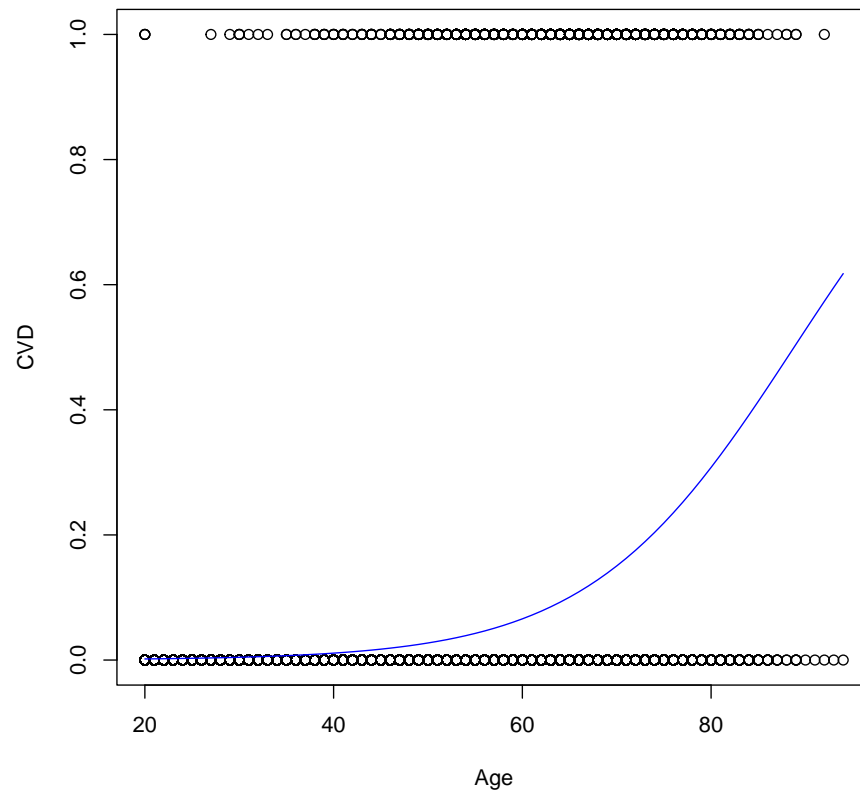

```
#Boxplot  
boxplot(nmc$age~nmc$cvd, xlab="CVD", ylab="Age")
```



- Il boxplot ci mostra come la media delle persone che hanno problemi cardiovascolari, all'interno del dataset, è quella della fascia tra i 60 e 80 anni.
- La media delle persone che non hanno un problema cardiovascolare è quella tra i 40 e 60 anni.
- I problemi cardiovascolari sono più frequenti nella fascia anziana della popolazione.

Eseguiamo il plot del modello con la sola variabile Age.
Modello: $CVD \sim Age$.

```
#Plot
pstim.age <- fit.age$fitted.values
plot(nmc$age, nmc$cvd, xlab="Age", ylab="CVD")
lines(sort(nmc$age), pstim.age[order(nmc$age)], col="blue")
```



Il modello e il grafico suggeriscono come, all'aumentare dell'età, ci sia un aumento esponenziale nelle probabilità nell'incorrere in un problema cardiovascolare. In particolare possiamo notare, come visualizzato anche dal boxplot, che superata la soglia dei 40 anni si ha un notevole aumento nella probabilità di CVD, confermando quindi come questo problema sia legato principalmente ad un fattore di età.

3.2 Sex

```
#Regressioni logistiche semplici
#Sex
fit.sex <- glm(nmc$cvd ~ nmc$sex, family=binomial)
summary(fit.sex)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4152  -0.4152  -0.2668  -0.2668   2.5898
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.31797    0.03654  -90.80  <2e-16 ***
## nmc$sexMale   0.91004    0.05021   18.12  <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: 13400  on 33326  degrees of freedom
## Residual deviance: 13073  on 33325  degrees of freedom
## AIC: 13077
##
## Number of Fisher Scoring iterations: 6
```

- Nella regressione logistica semplice, il sesso Maschile sembra aumentare notevolmente la possibilità di incorrere in un CVD rispetto al sesso Femminile, con valore stimato: $\text{SexMale} \sim 0.910$.
- La variabile Sex risulta significativa secondo il *p-value*, superando quindi il 5% di significatività.

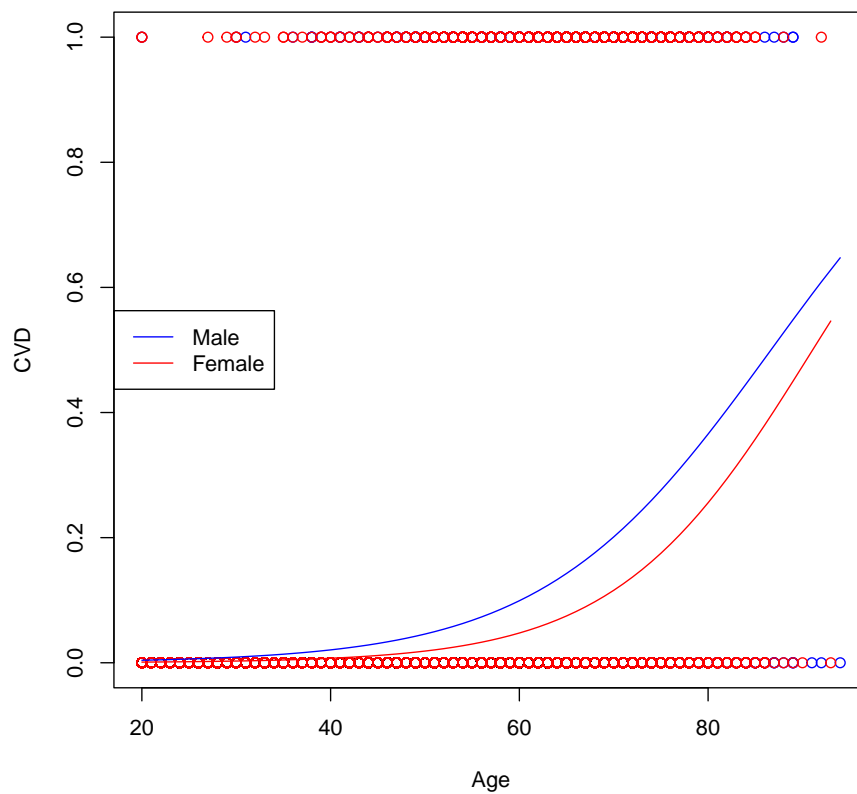
Valutiamo quanto il sesso possa influire nella presenza o meno di CVD.

```
#Modello per Maschio
fit.sex.male <- glm(nmc$cvd[nmc$sex=="Male"] ~
                    nmc$age[nmc$sex=="Male"],
                    family=binomial)
pstim.sex.male <- fit.sex.male$fitted.values
#Modello per Femmina
```

```

fit.sex.female <- glm(nmc$cvd[nmc$sex=="Female"] ~
                      nmc$age[nmc$sex=="Female"],
                      family=binomial)
pstim.sex.female <- fit.sex.female$fitted.values
#Plot
plot(nmc$age[nmc$sex=="Male"], nmc$cvd[nmc$sex=="Male"],
     xlab="Age", ylab="CVD", col="blue")
points(nmc$age[nmc$sex=="Female"], nmc$cvd[nmc$sex=="Female"],
       col="red")
lines(nmc$age[nmc$sex=="Male"], pstim.sex.male, col="blue")
lines(nmc$age[nmc$sex=="Female"], pstim.sex.female, col="red")
legend(x="left", legend=c("Male", "Female"), lty=c(1, 1),
      col=c("blue", "red"), lwd=1)

```



Il grafico ci conferma come il sesso maschile sia più a rischio di problemi cardiovascolari rispetto al sesso femminile.

3.3 BMI

```
#BMI
fit.bmi <- glm(nmc$cvd ~ nmc$bmi, family=binomial)
summary(fit.bmi)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$bmi, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3614  -0.3201  -0.3201  -0.3201   2.4481
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.94542     0.02605 -113.070 < 2e-16 ***
## nmc$bmi      0.24948     0.08995   2.773  0.00555 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 13393  on 33325  degrees of freedom
## AIC: 13397
##
## Number of Fisher Scoring iterations: 5
```

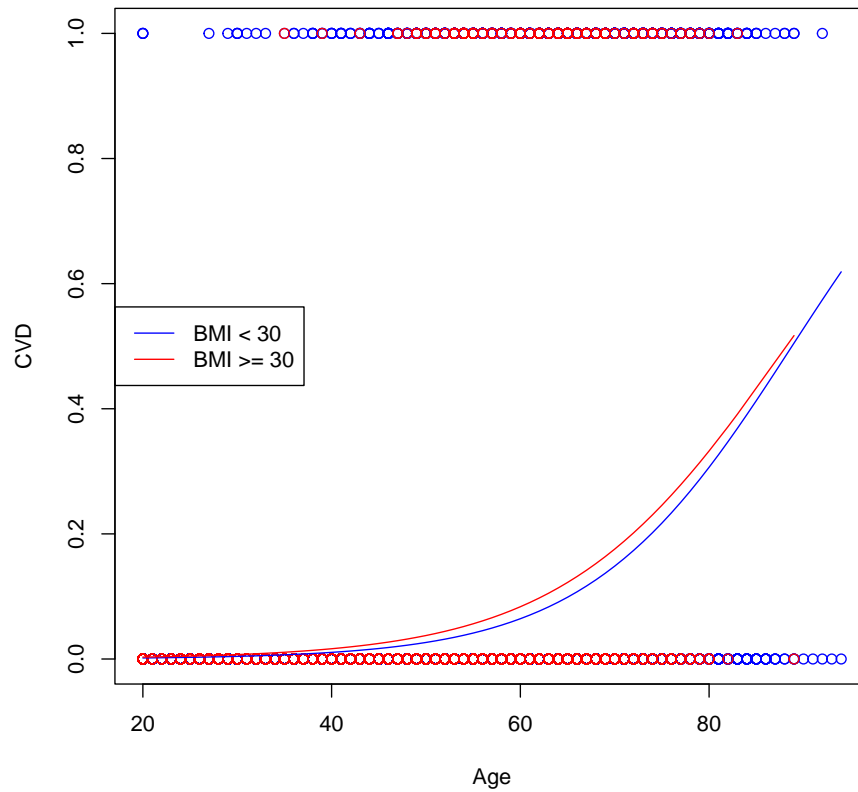
- La variabile BMI risulta positiva nell'insorgenza di un CVD con valore stimato: BMI \sim 0.249.
- La variabile BMI risulta significativa secondo il *p-value*.

Visualizziamo come il BMI possa influenzare nell'avanzamento dell'età.

```
#BMI 0
fit.bmi.0 <- glm(nmc$cvd[nmc$bmi==0] ~ nmc$age[nmc$bmi==0],
                 family=binomial)
pstim.bmi.0 <- fit.bmi.0$fitted.values

#BMI 1
fit.bmi.1 <- glm(nmc$cvd[nmc$bmi==1] ~ nmc$age[nmc$bmi==1],
                 family=binomial)
pstim.bmi.1 <- fit.bmi.1$fitted.values
```

```
#Plot
plot(nmc$age[nmc$bmi==0], nmc$cvd[nmc$bmi==0],
      xlab="Age", ylab="CVD", col="blue")
points(nmc$age[nmc$bmi==1], nmc$cvd[nmc$bmi==1], col="red")
lines(nmc$age[nmc$bmi==0], pstima.bmi.0, col="blue")
lines(nmc$age[nmc$bmi==1], pstima.bmi.1, col="red")
legend(x="left", legend=c("BMI < 30", "BMI >= 30"), lty=c(1, 1),
      col=c("blue", "red"), lwd=1)
```



Le due curve sono approssivamente simili, con un leggero aumento per coloro che hanno un indice di massa corporea maggiore di 30.

3.4 Fitness

```
#Fitness
fit.fitness <- glm(nmc$cvd ~ nmc$fitness, family=binomial)
summary(fit.fitness)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$fitness, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3438  -0.3299  -0.3166  -0.3166   2.5218
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.22195    0.09918  -32.487  <2e-16 ***
## nmc$fitness   0.08459    0.02723   3.106   0.0019 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 13390  on 33325  degrees of freedom
## AIC: 13394
##
## Number of Fisher Scoring iterations: 5
```

Contrariamente a quello che ci si potesse aspettare, per il solo modello di regressione logistica semplice, la variabile ordinale Fitness risulta, anche se di poco, positiva per l'insorgenza di un problema cardiovascolare.

Verifichiamo quindi se ci siano delle differenze nel modello di regressione logistica semplice con la variabile categoriale di Fitness.

```
#Fitness: Catoriale
fit.fitness.cat <- glm(nmc$cvd ~ fitness, family=binomial)
summary(fit.fitness.cat)

##
## Call:
## glm(formula = nmc$cvd ~ fitness, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3404  -0.3404  -0.3083  -0.3083   2.4894
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.81935     0.04066 -69.344 < 2e-16 ***
## fitnessJust as good -0.20312     0.05783  -3.512 0.000444 ***
## fitnessLittle Worse -0.23313     0.09169  -2.542 0.011009 *
## fitnessMuch better  -0.03049     0.07762  -0.393 0.694406
## fitnessMuch Worse   -0.09915     0.17360  -0.571 0.567914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 13384  on 33322  degrees of freedom
## AIC: 13394
##
## Number of Fisher Scoring iterations: 5
```

- Con la variabile categoriale di Fitness notiamo come ci sia una diminuzione nell'insorgenza di CVD per tutte le categorie.
- Solamente le categorie Fitness:Just as good e Fitness:Little Worse risultano significative.

Visulizziamo il comportamento della variabile Fitness all'aumentare dell'età.

```
#Fitness Much Worse
fit.fitness.muchworse <- glm(nmc$cvd[fitness=="Much Worse"] ~
                             nmc$age[fitness=="Much Worse"],
                             family=binomial)
pstim.fitness.muchworse <- fit.fitness.muchworse$fitted.values

#Fitness LittleWorse
fit.fitness.littleworse <- glm(nmc$cvd[fitness=="Little Worse"] ~
                               nmc$age[fitness=="Little Worse"],
                               family=binomial)
pstim.fitness.littleworse <- fit.fitness.littleworse$fitted.values

#Fitness Just as good
fit.fitness.justasgood<- glm(nmc$cvd[fitness=="Just as good"] ~
                             nmc$age[fitness=="Just as good"],
                             family=binomial)
pstim.fitness.justasgood <- fit.fitness.justasgood$fitted.values
```



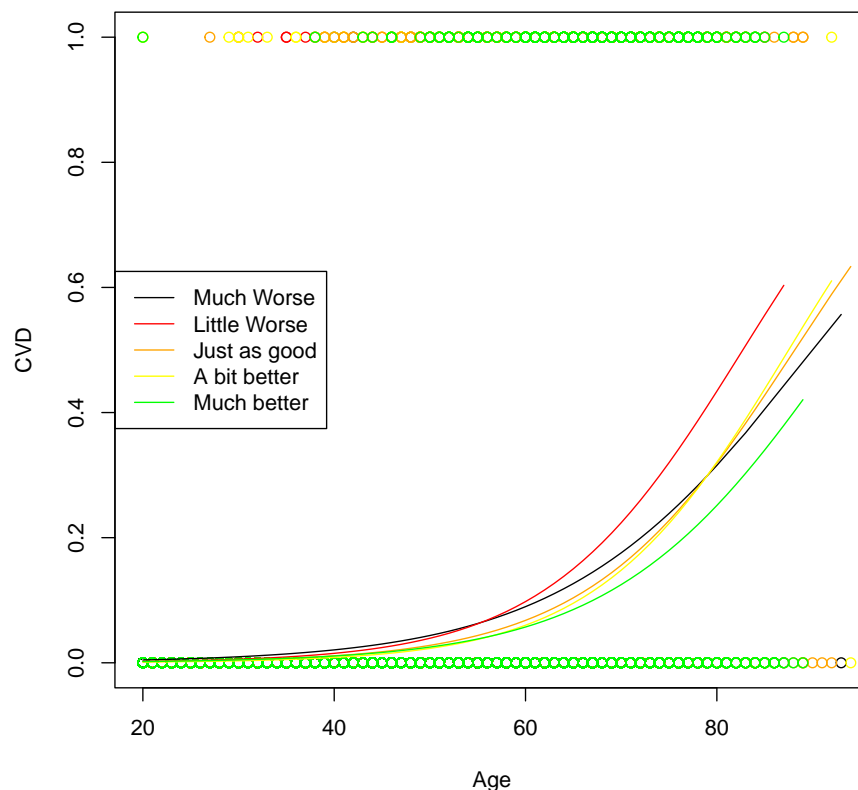
```

#Fitness A bit better
fit.fitness.abitbetter <- glm(nmc$cvd[fitness=="A bit better"] ~
                             nmc$age[fitness=="A bit better"],
                             family=binomial)
pstim.fitness.abitbetter <- fit.fitness.abitbetter$fitted.values

#Fitness Much better
fit.fitness.muchbetter <- glm(nmc$cvd[fitness=="Much better"] ~
                              nmc$age[fitness=="Much better"],
                              family=binomial)
pstim.fitness.muchbetter <- fit.fitness.muchbetter$fitted.values

#Plot
plot(nmc$age[fitness=="Much Worse"],nmc$cvd[fitness=="Much Worse"],
     xlab="Age", ylab="CVD", col="black")
points(nmc$age[fitness=="Little Worse"], nmc$cvd[fitness=="Little Worse"],
       col="red")
points(nmc$age[fitness=="Just as good"], nmc$cvd[fitness=="Just as good"],
       col="yellow")
points(nmc$age[fitness=="A bit better"], nmc$cvd[fitness=="A bit better"],
       col="orange")
points(nmc$age[fitness=="Much better"],nmc$cvd[fitness=="Much better"],
       col="green")
lines(nmc$age[fitness=="Much Worse"], pstim.fitness.muchworse,
      col="black")
lines(nmc$age[fitness=="Little Worse"], pstim.fitness.littleworse,
      col="red")
lines(nmc$age[fitness=="Just as good"], pstim.fitness.justasgood,
      col="orange")
lines(nmc$age[fitness=="A bit better"], pstim.fitness.abitbetter,
      col="yellow")
lines(nmc$age[fitness=="Much better"], pstim.fitness.muchbetter,
      col="green")
legend(x="left",
      legend=c("Much Worse", "Little Worse", "Just as good",
               "A bit better", "Much better"),
      lty=c(1, 1, 1, 1, 1),
      col=c("black","red", "orange", "yellow", "green"), lwd=1)

```



Attraverso il grafico notiamo che la categoria Fitness:Much better è quella meno soggetta rispetto a tutte le altre. Viceversa la categoria Fitness:Little Worse ha più probabilità di incorrere in un problema cardiovascolare.

Chi è della categoria Fitness:Much Worse ha meno probabilità rispetto alla categoria Fitness:Little Worse evidenziando come un problema cardiovascolare non è associato per forza a una pessima condizione di salute.

In conclusione, per il solo modello di regressione logistica semplice, consideriamo la variabile ordinale Fitness come significativa.

Nei successivi capitoli considereremo unicamente la variabile ordinale Fitness.

3.5 PA

```
#PA
fit.pa <- glm(nmc$cvd ~ nmc$pa, family=binomial)
summary(fit.pa)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$pa, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3242 -0.3242 -0.3242 -0.3242  2.4754
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.91978    0.02581 -113.126  <2e-16 ***
## nmc$pa       -0.09610    0.09974   -0.963    0.335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 13399  on 33325  degrees of freedom
## AIC: 13403
##
## Number of Fisher Scoring iterations: 5
```

Secondo la valutazione del *p-value* la variabile PA, nonostante influisca negativamente per la CVD, non supera il 5% di significatività, risultando non significativa.

3.6 Smoke

```
#Smoke
fit.smoke <- glm(nmc$cvd ~ nmc$smoke, family=binomial)
summary(fit.smoke)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$smoke, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.3402 -0.3186 -0.3186 -0.3186 2.4946
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.06590    0.09377 -32.696  <2e-16 ***
## nmc$smokeFormer  0.24571    0.10465  2.348  0.0189 *
## nmc$smokeNO      0.11061    0.09880  1.119  0.2629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 13392  on 33324  degrees of freedom
## AIC: 13398
##
## Number of Fisher Scoring iterations: 5
```

- Le categorie Smoke:Former e Smoke:NO sembrano influire positivamente sull'insorgenza di CVD.
- Risulta significativa solo la categoria Smoke:Former con valore stimato: Smoke:Former ~ 0.246 .

Verifichiamo ora il modello di regressione logistica semplice nel caso della variabile ordinale Smoke.

```
#Smoke: Ordinale
fit.smoke.ord <- glm(nmc$cvd ~ smoke.ord, family=binomial)
summary(fit.smoke.ord)

##
## Call:
## glm(formula = nmc$cvd ~ smoke.ord, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3281  -0.3249  -0.3218  -0.3218   2.4441
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.95522    0.06099 -48.454  <2e-16 ***
## smoke.ord    0.02015    0.03893  0.518   0.605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 13400  on 33325  degrees of freedom
## AIC: 13404
##
## Number of Fisher Scoring iterations: 5
```

- La variabile ordinale Smoke risulta positiva nell'insorgenza di CVD.
- Nonostante ciò la variabile Smoke ordinale risulta non significativa secondo il *p-value*.

Analizziamo se ci siano delle differenze tra le varie categorie di fumatori con l'avanzare dell'età.

```
#Smoke NO
fit.smoke.no <- glm(nmc$cvd[nmc$smoke=="NO"] ~
                    nmc$age[nmc$smoke=="NO"],
                    family=binomial)
pstim.smoke.no <- fit.smoke.no$fitted.values

#Smoke Former
fit.smoke.former <- glm(nmc$cvd[nmc$smoke=="Former"] ~
                        nmc$age[nmc$smoke=="Former"],
                        family=binomial)
pstim.smoke.former <- fit.smoke.former$fitted.values

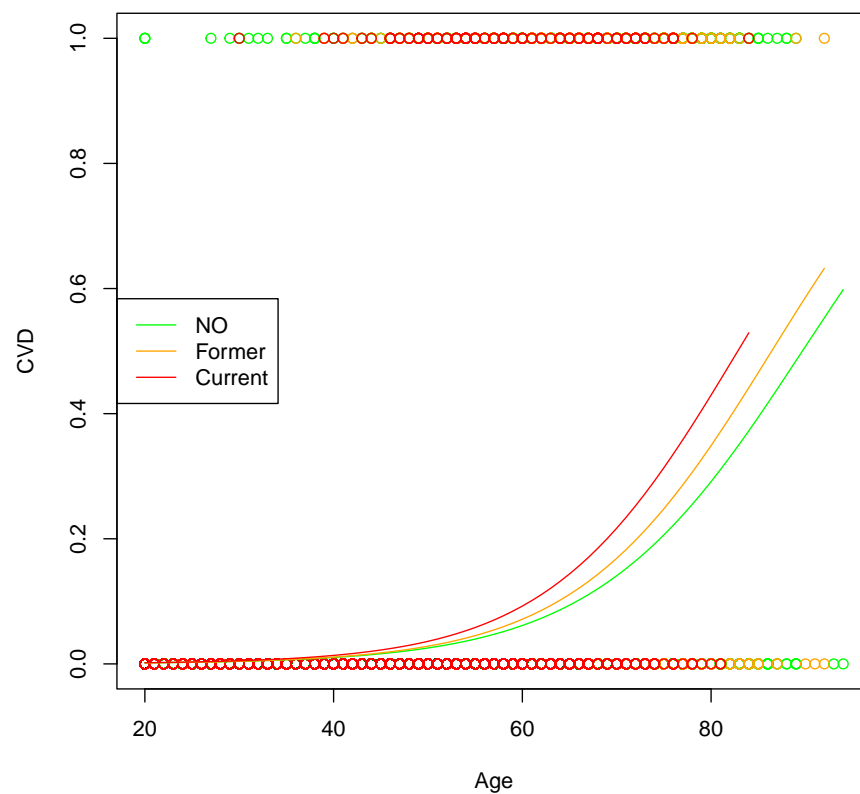
#Smoke Current
fit.smoke.current <- glm(nmc$cvd[nmc$smoke=="Current"] ~
                         nmc$age[nmc$smoke=="Current"],
                         family=binomial)
pstim.smoke.current <- fit.smoke.current$fitted.values

#Plot
plot(nmc$age[nmc$smoke=="NO"], nmc$cvd[nmc$smoke=="NO"],
     xlab="Age", ylab="CVD", col="green")
points(nmc$age[nmc$smoke=="Former"], nmc$cvd[nmc$smoke=="Former"],
       col="orange")
points(nmc$age[nmc$smoke=="Current"], nmc$cvd[nmc$smoke=="Current"],
       col="red")
lines(nmc$age[nmc$smoke=="NO"], pstim.smoke.no,
      col="green")
lines(nmc$age[nmc$smoke=="Former"], pstim.smoke.former,
```

```

col="orange")
lines(nmc$age[nmc$smoke=="Current"], pstim.smoke.current,
col="red")
legend(x="left",
legend=c("NO", "Former", "Current"),
lty=c(1, 1, 1),
col=c("green", "orange", "red"), lwd=1)

```



Possiamo notare come un un fumatore, rispetto alle altre categorie, abbia una maggiore probabilità di incorrere nella malattia con il passare del tempo.

Viceversa, il non fumatore ha meno probabilità rispetto alle altre categorie di incorrere nella malattia.

Nei successivi capitoli considereremo unicamente la variabile categoriale di Smoke.

3.7 Alchol

```
#Alchol
fit.alc <- glm(nmc$cvd ~ nmc$alc, family=binomial)
summary(fit.alc)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$alc, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3241  -0.3235  -0.3230  -0.3230   2.4425
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.934597   0.084928  -34.55  <2e-16 ***
## nmc$alc      0.003563   0.035652   0.10    0.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 13400  on 33325  degrees of freedom
## AIC: 13404
##
## Number of Fisher Scoring iterations: 5
```

La variabile Alchol, secondo la valutazione del *p-value*, non supera il 5% di significatività, risultando non significativa.

3.8 Commento

Nei soli modelli con regressione logistica semplice abbiamo che:

- Le variabili che risultano essere significative secondo la valutazione del *p-value* sono: Sex, Age, BMI e Fitness.
- Sempre secondo la valutazione del *p-value*, le variabili che invece risultano non significative sono: PA, Smoke e Alchol.
- Le variabili Sex:Male, Age e BMI aumentano il rischio di CVD.
- La variabile Fitness evidenzia il fatto che chi sta bene è meno soggetto alla problematica.
- Il fumatore è più soggetto alla malattia rispetto alle altre categorie.

4 Regressioni Logistiche Multiple

Consideriamo ora la regressione logistica multipla includendo tutte le variabili che sono presenti all'interno del dataset, verificando quali di esse sono più o meno significative per la visualizzazione di un primo modello unico.

4.1 Modello Completo

```
#Regressioni logistiche multiple
#Modello Completo
#Variabili: Sex, Age, BMI, Fitness, PA, Smoke, Alcohol
fit.all <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+nmc$fitness+
               nmc$pa+nmc$smoke+nmc$alc,
               family=binomial)

summary(fit.all)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$pa + nmc$smoke + nmc$alc, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5967  -0.3394  -0.1937  -0.0950   3.6484
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -7.475667   0.213543  -35.008 < 2e-16 ***
## nmc$sexMale     0.799132   0.054689   14.612 < 2e-16 ***
## nmc$age         0.092680   0.002446   37.896 < 2e-16 ***
## nmc$bmi         0.235120   0.096986    2.424 0.015339 *
## nmc$fitness    -0.181741   0.031070   -5.849 4.93e-09 ***
## nmc$pa         0.035563   0.108422    0.328 0.742909
## nmc$smokeFormer -0.332158   0.111102   -2.990 0.002793 **
## nmc$smokeNO    -0.374001   0.106486   -3.512 0.000444 ***
## nmc$alc        -0.056404   0.035625   -1.583 0.113368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10883  on 33318  degrees of freedom
## AIC: 10901
##
```



```
## Number of Fisher Scoring iterations: 7
```

Per il modello che include tutte le variabili:

Modello: $CVD \sim Sex + Age + BMI + Fitness + PA + Smoke + Alcohol$

- Risultano essere significative, secondo il *p-value*, le variabili: Sex, Age, BMI, Fitness e Smoke.
- Risultano essere non significative, non superando il 5% di significatività del *p-value*, le variabili: PA e Alcohol.
- I parametri stimati nella regressione logistica multipla differiscono da quelli presenti nelle regressioni logistiche semplici precedentemente analizzate.
- Gli errori standard non differiscono molto da quelli presenti nei modelli con regressione logistica semplice.
- La variabile Sex mostra ancora come il sesso Maschile influisca positivamente nella presenza di CVD con valore stimato: Sex:Male ~ 0.788 .
- Anche le variabili BMI e Smoke mostrano un aumento nelle possibilità di insorgenza di un CVD.
- La variabile Fitness aumenta di significatività, rispetto al modello di regressione logistica semplice, riducendo la probabilità di CVD con valore stimato: Fitness ~ -0.184 .

4.2 Modello Significativo

Dato che nel modello completo sono presenti variabili non significative, le andremo ad eliminare gradualmente dalla formula del modello fino ad ottenere un modello con solo variabili significative.

Iniziamo eliminando la variabile non significativa PA.

```
#Modello senza PA
#Variabili: Sex, Age, BMI, Fitness, Smoke, Alcohol
fit.npa <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+nmc$fitness+
               nmc$smoke+nmc$alc,
               family=binomial)
summary(fit.npa)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$alc, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.5978 -0.3371 -0.1941 -0.0950 3.6471
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.462934   0.209921 -35.551 < 2e-16 ***
## nmc$sexMale    0.799887   0.054643  14.638 < 2e-16 ***
## nmc$age        0.092640   0.002442  37.930 < 2e-16 ***
## nmc$bmi        0.235857   0.096958   2.433 0.014992 *
## nmc$fitness    -0.183877   0.030378  -6.053 1.42e-09 ***
## nmc$smokeFormer -0.332592   0.111097  -2.994 0.002756 **
## nmc$smokeNO     -0.374525   0.106476  -3.517 0.000436 ***
## nmc$alc        -0.056553   0.035625  -1.587 0.112413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10883  on 33319  degrees of freedom
## AIC: 10899
##
## Number of Fisher Scoring iterations: 7
```

Tutte le variabili che erano significative nel modello completo risultano ancora significative.

Eliminiamo la variabile Alchol, che risulta ancora non significativa, all'interno della formula.

```
#Modello significativo
#Variabili: Sex, Age, BMI, Fitness, Smoke
fit <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+nmc$fitness+
           nmc$smoke, family=binomial)
summary(fit)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6215  -0.3381  -0.1935  -0.0943   3.6515
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.614728   0.187445 -40.624 < 2e-16 ***
```

```
## nmc$sexMale      0.786417    0.053959   14.574   < 2e-16 ***
## nmc$age          0.092988    0.002437   38.159   < 2e-16 ***
## nmc$bmi          0.240200    0.096914    2.478 0.013194 *
## nmc$fitness      -0.186214    0.030344   -6.137 8.42e-10 ***
## nmc$smokeFormer -0.331879    0.111118   -2.987 0.002820 **
## nmc$smokeNO      -0.351977    0.105515   -3.336 0.000851 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10886  on 33320  degrees of freedom
## AIC: 10900
##
## Number of Fisher Scoring iterations: 7
```

Il modello risultate è:
Modello: CVD ~ Sex + Age + BMI + Fitness + Smoke

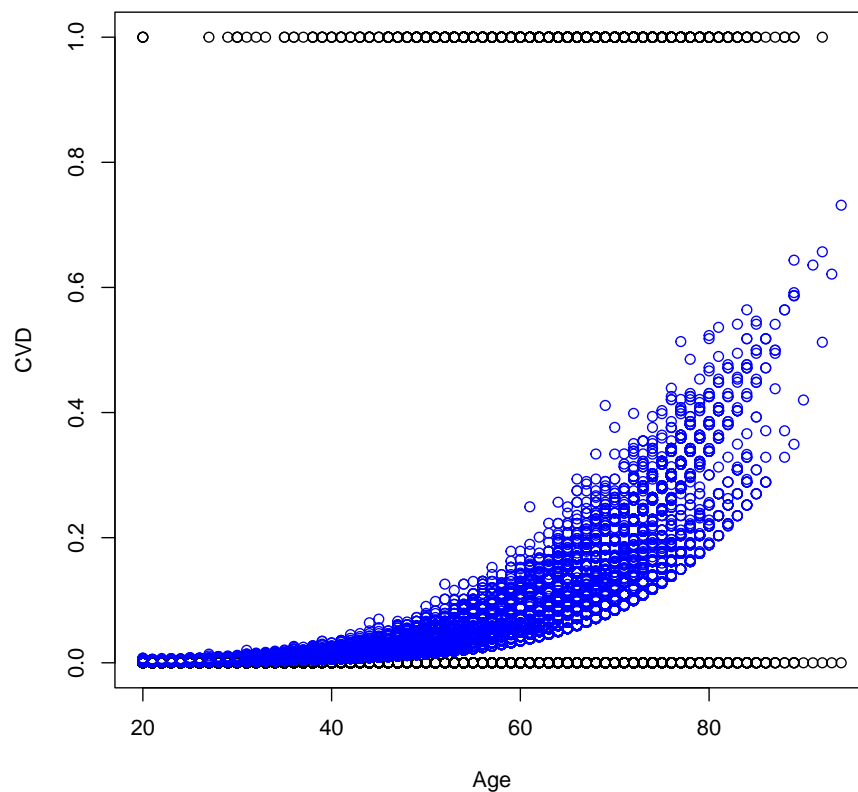
- Le variabili risultano essere tutte significative secondo il *p-value*.
- I parametri stimati e gli errori standard non differiscono molto dal modello completo.

Il modello con solo variabili significative sembra mostrare un buon adattamento.

Visualizziamo il grafico dell'andamento del modello stimato.

```
pstima <- fit$fitted.values

#Plot
plot(nmc$age, nmc$cvd, xlab="Age", ylab="CVD")
points(sort(nmc$age), pstima[order(nmc$age)], col="blue")
```



4.3 Commento

- Il modello risulta essere:
Modello: $CVD \sim \text{Sex} + \text{Age} + \text{BMI} + \text{Fitness} + \text{Smoke}$
- Come visto nelle regressioni logistiche semplici, le variabili sexMale, Age e BMI continuano ad influenzare positivamente la comparsa di problemi cardiovascolari.
- Al contrario, le variabili significative Fitness e Smoke (per la categoria "Former" e la categoria "NO") riducono la possibilità di avere un CVD.
- Di conseguenza la categoria Smoke:Current ha una probabilità maggiore nell'insorgenza di CVD.

4.4 Dati di esempio

Effettuiamo una valutazione della probabilità su degli individui casuali in base ai modelli precedentemente analizzati.

Visualizziamo i coefficienti del modello completo e quello significativo.

```
#Coefficienti Modello Completo
coef(fit.all)
```

	(Intercept)	nmc\$sexMale	nmc\$age	nmc\$bmi	nmc\$fitness
##	-7.47566727	0.79913169	0.09267966	0.23512007	-0.18174054
	nmc\$pa	nmc\$smokeFormer	nmc\$smokeNO	nmc\$alc	
##	0.03556271	-0.33215760	-0.37400050	-0.05640352	

```
#Coefficienti Modello Significativo
coef(fit)
```

	(Intercept)	nmc\$sexMale	nmc\$age	nmc\$bmi	nmc\$fitness
##	-7.61472806	0.78641659	0.09298766	0.24020007	-0.18621360
	nmc\$smokeFormer	nmc\$smokeNO			
##	-0.33187854	-0.35197684			

Verifichiamo la probabilità di avere un CVD per un Uomo (Sex 1) di 45 anni fumatore, con BMI pari a 32 (BMI dicotomizzato a 1), che non fa consumo di alcohol (Alchol 1), in ottima salute (Fitness 5) e PA=1.

```
#Stima per il Modello Completo
#Dato
#Intercetta: 1, Sex: 1, Age: 45, BMI: 1, Fitness: 5,
#PA: 1, Smoke:Former: 0, Smoke:NO: 0, Alchol: 1
man45.all <- c(1, 1, 45, 1, 5, 1, 0, 0, 1)
stima.man45.all <- exp(coef(fit.all)%*%man45.all)/
  (1+exp(coef(fit.all)%*%man45.all))
stima.man45.all
```

##	[,1]
##	[1,] 0.03915164

```
#Stima per il Modello Significativo
#Dato
#Intercetta: 1, Sex: 1, Age: 45, BMI: 1, Fitness: 5,
#Smoke:Former: 0, Smoke:NO: 0
man45 <- c(1, 1, 45, 1, 5, 0, 0)
stima.man45 <- exp(coef(fit)%*%man45)/(1+exp(coef(fit)%*%man45))
stima.man45
```

```
##           [,1]
## [1,] 0.03439862
```

Risulta che:

- per il Modello Completo: $\hat{\pi} = 0.0391$
- per il Modello Significativo: $\hat{\pi} = 0.0343$

Verifichiamo adesso la probabilità di avere un CVD per una Donna (Sex 0) di 60 anni ex fumatrice, con BMI pari a 35 (BMI dicotomizzato 1), che beve alcohol nella media (Alchol 4), in buona salute (Fitness 4), e PA = 1.

```
#Stima per il Modello Completo
#Dato
#Intercetta: 1, Sex: 0, Age: 60, BMI: 1, Fitness: 4,
#PA: 1, Smoke:Former: 1, Smoke:NO: 0, Alchol: 4
woman60.all <- c(1, 0, 60, 1, 4, 1, 1, 0, 4)
stima.woman60.all <- exp(coef(fit.all)%*%woman60.all)/
                    (1+exp(coef(fit.all)%*%woman60.all))

stima.woman60.all

##           [,1]
## [1,] 0.05074142
```

```
#Stima per il Modello Significativo
#Intercetta: 1, Sex: 0, Age: 60, BMI: 1, Fitness: 4,
#Smoke:Former: 1, Smoke:NO: 0
woman60 <- c(1, 0, 60, 1, 4, 1, 0)
stima.woman60 <- exp(coef(fit)%*%woman60)/(1+exp(coef(fit)%*%woman60))
stima.woman60

##           [,1]
## [1,] 0.05355512
```

Risulta che:

- Modello Completo: $\hat{\pi} = 0.0507$
- Modello Significativo: $\hat{\pi} = 0.0535$

In conclusione possiamo vedere come la probabilità del modello completo e del modello con solo variabili significative si mostrino particolarmente simili.

5 Considerazioni sul Modello

In questo capitolo confronteremo il comportamento del modello su una specifica categoria di individui.

5.1 Maschio e Femmina

Nei precedenti capitoli, durante l'analisi delle singole variabili e dei vari modelli, abbiamo notato come ci siano state sempre diverse probabilità tra il sesso maschile e il sesso femminile.

Valutiamo ancora all'interno di un grafico se questa nostra ipotesi si verifica in un modello più complesso di quello marginale.

```
#Modello Sesso Maschile
#Sex: Male
fit.male <- glm(nmc$cvd[nmc$sex=="Male"] ~
               nmc$age[nmc$sex=="Male"]+
               nmc$bmi[nmc$sex=="Male"]+
               nmc$fitness[nmc$sex=="Male"]+
               nmc$smoke[nmc$sex=="Male"],
               family=binomial)
summary(fit.male)

##
## Call:
## glm(formula = nmc$cvd[nmc$sex == "Male"] ~ nmc$age[nmc$sex ==
##      "Male"] + nmc$bmi[nmc$sex == "Male"] + nmc$fitness[nmc$sex ==
##      "Male"] + nmc$smoke[nmc$sex == "Male"], family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5009  -0.4634  -0.2744  -0.1136   3.4503
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -6.310134    0.257771  -24.480  < 2e-16 ***
## nmc$age[nmc$sex == "Male"]    0.086098    0.003316   25.968  < 2e-16 ***
## nmc$bmi[nmc$sex == "Male"]    0.135192    0.157317    0.859 0.390142
## nmc$fitness[nmc$sex == "Male"] -0.156521    0.042188   -3.710 0.000207 ***
## nmc$smoke[nmc$sex == "Male"]Former -0.472766    0.146244   -3.233 0.001226 **
## nmc$smoke[nmc$sex == "Male"]NO    -0.579023    0.142624   -4.060 4.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
## Null deviance: 6344.1 on 11129 degrees of freedom
## Residual deviance: 5297.3 on 11124 degrees of freedom
## AIC: 5309.3
##
## Number of Fisher Scoring iterations: 6

pstima.male <- fit.male$fitted.values
```

```
#Modello Sesso Femminile
#Sex: Female
fit.female <- glm(nmc$cvd[nmc$sex=="Female"] ~
  nmc$age[nmc$sex=="Female"]+
  nmc$bmi[nmc$sex=="Female"]+
  nmc$fitness[nmc$sex=="Female"]+
  nmc$smoke[nmc$sex=="Female"],
  family=binomial)
summary(fit.female)

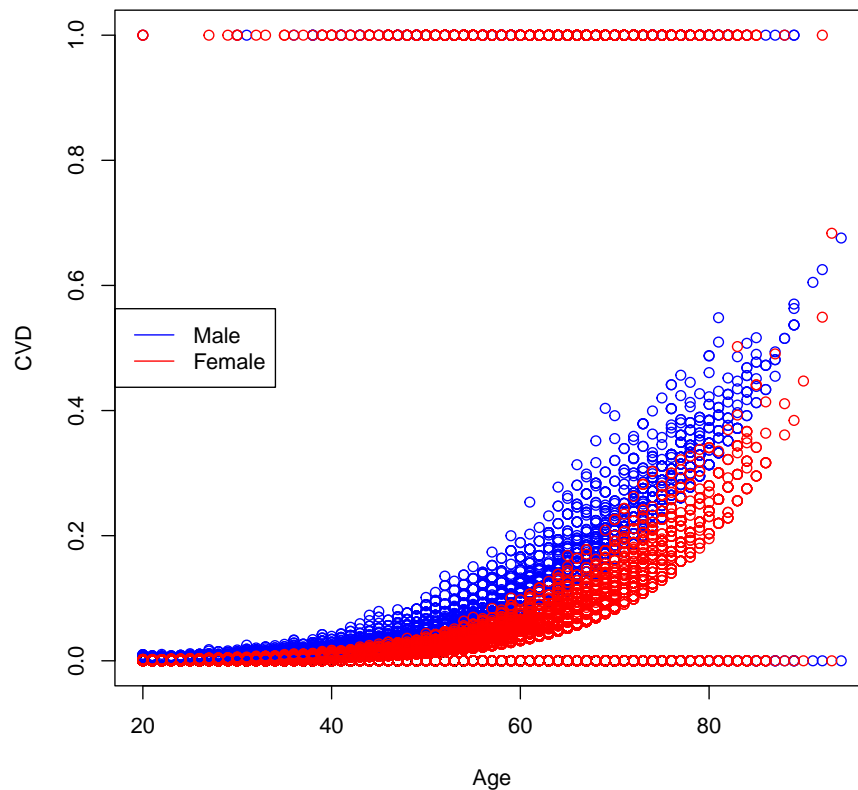
##
## Call:
## glm(formula = nmc$cvd[nmc$sex == "Female"] ~ nmc$age[nmc$sex ==
## "Female"] + nmc$bmi[nmc$sex == "Female"] + nmc$fitness[nmc$sex ==
## "Female"] + nmc$smoke[nmc$sex == "Female"], family = binomial)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.5166 -0.2829 -0.1632 -0.0847 3.7175
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.132620 0.277110 -29.348 < 2e-16 ***
## nmc$age[nmc$sex == "Female"] 0.099409 0.003598 27.630 < 2e-16 ***
## nmc$bmi[nmc$sex == "Female"] 0.284497 0.123459 2.304 0.0212 *
## nmc$fitness[nmc$sex == "Female"] -0.210839 0.043547 -4.842 1.29e-06 ***
## nmc$smoke[nmc$sex == "Female"]Former -0.182839 0.174076 -1.050 0.2936
## nmc$smoke[nmc$sex == "Female"]NO -0.132124 0.160494 -0.823 0.4104
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 6729.3 on 22196 degrees of freedom
## Residual deviance: 5573.1 on 22191 degrees of freedom
## AIC: 5585.1
##
```



```
## Number of Fisher Scoring iterations: 7

pstima.female <- fit.female$fitted.values
```

```
#Plot
plot(nmc$age[nmc$sex=="Male"], nmc$cvd[nmc$sex=="Male"],
     xlab="Age", ylab="CVD", col="blue")
points(nmc$age[nmc$sex=="Female"], nmc$cvd[nmc$sex=="Female"],
       col="red")
points(sort(nmc$age[nmc$sex=="Male"]),
       pstima.male[order(nmc$age[nmc$sex=="Male"])] ,
       col="blue")
points(sort(nmc$age[nmc$sex=="Female"]),
       pstima.female[order(nmc$age[nmc$sex=="Female"])] ,
       col="red")
legend(x="left", legend=c("Male", "Female"), lty=c(1, 1),
      col=c("blue", "red"), lwd=1)
```



Possiamo notare ancora come statisticamente il sesso maschile rimanga il soggetto che ha più rischi di CVD rispetto al genere femminile.

5.2 Attività Fisica

Dato che la variabile PA nelle valutazioni dei modelli è sempre stata scartata, verifichiamo se all'interno del nostro modello possono esserci delle differenze tra le due categorie di PA per il calcolo del CVD.

```
#Modello PA 0
fit.pa.0 <- glm(nmc$cvd[nmc$pa==0] ~ nmc$sex[nmc$pa==0] +
               nmc$age[nmc$pa==0] + nmc$bmi[nmc$pa==0] +
               nmc$fitness[nmc$pa==0] + nmc$smoke[nmc$pa==0],
               family=binomial)
summary(fit.pa.0)
```

```
##
## Call:
## glm(formula = nmc$cvd[nmc$pa == 0] ~ nmc$sex[nmc$pa == 0] + nmc$age[nmc$pa ==
## 0] + nmc$bmi[nmc$pa == 0] + nmc$fitness[nmc$pa == 0] + nmc$smoke[nmc$pa ==
## 0], family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6202  -0.3396  -0.1936  -0.0931   3.6511
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -7.644393    0.199721 -38.275 < 2e-16 ***
## nmc$sex[nmc$pa == 0]Male    0.782201    0.055911  13.990 < 2e-16 ***
## nmc$age[nmc$pa == 0]       0.092985    0.002545  36.533 < 2e-16 ***
## nmc$bmi[nmc$pa == 0]       0.187919    0.104915   1.791  0.07327 .
## nmc$fitness[nmc$pa == 0]   -0.175433    0.032345  -5.424 5.84e-08 ***
## nmc$smoke[nmc$pa == 0]Former -0.346614    0.116950  -2.964  0.00304 **
## nmc$smoke[nmc$pa == 0]NO    -0.353123    0.111011  -3.181  0.00147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 12486  on 30907  degrees of freedom
## Residual deviance: 10141  on 30901  degrees of freedom
## AIC: 10155
##
## Number of Fisher Scoring iterations: 7

pstimapa.0 <- fit.pa.0$fitted.values
```

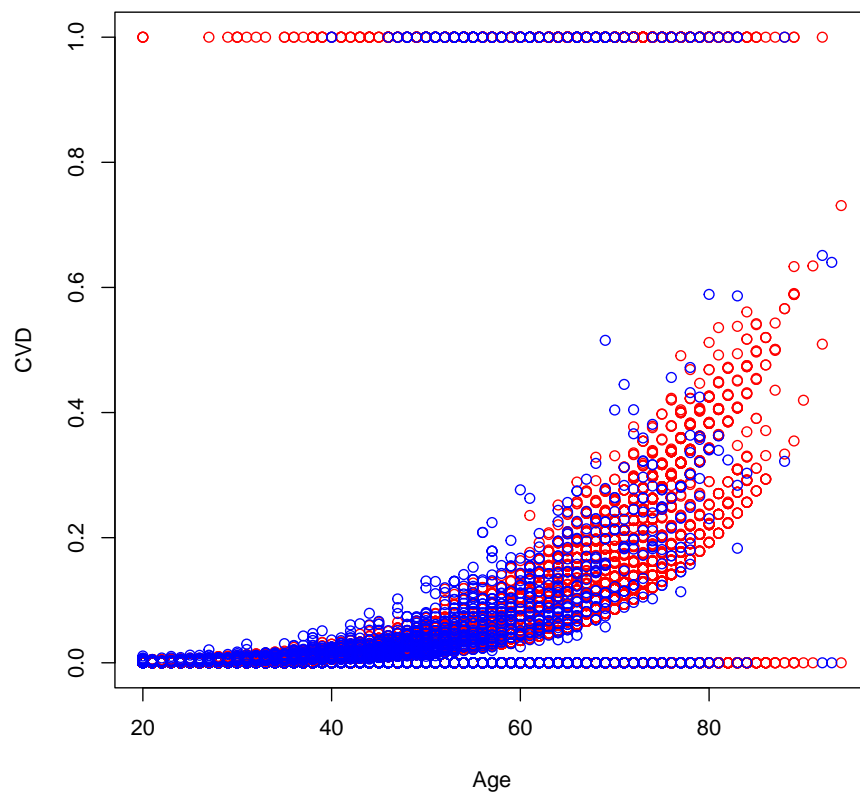
```
#Modello PA 1
fit.pa.1 <- glm(nmc$cvd[nmc$pa==1] ~ nmc$sex[nmc$pa==1] +
               nmc$age[nmc$pa==1] + nmc$bmi[nmc$pa==1] +
               nmc$fitness[nmc$pa==1] + nmc$smoke[nmc$pa==1],
               family=binomial)
summary(fit.pa.1)

##
## Call:
## glm(formula = nmc$cvd[nmc$pa == 1] ~ nmc$sex[nmc$pa == 1] + nmc$age[nmc$pa ==
## 1] + nmc$bmi[nmc$pa == 1] + nmc$fitness[nmc$pa == 1] + nmc$smoke[nmc$pa ==
## 1], family = binomial)
##
```

```
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.4512  -0.3135  -0.1902  -0.1060   2.9592
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -7.463732    0.614386  -12.148 < 2e-16 ***
## nmc$sex[nmc$pa == 1]Male    0.833173    0.210572   3.957 7.6e-05 ***
## nmc$age[nmc$pa == 1]      0.093239    0.008655  10.773 < 2e-16 ***
## nmc$bmi[nmc$pa == 1]      0.543559    0.264374   2.056 0.0398 *
## nmc$fitness[nmc$pa == 1]  -0.284719    0.111611  -2.551 0.0107 *
## nmc$smoke[nmc$pa == 1]Former -0.184548    0.359510  -0.513 0.6077
## nmc$smoke[nmc$pa == 1]NO   -0.347033    0.343785  -1.009 0.3128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 913.04  on 2418  degrees of freedom
## Residual deviance: 740.78  on 2412  degrees of freedom
## AIC: 754.78
##
## Number of Fisher Scoring iterations: 7

pstimapa.1 <- fit.pa.1$fitted.values

#Plot
plot(nmc$age[nmc$pa==0], nmc$cvd[nmc$pa==0], xlab="Age", ylab="CVD", col="red")
points(nmc$age[nmc$pa==1], nmc$cvd[nmc$pa==1], col="blue")
points(nmc$age[nmc$pa==0], pstimapa.0, col="red")
points(nmc$age[nmc$pa==1], pstimapa.1, col="blue")
```



Come è possibile notare, anche nella visualizzazione del grafico, non sembrano mostrarci differenze tra le due categorie di PA.

6 Interazioni fra le variabili

Valutiamo se all'interno del modello ci sia la possibilità di interazioni fra le variabili.

Consideriamo i casi nei quali le variabili come Smoke, Alchol, PA o Sex possano interagire con le altre variabili, limitandoci unicamente nelle interazioni del secondo ordine.

6.1 Smoke e Alchol

Analizziamo il caso nel quale il consumo di alchol, combinato con l'uso di sigaretta, possa o meno aumentare le probabilità di CVD.

```
#Modello con interazione: Smoke e Alchol
fit.smokealchol <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+
                        nmc$fitness+nmc$smoke+
                        nmc$smoke*nmc$alc,
                        family=binomial)

summary(fit.smokealchol)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$smoke * nmc$alc, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6094  -0.3392  -0.1936  -0.0952   3.6495
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.777741    0.409428  -18.997  < 2e-16 ***
## nmc$sexMale      0.800386    0.054612   14.656  < 2e-16 ***
## nmc$age          0.092757    0.002449   37.869  < 2e-16 ***
## nmc$bmi          0.233538    0.097010    2.407   0.0161 *
## nmc$fitness     -0.183944    0.030378   -6.055  1.4e-09 ***
## nmc$smokeFormer  0.239303    0.425957    0.562   0.5743
## nmc$smokeNO     -0.120962    0.395698   -0.306   0.7598
## nmc$alc          0.062892    0.142152    0.442   0.6582
## nmc$smokeFormer:nmc$alc -0.223037    0.158794   -1.405   0.1602
## nmc$smokeNO:nmc$alc  -0.094178    0.148217   -0.635   0.5252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10880  on 33317  degrees of freedom
## AIC: 10900
##
## Number of Fisher Scoring iterations: 7
```

I dati sembrano non mostrare l'interazione fra Smoke e Alchol.

6.2 Smoke e BMI

Vediamo se l'uso di sigaretta per una persona con un alto indice di massa corporea possa aumentarne le probabilità.

```
#Modello con interazione: Smoke e BMI
fit.smokebmi <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+
                    nmc$fitness+nmc$smoke+
                    nmc$smoke*nmc$bmi,
                    family=binomial)
summary(fit.smokebmi)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$smoke * nmc$bmi, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6167  -0.3389  -0.1923  -0.0938   3.6547
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.538233    0.187909  -40.116  < 2e-16 ***
## nmc$sexMale      0.789657    0.054027   14.616  < 2e-16 ***
## nmc$age          0.092953    0.002438   38.128  < 2e-16 ***
## nmc$bmi         -1.495180    0.726032   -2.059  0.039457 *
## nmc$fitness     -0.186487    0.030356   -6.143  8.08e-10 ***
## nmc$smokeFormer -0.400699    0.113377   -3.534  0.000409 ***
## nmc$smokeNO     -0.438392    0.107190   -4.090  4.32e-05 ***
## nmc$bmi:nmc$smokeFormer 1.667155    0.744567    2.239  0.025150 *
## nmc$bmi:nmc$smokeNO    1.874406    0.734922    2.550  0.010757 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
```

```
## Residual deviance: 10874 on 33318 degrees of freedom
## AIC: 10892
##
## Number of Fisher Scoring iterations: 7
```

A differenza di Smoke e Alchol, l'interazione tra Smoke e BMI sembra mostrare un'interazione significativa, variando il valore atteso e il significato della variabile BMI.

In questo caso la variabile BMI assume valore negativo, influenzando negativamente nella comparsa di CVD.

6.3 Alchol e BMI

Come per il caso di Smoke, verifichiamo se il consumo di alchol associato ad un maggior indice di massa corporea influisce nella probabilità di CVD.

```
#Modello con interazione: Alchol e BMI
fit.alcholbmi <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+
                    nmc$fitness+nmc$smoke+
                    nmc$alc*nmc$bmi,
                    family=binomial)

summary(fit.alcholbmi)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$alc * nmc$bmi, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5941  -0.3386  -0.1936  -0.0950   3.6460
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -7.436723    0.211379  -35.182  < 2e-16 ***
## nmc$sexMale     0.798024    0.054655   14.601  < 2e-16 ***
## nmc$age         0.092645    0.002442   37.932  < 2e-16 ***
## nmc$bmi        -0.031850    0.282019   -0.113  0.910080
## nmc$fitness    -0.184203    0.030379   -6.063  1.33e-09 ***
## nmc$smokeFormer -0.332279    0.111100   -2.991  0.002782 **
## nmc$smokeNO    -0.374629    0.106482   -3.518  0.000434 ***
## nmc$alc        -0.067192    0.037109   -1.811  0.070191 .
## nmc$bmi:nmc$alc  0.120560    0.118070    1.021  0.307213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```



```
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10882  on 33318  degrees of freedom
## AIC: 10900
##
## Number of Fisher Scoring iterations: 7
```

A differenza di Smoke*BMI, l'interazione tra Alcohol e BMI non è supportata.

6.4 Sex e Smoke

Verifichiamo se l'utilizzo di sigaretta sia peggiorativo in uno dei due sessi.

```
#Modello con interazione: Sex e Smoke
fit.sexsmoke <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+
                    nmc$fitness+nmc$smoke+
                    nmc$sex*nmc$smoke,
                    family=binomial)
summary(fit.sexsmoke)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$sex * nmc$smoke, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5936  -0.3413  -0.1902  -0.0948   3.6359
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.823919    0.218529  -35.803  < 2e-16 ***
## nmc$sexMale      1.213811    0.200893   6.042 1.52e-09 ***
## nmc$age          0.092548    0.002447  37.822  < 2e-16 ***
## nmc$bmi          0.237345    0.096938   2.448  0.0143 *
## nmc$fitness     -0.182865    0.030381  -6.019 1.75e-09 ***
## nmc$smokeFormer -0.168152    0.173095  -0.971  0.3313
## nmc$smokeNO     -0.086983    0.158493  -0.549  0.5831
## nmc$sexMale:nmc$smokeFormer -0.332887    0.226013  -1.473  0.1408
## nmc$sexMale:nmc$smokeNO    -0.513869    0.211536  -2.429  0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10879  on 33318  degrees of freedom
## AIC: 10897
##
## Number of Fisher Scoring iterations: 7
```

L'interazione fra le variabili Sex e Smoke risulta non significativa.

6.5 Sex e Age

Analizziamo ora il caso nel quale l'aumento dell'età possa influenzare in maniera differente tra i due sessi.

```
#Modello con interazione: Sex e Age
fit.sexage <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+
                  nmc$fitness+nmc$smoke+
                  nmc$sex*nmc$age,
                  family=binomial)
summary(fit.sexage)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$sex * nmc$age, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5257  -0.3426  -0.1892  -0.0925   3.7386
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.072593   0.247550 -32.610 < 2e-16 ***
## nmc$sexMale     1.686372   0.306886   5.495 3.90e-08 ***
## nmc$age         0.100329   0.003529  28.427 < 2e-16 ***
## nmc$bmi         0.233344   0.096960   2.407 0.016102 *
## nmc$fitness    -0.185576   0.030249  -6.135 8.52e-10 ***
## nmc$smokeFormer -0.328928   0.111095  -2.961 0.003069 **
## nmc$smokeNO    -0.364822   0.105629  -3.454 0.000553 ***
## nmc$sexMale:nmc$age -0.014186   0.004760  -2.980 0.002879 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10877  on 33319  degrees of freedom
```

```
## AIC: 10893
##
## Number of Fisher Scoring iterations: 7
```

Contrariamente a quello che ci si poteva aspettare, esiste un interazione significativa tra la variabile Sex e Age. Per il sesso maschile con l'aumentare dell'età ha, anche se piccola, una riduzione nella probabilità di CVD.

Analizzeremo successivamente se questa interazione può risultare utile ai fini del nostro problema.

6.6 PA e Age

Verifichiamo se l'attività fisica di un individuo è influenzata in base alla sua età.

```
#Modello con interazione PA e Age
fit.sexsmoke <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+
                    nmc$fitness+nmc$smoke+
                    nmc$pa*nmc$age,
                    family=binomial)
summary(fit.sexsmoke)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$pa * nmc$age, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6223  -0.3380  -0.1931  -0.0944   3.6547
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -7.637524   0.196623  -38.844 < 2e-16 ***
## nmc$sexMale     0.785270   0.054030   14.534 < 2e-16 ***
## nmc$age         0.093182   0.002543   36.636 < 2e-16 ***
## nmc$bmi         0.239306   0.096938    2.469 0.013563 *
## nmc$fitness    -0.183966   0.031043   -5.926 3.1e-09 ***
## nmc$smokeFormer -0.331027   0.111137   -2.979 0.002896 **
## nmc$smokeNO     -0.351321   0.105525   -3.329 0.000871 ***
## nmc$pa         0.150329   0.532956    0.282 0.777893
## nmc$age:nmc$pa -0.001873   0.008691   -0.216 0.829356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10886  on 33318  degrees of freedom
## AIC: 10904
##
## Number of Fisher Scoring iterations: 7
```

Non è verificata l'interazione fra le variabili PA e Age.

6.7 PA e Fitness

Analizziamo il caso nel quale l'attività fisica e lo stato di salute di un individuo possano aumentare le probabilità di CVD.

```
#Modello con interazione PA e Fitness
fit.sexsmoke <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+
                    nmc$fitness+nmc$smoke+
                    nmc$pa*nmc$fitness,
                    family=binomial)
summary(fit.sexsmoke)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$pa * nmc$fitness, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6172  -0.3387  -0.1939  -0.0944   3.6542
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.66807    0.19391  -39.545 < 2e-16 ***
## nmc$sexMale      0.78531    0.05400   14.543 < 2e-16 ***
## nmc$age          0.09301    0.00244   38.121 < 2e-16 ***
## nmc$bmi          0.23582    0.09706    2.430 0.015113 *
## nmc$fitness     -0.17257    0.03220   -5.358 8.4e-08 ***
## nmc$smokeFormer -0.33083    0.11115   -2.976 0.002917 **
## nmc$smokeNO     -0.34965    0.10557   -3.312 0.000926 ***
## nmc$pa           0.44155    0.31777    1.390 0.164666
## nmc$fitness:nmc$pa -0.14624    0.10964   -1.334 0.182252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
## Null deviance: 13400 on 33326 degrees of freedom
## Residual deviance: 10884 on 33318 degrees of freedom
## AIC: 10902
##
## Number of Fisher Scoring iterations: 7
```

Il modello mostra come non ci sia interazione fra le variabili PA e Fitness.

6.8 Modello con interazioni

Analizziamo ora il modello con solo variabili significative aggiungendo le interazioni che precedentemente abbiamo valutato come significative.

Il modello da valutare sarà quindi:

Modello: $CVD \sim Sex + Age + BMI + Fitness + Smoke + Sex*Age + Smoke*BMI$.

```
#Modello con interazione: Sex*Age + Smoke*BMI
fit.int <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+
               nmc$fitness+nmc$smoke+
               nmc$sex*nmc$age+
               nmc$smoke*nmc$bmi,
               family=binomial)
summary(fit.int)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$sex * nmc$age + nmc$smoke * nmc$bmi, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5214  -0.3437  -0.1885  -0.0913   3.7412
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.994945    0.248199  -32.212  < 2e-16 ***
## nmc$sexMale      1.684991    0.307149   5.486 4.11e-08 ***
## nmc$age          0.100260    0.003532  28.386  < 2e-16 ***
## nmc$bmi         -1.494470    0.726361  -2.057 0.039641 *
## nmc$fitness     -0.185813    0.030262  -6.140 8.24e-10 ***
## nmc$smokeFormer -0.396574    0.113345  -3.499 0.000467 ***
## nmc$smokeNO     -0.450336    0.107284  -4.198 2.70e-05 ***
## nmc$sexMale:nmc$age -0.014114    0.004764  -2.963 0.003050 **
## nmc$bmi:nmc$smokeFormer 1.656064    0.744884   2.223 0.026199 *
## nmc$bmi:nmc$smokeNO   1.867843    0.735274   2.540 0.011075 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10866  on 33317  degrees of freedom
## AIC: 10886
##
## Number of Fisher Scoring iterations: 7
```

6.9 Commento

Nonostante il modello con le due interazioni Smoke*BMI e Sex*Age risulti significativo, vediamo come questo si comporti in maniera differente dalle valutazioni che abbiamo analizzato precedentemente.

Il modello con interazioni mostra una minor probabilità per un individuo che fuma e con alto indice di massa corporea, o come un individuo di sesso maschile abbia una minima riduzione di probabilità con l'aumentare dell'età.

Inoltre il significato della variabile BMI varia rispetto al modello con solo variabili significative e al modello con la sola regressione logistica semplice, diminuendone anche la significatività.

Probabilmente la comparsa di queste interazioni potrebbero essere causate da un problema di overfitting, dove il modello cerca di avvicinarsi di più ai dati del dataset e meno ad una probabilità generica nella valutazione di un CVD.

Decido quindi di non considerare questo modello perché, oltre ad aumentarne il grado, non fornire un contributo decisivo per il nostro problema, andando contro anche alle analisi che fino a qui abbiamo valutato.

7 Selezione del Modello

Utilizziamo adesso i metodi Backward, Forward e Both basati sui criteri di penalizzazione AIC e BIC per la selezione del modello.

Per eseguire le varie procedure, prenderemo in considerazione la formula base con solo l'intercetta e il modello che comprende tutte le variabili fornite dal dataset, senza interazioni.

```
#Inizilizziamo la formula base con intercetta  
fit.0 <- glm(nmc$cvd ~ 1, family= "binomial")
```

7.1 Backward

Verifichiamo le formule della procedura Backward con AIC e BIC.

7.1.1 AIC

```
#Backward: AIC  
backward.AIC <- step(fit.all, direction="backward", k=2)  
  
## Start:  AIC=10901.21  
## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness + nmc$pa +  
##      nmc$smoke + nmc$alc  
##  
##           Df Deviance   AIC  
## - nmc$pa      1    10883 10899  
## <none>         10883 10901  
## - nmc$alc      1    10886 10902  
## - nmc$bmi      1    10889 10905  
## - nmc$smoke    2    10895 10909  
## - nmc$fitness  1    10917 10933  
## - nmc$sex      1    11097 11113  
## - nmc$age      1    13043 13059  
##  
## Step:  AIC=10899.32  
## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness + nmc$smoke +  
##      nmc$alc  
##  
##           Df Deviance   AIC  
## <none>         10883 10899  
## - nmc$alc      1    10886 10900  
## - nmc$bmi      1    10889 10903  
## - nmc$smoke    2    10895 10907  
## - nmc$fitness  1    10920 10934
```

```
## - nmc$sex      1      11098 11112
## - nmc$age      1      13045 13059

formula(backward.AIC)

## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness + nmc$smoke +
##      nmc$alc

summary(backward.AIC)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##      nmc$smoke + nmc$alc, family = binomial)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5978  -0.3371  -0.1941  -0.0950   3.6471
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -7.462934   0.209921 -35.551 < 2e-16 ***
## nmc$sexMale    0.799887   0.054643  14.638 < 2e-16 ***
## nmc$age        0.092640   0.002442  37.930 < 2e-16 ***
## nmc$bmi        0.235857   0.096958   2.433 0.014992 *
## nmc$fitness    -0.183877   0.030378  -6.053 1.42e-09 ***
## nmc$smokeFormer -0.332592   0.111097  -2.994 0.002756 **
## nmc$smokeNO    -0.374525   0.106476  -3.517 0.000436 ***
## nmc$alc       -0.056553   0.035625  -1.587 0.112413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10883  on 33319  degrees of freedom
## AIC: 10899
##
## Number of Fisher Scoring iterations: 7
```

7.1.2 BIC

```
#Backward: BIC
backward.BIC <- step(fit.all, direction="backward",
```



```

                                k=log(length(nmc$cvd)))
## Start:  AIC=10976.94
## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness + nmc$pa +
##          nmc$smoke + nmc$alc
##
##              Df Deviance    AIC
## - nmc$pa      1     10883 10967
## - nmc$smoke    2     10895 10968
## - nmc$alc      1     10886 10969
## - nmc$bmi      1     10889 10972
## <none>         10883 10977
## - nmc$fitness  1     10917 11001
## - nmc$sex      1     11097 11180
## - nmc$age      1     13043 13127
##
## Step:  AIC=10966.63
## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness + nmc$smoke +
##          nmc$alc
##
##              Df Deviance    AIC
## - nmc$smoke    2     10895 10957
## - nmc$alc      1     10886 10959
## - nmc$bmi      1     10889 10962
## <none>         10883 10967
## - nmc$fitness  1     10920 10993
## - nmc$sex      1     11098 11170
## - nmc$age      1     13045 13118
##
## Step:  AIC=10957.36
## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness + nmc$alc
##
##              Df Deviance    AIC
## - nmc$alc      1     10896 10948
## - nmc$bmi      1     10900 10952
## <none>         10895 10957
## - nmc$fitness  1     10936 10988
## - nmc$sex      1     11114 11167
## - nmc$age      1     13050 13102
##
## Step:  AIC=10948.37
## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness
##
##              Df Deviance    AIC
## - nmc$bmi      1     10902 10943
## <none>         10896 10948

```

```

## - nmc$fitness 1 10938 10979
## - nmc$sex 1 11117 11159
## - nmc$age 1 13059 13100
##
## Step: AIC=10943.43
## nmc$cvd ~ nmc$sex + nmc$age + nmc$fitness
##
## Df Deviance AIC
## <none> 10902 10943
## - nmc$fitness 1 10952 10983
## - nmc$sex 1 11120 11151
## - nmc$age 1 13073 13104

formula(backward.BIC)

## nmc$cvd ~ nmc$sex + nmc$age + nmc$fitness

summary(backward.BIC)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$fitness, family = binomial)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.6340 -0.3381 -0.1940 -0.0966 3.6228
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.771416 0.170954 -45.46 < 2e-16 ***
## nmc$sexMale 0.783860 0.053038 14.78 < 2e-16 ***
## nmc$age 0.091980 0.002398 38.35 < 2e-16 ***
## nmc$fitness -0.209655 0.029570 -7.09 1.34e-12 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 13400 on 33326 degrees of freedom
## Residual deviance: 10902 on 33323 degrees of freedom
## AIC: 10910
##
## Number of Fisher Scoring iterations: 7

```

7.2 Forward

Verifichiamo adesso le formule della procedura Forward con AIC e BIC.

7.2.1 AIC

```
#Forward: AIC
forward.AIC <- step(fit.0, scope=formula(fit.all),
                    direction="forward", k=2)

## Start:  AIC=13402.04
## nmc$cvd ~ 1
##
##           Df Deviance  AIC
## + nmc$age    1    11150 11154
## + nmc$sex     1    13073 13077
## + nmc$fitness 1    13390 13394
## + nmc$bmi     1    13393 13397
## + nmc$smoke   2    13392 13398
## <none>         13400 13402
## + nmc$pa      1    13399 13403
## + nmc$alc     1    13400 13404
##
## Step:  AIC=11153.6
## nmc$cvd ~ nmc$age
##
##           Df Deviance  AIC
## + nmc$sex    1    10952 10958
## + nmc$fitness 1    11120 11126
## + nmc$smoke   2    11128 11136
## + nmc$bmi     1    11142 11148
## + nmc$pa      1    11145 11151
## <none>         11150 11154
## + nmc$alc     1    11148 11154
##
## Step:  AIC=10957.65
## nmc$cvd ~ nmc$age + nmc$sex
##
##           Df Deviance  AIC
## + nmc$fitness 1    10902 10910
## + nmc$bmi     1    10938 10946
## + nmc$smoke   2    10938 10948
## + nmc$pa      1    10948 10956
## + nmc$alc     1    10949 10957
## <none>         10952 10958
##
```

```

## Step: AIC=10909.77
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness
##
##           Df Deviance   AIC
## + nmc$smoke 2    10892 10904
## + nmc$bmi   1    10896 10906
## <none>           10902 10910
## + nmc$alc   1    10900 10910
## + nmc$pa    1    10902 10912
##
## Step: AIC=10903.71
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke
##
##           Df Deviance   AIC
## + nmc$bmi   1    10886 10900
## + nmc$alc   1    10889 10903
## <none>           10892 10904
## + nmc$pa    1    10892 10906
##
## Step: AIC=10899.84
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke + nmc$bmi
##
##           Df Deviance   AIC
## + nmc$alc   1    10883 10899
## <none>           10886 10900
## + nmc$pa    1    10886 10902
##
## Step: AIC=10899.32
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke + nmc$bmi +
##           nmc$alc
##
##           Df Deviance   AIC
## <none>           10883 10899
## + nmc$pa    1    10883 10901

formula(forward.AIC)

## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke + nmc$bmi +
##           nmc$alc

summary(forward.AIC)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke +
##       nmc$bmi + nmc$alc, family = "binomial")

```

```
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5978  -0.3371  -0.1941  -0.0950   3.6471
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -7.462934   0.209921  -35.551 < 2e-16 ***
## nmc$age         0.092640   0.002442   37.930 < 2e-16 ***
## nmc$sexMale     0.799887   0.054643   14.638 < 2e-16 ***
## nmc$fitness    -0.183877   0.030378   -6.053 1.42e-09 ***
## nmc$smokeFormer -0.332592   0.111097   -2.994 0.002756 **
## nmc$smokeNO     -0.374525   0.106476   -3.517 0.000436 ***
## nmc$bmi         0.235857   0.096958    2.433 0.014992 *
## nmc$alc        -0.056553   0.035625   -1.587 0.112413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10883  on 33319  degrees of freedom
## AIC: 10899
##
## Number of Fisher Scoring iterations: 7
```

7.2.2 BIC

```
#Forward: BIC
forward.BIC <- step(fit.0, scope=formula(fit.all),
                    direction="forward",
                    k=log(length(nmc$cvd)))

## Start:  AIC=13410.45
## nmc$cvd ~ 1
##
##           Df Deviance   AIC
## + nmc$age    1    11150 11170
## + nmc$sex     1    13073 13094
## <none>         13400 13410
## + nmc$fitness 1    13390 13411
## + nmc$bmi     1    13393 13414
## + nmc$pa      1    13399 13420
## + nmc$alc     1    13400 13421
```

```

## + nmc$smoke      2      13392 13423
##
## Step:  AIC=11170.43
## nmc$cvd ~ nmc$age
##
##              Df Deviance   AIC
## + nmc$sex      1      10952 10983
## + nmc$fitness  1      11120 11151
## + nmc$smoke    2      11128 11169
## <none>          11150 11170
## + nmc$bmi      1      11142 11173
## + nmc$pa       1      11145 11176
## + nmc$alc      1      11148 11179
##
## Step:  AIC=10982.89
## nmc$cvd ~ nmc$age + nmc$sex
##
##              Df Deviance   AIC
## + nmc$fitness  1      10902 10943
## + nmc$bmi      1      10938 10979
## <none>          10952 10983
## + nmc$pa       1      10948 10989
## + nmc$smoke    2      10938 10990
## + nmc$alc      1      10949 10991
##
## Step:  AIC=10943.43
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness
##
##              Df Deviance   AIC
## <none>          10902 10943
## + nmc$bmi      1      10896 10948
## + nmc$alc      1      10900 10952
## + nmc$pa       1      10902 10954
## + nmc$smoke    2      10892 10954

formula(forward.BIC)

## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness

summary(forward.BIC)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness, family = "binomial")
##
## Deviance Residuals:

```

```
##      Min      1Q   Median      3Q      Max
## -1.6340 -0.3381 -0.1940 -0.0966  3.6228
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.771416   0.170954 -45.46 < 2e-16 ***
## nmc$age      0.091980   0.002398  38.35 < 2e-16 ***
## nmc$sexMale  0.783860   0.053038  14.78 < 2e-16 ***
## nmc$fitness -0.209655   0.029570  -7.09 1.34e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10902  on 33323  degrees of freedom
## AIC: 10910
##
## Number of Fisher Scoring iterations: 7
```

7.3 Both

Infine vediamo le formule della procedura Both con AIC e BIC.

7.3.1 AIC

```
#Both: AIC
both.AIC <- step(fit.0, scope=formula(fit.all),
                 direction="both",
                 k=2)

## Start:  AIC=13402.04
## nmc$cvd ~ 1
##
##              Df Deviance   AIC
## + nmc$age      1    11150 11154
## + nmc$sex      1    13073 13077
## + nmc$fitness  1    13390 13394
## + nmc$bmi      1    13393 13397
## + nmc$smoke    2    13392 13398
## <none>         13400 13402
## + nmc$pa       1    13399 13403
## + nmc$alc      1    13400 13404
##
```

```

## Step: AIC=11153.6
## nmc$cvd ~ nmc$age
##
##           Df Deviance  AIC
## + nmc$sex    1    10952 10958
## + nmc$fitness 1    11120 11126
## + nmc$smoke   2    11128 11136
## + nmc$bmi     1    11142 11148
## + nmc$pa      1    11145 11151
## <none>         11150 11154
## + nmc$alc     1    11148 11154
## - nmc$age     1    13400 13402
##
## Step: AIC=10957.65
## nmc$cvd ~ nmc$age + nmc$sex
##
##           Df Deviance  AIC
## + nmc$fitness 1    10902 10910
## + nmc$bmi     1    10938 10946
## + nmc$smoke   2    10938 10948
## + nmc$pa      1    10948 10956
## + nmc$alc     1    10949 10957
## <none>         10952 10958
## - nmc$sex     1    11150 11154
## - nmc$age     1    13073 13077
##
## Step: AIC=10909.77
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness
##
##           Df Deviance  AIC
## + nmc$smoke   2    10892 10904
## + nmc$bmi     1    10896 10906
## <none>         10902 10910
## + nmc$alc     1    10900 10910
## + nmc$pa      1    10902 10912
## - nmc$fitness 1    10952 10958
## - nmc$sex     1    11120 11126
## - nmc$age     1    13073 13079
##
## Step: AIC=10903.71
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke
##
##           Df Deviance  AIC
## + nmc$bmi     1    10886 10900
## + nmc$alc     1    10889 10903

```



```

## <none>          10892 10904
## + nmc$pa        1    10892 10906
## - nmc$smoke      2    10902 10910
## - nmc$fitness    1    10938 10948
## - nmc$sex        1    11101 11111
## - nmc$age        1    13069 13079
##
## Step:  AIC=10899.84
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke + nmc$bmi
##
##              Df Deviance   AIC
## + nmc$alc      1    10883 10899
## <none>          10886 10900
## + nmc$pa        1    10886 10902
## - nmc$bmi        1    10892 10904
## - nmc$smoke      2    10896 10906
## - nmc$fitness    1    10923 10935
## - nmc$sex        1    11098 11110
## - nmc$age        1    13056 13068
##
## Step:  AIC=10899.32
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke + nmc$bmi +
##           nmc$alc
##
##              Df Deviance   AIC
## <none>          10883 10899
## - nmc$alc      1    10886 10900
## + nmc$pa        1    10883 10901
## - nmc$bmi        1    10889 10903
## - nmc$smoke      2    10895 10907
## - nmc$fitness    1    10920 10934
## - nmc$sex        1    11098 11112
## - nmc$age        1    13045 13059

formula(both.AIC)

## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke + nmc$bmi +
##           nmc$alc

summary(both.AIC)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness + nmc$smoke +
##       nmc$bmi + nmc$alc, family = "binomial")
##

```

```
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5978 -0.3371 -0.1941 -0.0950  3.6471
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -7.462934   0.209921 -35.551 < 2e-16 ***
## nmc$age        0.092640   0.002442  37.930 < 2e-16 ***
## nmc$sexMale    0.799887   0.054643  14.638 < 2e-16 ***
## nmc$fitness   -0.183877   0.030378  -6.053 1.42e-09 ***
## nmc$smokeFormer -0.332592   0.111097  -2.994 0.002756 **
## nmc$smokeNO    -0.374525   0.106476  -3.517 0.000436 ***
## nmc$bmi        0.235857   0.096958   2.433 0.014992 *
## nmc$alc       -0.056553   0.035625  -1.587 0.112413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10883  on 33319  degrees of freedom
## AIC: 10899
##
## Number of Fisher Scoring iterations: 7
```

7.3.2 BIC

```
#Bot: BIC
both.BIC <- step(fit.0, scope=formula(fit.all),
                 direction="both",
                 k=log(length(nmc$cvd)))

## Start:  AIC=13410.45
## nmc$cvd ~ 1
##
##           Df Deviance   AIC
## + nmc$age    1    11150 11170
## + nmc$sex    1    13073 13094
## <none>              13400 13410
## + nmc$fitness 1    13390 13411
## + nmc$bmi     1    13393 13414
## + nmc$pa      1    13399 13420
## + nmc$alc     1    13400 13421
## + nmc$smoke   2    13392 13423
```

```
##
## Step: AIC=11170.43
## nmc$cvd ~ nmc$age
##
##           Df Deviance   AIC
## + nmc$sex      1    10952 10983
## + nmc$fitness  1    11120 11151
## + nmc$smoke    2    11128 11169
## <none>          1    11150 11170
## + nmc$bmi      1    11142 11173
## + nmc$pa       1    11145 11176
## + nmc$alc      1    11148 11179
## - nmc$age      1    13400 13410
##
## Step: AIC=10982.89
## nmc$cvd ~ nmc$age + nmc$sex
##
##           Df Deviance   AIC
## + nmc$fitness  1    10902 10943
## + nmc$bmi      1    10938 10979
## <none>          1    10952 10983
## + nmc$pa       1    10948 10989
## + nmc$smoke    2    10938 10990
## + nmc$alc      1    10949 10991
## - nmc$sex      1    11150 11170
## - nmc$age      1    13073 13094
##
## Step: AIC=10943.43
## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness
##
##           Df Deviance   AIC
## <none>          1    10902 10943
## + nmc$bmi      1    10896 10948
## + nmc$alc      1    10900 10952
## + nmc$pa       1    10902 10954
## + nmc$smoke    2    10892 10954
## - nmc$fitness  1    10952 10983
## - nmc$sex      1    11120 11151
## - nmc$age      1    13073 13104

formula(both.BIC)

## nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness

summary(both.BIC)

##
```

```
## Call:
## glm(formula = nmc$cvd ~ nmc$age + nmc$sex + nmc$fitness, family = "binomial")
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6340  -0.3381  -0.1940  -0.0966   3.6228
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.771416   0.170954  -45.46 < 2e-16 ***
## nmc$age      0.091980   0.002398   38.35 < 2e-16 ***
## nmc$sexMale  0.783860   0.053038   14.78 < 2e-16 ***
## nmc$fitness -0.209655   0.029570   -7.09 1.34e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 10902  on 33323  degrees of freedom
## AIC: 10910
##
## Number of Fisher Scoring iterations: 7
```

7.4 Commento

Le formule ottenute dalle tre procedure sono:

- Le procedure FORWARD, BACKWARD e BOTH AIC:
CVD ~ Age + Sex + Fitness + Smoke + BMI + Alchol
- Le procedure FORWARD, BACKWARD e BOTH BIC:
CVD ~ Age + Sex + Fitness

In questo caso, basandomi sulle procedure AIC e BIC, seleziono il risultato dalla procedura AIC ed elimino, secondo il *p-value*, la variabile non significativa Alchol, ottenendo un modello che si pone in mezzo alle due procedure AIC e BIC.

Il modello risultate da questa analisi risulta essere quello analizzato nel capitolo delle regressioni logistiche multiple.

Modello: CVD ~ Sex + Age + Fitness + Smoke + BMI.

8 Test sul Modello

Verifichiamo adesso come il modello tende a calcolare la probabilità per un campione di individui.

8.1 Age

Calcoliamo la probabilità di un ragazzo di 26 anni e di Uomo di 42 anni, entrambi fumatori (Smoke:Current), con BMI pari a 25(BMI 0) che stanno male (Fitness 2).

```
#Dato ragazzo
#Intercetta: 1, SexMale: 1, Age: 26, BMI: 0, Fitness: 2,
#Smoke:Former: 0, Smoke:NO: 0
t.age.boy <- c(1, 1, 26, 0, 2, 0, 0)

#Dato Uomo
#Intercetta: 1, SexMale: 1, Age: 42, BMI: 0, Fitness: 2,
#Smoke:Former: 0, Smoke:NO: 0
t.age.man <- c(1, 1, 42, 0, 2, 0, 0)
```

```
#Stima per il ragazzo
stima.age.boy <- exp(coef(fit)%*%t.age.boy)/
               (1+exp(coef(fit)%*%t.age.boy))
stima.age.boy

##           [,1]
## [1,] 0.008300869

#Stima per l'uomo
stima.age.man <- exp(coef(fit)%*%t.age.man)/
               (1+exp(coef(fit)%*%t.age.man))
stima.age.man

##           [,1]
## [1,] 0.03573426
```

- Probabilità per il ragazzo: $\hat{\pi} = 0.008$
- Probabilità per l'uomo: $\hat{\pi} = 0.036$

8.2 Sex

Calcoliamo la probabilità di una Donna e di un Uomo di 55 anni, ex-fumatori (Smoke:Former 1), con BMI pari a 32(BMI 1) che godono di ottima salute(Fitness 5).

```
#Dato Uomo
#Intercetta: 1, sexMale: 1, Age: 55, BMI: 0, Fitness: 5,
#Smoke:Former: 1, Smoke:NO: 0
t.sex.man <- c(1, 1, 55, 0, 5, 1, 0)

#Dato Donna
#Intercetta: 1, sexMale: 0, Age: 55, BMI: 0, Fitness: 5,
#Smoke:Former: 1, Smoke:NO: 0
t.sex.woman <- c(1, 0, 55, 0, 5, 1, 0)
```

```
#Stima per l'uomo
stima.sex.man <- exp(coef(fit)%*%t.sex.man)/
  (1+exp(coef(fit)%*%t.sex.man))

stima.sex.man

##           [,1]
## [1,] 0.04847874

#Stima per la donna
stima.sex.woman <- exp(coef(fit)%*%t.sex.woman)/
  (1+exp(coef(fit)%*%t.sex.woman))

stima.sex.woman

##           [,1]
## [1,] 0.0226795
```

- Probabilità per l'uomo: $\hat{\pi} = 0.048$
- Probabilità per la donna: $\hat{\pi} = 0.023$

8.3 Smoke

Calcoliamo la probabilità di una Donna di 36 anni, fumatrice, ex-fumatrice e non fumatrice, con BMI pari a 25(BMI 0) che è in buona salute (Fitness 4).

```
#Dato SmokeCurrent
#Intercetta: 1, Sex:0, Age: 36, BMI: 0, Fitness: 4,
#Smoke:Former: 0, Smoke:NO: 0
t.smoke.current <- c(1, 0, 36, 0, 4, 0, 0)

#Dato SmokeFormer
#Intercetta: 1, Sex:0, Age: 36, BMI: 0, Fitness: 4,
#Smoke:Former: 1, Smoke:NO: 0
t.smoke.former <- c(1, 0, 36, 0, 4, 1, 0)
```

```

#Dato SmokeNO
#Intercetta: 1, Sex:0, Age: 36, BMI: 0, Fitness: 4,
#Smoke:Former: 0, Smoke:NO: 1
t.smoke.no <- c(1, 0, 36, 0, 4, 0, 1)

#Stima SmokeCurrent
stima.smoke.current <- exp(coef(fit)%*%t.smoke.current)/
                      (1+exp(coef(fit)%*%t.smoke.current))
stima.smoke.current

##           [,1]
## [1,] 0.006613369

#Stima SmokeFormer
stima.smoke.former <- exp(coef(fit)%*%t.smoke.former)/
                      (1+exp(coef(fit)%*%t.smoke.former))
stima.smoke.former

##           [,1]
## [1,] 0.004754465

#Stima SmokeNO
stima.smoke.no <- exp(coef(fit)%*%t.smoke.no)/
                  (1+exp(coef(fit)%*%t.smoke.no))
stima.smoke.no

##           [,1]
## [1,] 0.004660303

```

- Probabilità per Fumatrice: $\hat{\pi} = 0.007$
- Probabilità per EX-Fumatrice: $\hat{\pi} = 0.005$
- Probabilità per Non Fumatrice: $\hat{\pi} = 0.005$

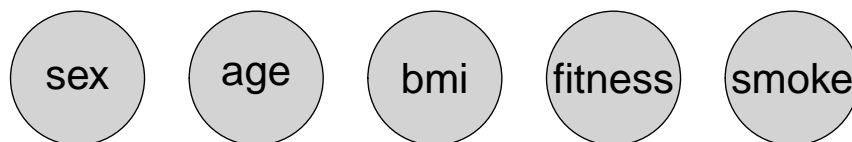
9 Visualizzazione del modello

Visualizziamo il grafo associato al modello scelto, che risulterà essere quello completamente indipendente.

```
#Modello scelto
m.fit <- dmod(~sex+age+bmi+fitness+smoke, data=nmc)
m.fit

## Model: A dModel with 5 variables
## -2logL      :      466894.73 mdim :    82 aic :      467058.73
## ideviance   :           0.00 idf  :     0 bic :      467748.69
## deviance    :      11626.84 df   :   4417

plot(m.fit)
```



10 Conclusioni

In conclusione, il modello scelto che più si adatta meglio al problema per il calcolo della probabilità di un problema cardiovascolare è:

Modello: $\sim \text{CVD} \sim \text{Sex} + \text{Age} + \text{BMI} + \text{Fitness} + \text{Smoke}$.

All'interno del modello sono presenti unicamente variabili significative e indipendenti.

E' stato scelto questo modello a differenza di altri modelli (come il modello con interazioni) perchè rappresenta al meglio lo studio svolto sulle varie categorie e contemporaneamente risulta essere un modello abbastanza semplice e significativo.