NMC - Foundations of Statistical Modelling

Lorenzo Baiardi

19 Aprile 2023

Indice

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")
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  25.6 22.9 22.9 22 24.4 ...
##  $ 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 ≥ 30
- FITNESS: statico di salute 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 nel consumo di alchol dell'individuo.
  1. Never
  2. Low
  3. Medium
  4. High

Per una maggiore comprensione del problema, convertiremo alcune variabili di tipo categoriale in variabili di tipo ordinale per 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
#Smoke: 1-NO, 2-FORMER, 3-CURRENT
#BMI: 0-<30, 1->=30

c.fit = c('Much Worse', 'Little Worse', 'Just as good',
            'A bit better', 'Much better')
c.alc =  c('Never', 'Low', 'Medium', 'High')
c.smoke<- c('NO', 'Former', 'Current')

#BMI dicotomizzata
bmi = nmc$bmi
nmc$bmi = as.numeric(nmc$bmi>=30)
#Variabili ordinali
```

```r
fitness <- nmc$fitness
nmc$fitness = as.numeric(ordered(nmc$fitness, c.fit))
nmc$alc = as.numeric(ordered(nmc$alc, c.alc))
smoke.ord <- as.numeric(ordered(nmc$smoke, c.smoke))

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

```r
#Tabella delle Frequenze del Dataset
ftable(sex+bmi+pa ~ cvd+smoke+alc+fitness, nmc)

##                          sex Female                Male
##                          bmi      0         1        0         1
##                          pa       0    1    0    1    0    1    0    1
## cvd smoke   alc fitness
## 0   Current 1   1                4    1    1    0    2    1    0    1
##                 2               11    6    6    0    2    2    2    0
##                 3               30    1    3    0    5    1    1    0
##                 4                6    1    0    0    3    0    0    0
##                 5                5    0    0    0    4    0    0    0
##             2   1               25    8   12    3    6    2    4    2
##                 2              163   43   25    9   25    4    2    1
##                 3              438   36   21    1   84    9    5    1
##                 4              188    4    8    1   48    1    1    0
##                 5               52    0    2    0   18    0    0    0
##             3   1                9    8    6    2   10    6    2    1
##                 2               72   20   16    3   46   15    1    2
##                 3              279   30   14    3  139   24   11    0
##                 4              198    8    5    0  115    4    3    0
##                 5               48    0    2    0   28    0    0    0
##             4   1                0    0    0    1    0    2    0    0
##                 2                4    1    0    0    5    0    1    0
##                 3               11    3    1    0   13    0    1    0
```

```
##                  4    9  1  1  0  4  0  1  0
##                  5    7  0  0  0  5  1  0  0
##      Former  1   1    2  4  5  3  2  2  0  1
##                  2   16  5 10  2  5  3  0  1
##                  3   79 14 17  1 31  1  6  0
##                  4   61  0  8  0 27  4  3  0
##                  5   34  0  0  0 12  0  0  0
##              2   1   34 17 16 10  9  7  8  4
##                  2  180 40 66 12 37 16 12  2
##                  3  982 67 122 8 243 37 26 5
##                  4  777 23 28  0 302 13 14 0
##                  5  282  1  3  1 122  0  4  1
##              3   1   10  5  9  7  8  5  6  3
##                  2  128 26 25  8 82 32 27 10
##                  3  830 80 56  5 505 53 45 9
##                  4  802 14 22  2 668 23 30 2
##                  5  276  2  3  0 290  2  4  0
##              4   1    3  0  0  0  2  1  0  0
##                  2    4  0  0  0  2  2  0  0
##                  3   38  2  3  0 41  6  5  1
##                  4   28  2  1  0 54  2  5  0
##                  5   12  0  0  0 21  0  0  0
##      NO      1   1   26 11 19 10 10  3  1  0
##                  2  203 36 42 10 63 21 10  1
##                  3  974 77 100 8 244 30 12 1
##                  4  657 19 41  0 336 18 5  1
##                  5  237  1  9  0 180  4  2  0
##              2   1   79 26 37 13 29  7  9  6
##                  2  600 129 133 18 183 52 23 13
##                  3 3073 254 231 16 755 87 46 4
##                  4 2467 47 78  3 991 34 19 1
##                  5  842  6 20  0 603 12  3  1
##              3   1   24  9 14  4 19 10  7  3
##                  2  202 42 38  9 120 47 20 5
##                  3 1254 105 76 11 671 94 36 8
##                  4 1281 29 32  2 1003 33 24 0
##                  5  424  3  2  0 526  6  5  0
##              4   1    2  0  0  0  3  2  0  0
##                  2    1  2  2  0  8  4  4  0
##                  3   30  5  1  0 41  5  4  0
##                  4   43  0  1  0 49  5  3  0
##                  5   12  0  0  0 42  0  1  0
## 1    Current 1   1    0  0  0  0  0  0  0  0
##                  2    0  0  0  0  1  0  0  0
##                  3    1  0  0  0  3  1  0  0
```

```
##                  4                 0    0    0    0    0    0    0    0
##                  5                 0    0    0    0    0    0    0    0
##              2   1                 1    0    0    0    0    1    0    0
##                  2                 3    1    0    0    1    2    0    0
##                  3                11    1    0    0   10    1    0    0
##                  4                 5    0    0    0    4    0    0    0
##                  5                 3    0    0    0    0    1    0    0
##              3   1                 2    0    0    0    0    0    0    0
##                  2                 2    1    0    0    3    0    1    0
##                  3                 8    2    0    0   17    0    0    0
##                  4                 2    0    0    0   14    0    0    0
##                  5                 1    0    0    0    4    0    0    0
##              4   1                 0    0    0    1    0    0    0    0
##                  2                 0    0    0    0    0    0    0    0
##                  3                 1    0    0    0    4    0    0    0
##                  4                 0    0    0    0    3    0    0    0
##                  5                 1    0    0    0    1    0    0    0
##     Former   1   1                 0    0    1    0    0    0    0    0
##                  2                 0    0    0    0    2    1    0    1
##                  3                 4    0    0    0   11    0    0    0
##                  4                 2    0    0    0    5    1    0    0
##                  5                 1    0    0    0    3    0    0    0
##              2   1                 1    1    1    2    0    1    0    0
##                  2                 5    1    4    0    2    0    0    1
##                  3                23    3    2    0   36    3    1    0
##                  4                22    1    5    0   46    1    2    0
##                  5                 8    0    0    0   18    0    0    0
##              3   1                 1    2    0    1    1    0    0    2
##                  2                 4    0    2    0   12    2    3    4
##                  3                16    1    1    1   58    4    2    2
##                  4                27    1    0    0   72    1    4    0
##                  5                 5    0    0    0   18    0    2    0
##              4   1                 0    0    0    0    0    0    0    0
##                  2                 0    0    0    0    1    0    0    0
##                  3                 1    0    0    0    4    1    1    1
##                  4                 0    0    0    0    6    0    0    0
##                  5                 2    0    0    0    3    0    0    0
##     NO       1   1                 4    2    2    0    0    0    0    1
##                  2                 6    2    7    0    5    1    0    0
##                  3                50    6   11    1   28    0    3    0
##                  4                47    0    3    0   50    3    1    0
##                  5                18    0    0    0   14    0    0    0
##              2   1                 2    2    1    0    0    1    0    0
##                  2                24    3    6    2    9    3    3    1
##                  3                96    4   15    2   50    5    2    0
```

```
##                 4  82  2  2  0  68  2  0  0
##                 5  34  0  1  0  25  0  0  0
##           3     1   0  0  0  0   2  0  0  0
##                 2  10  0  3  0   4  1  3  0
##                 3  34  4  6  0  44  7  4  1
##                 4  45  0  4  0  90  2  5  0
##                 5  17  1  2  0  51  0  1  0
##           4     1   1  0  0  0   0  0  0  0
##                 2   0  0  0  0   1  1  0  0
##                 3   5  0  1  0   4  0  1  0
##                 4   5  0  0  0   4  1  1  0
##                 5   2  0  0  0   5  0  0  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.

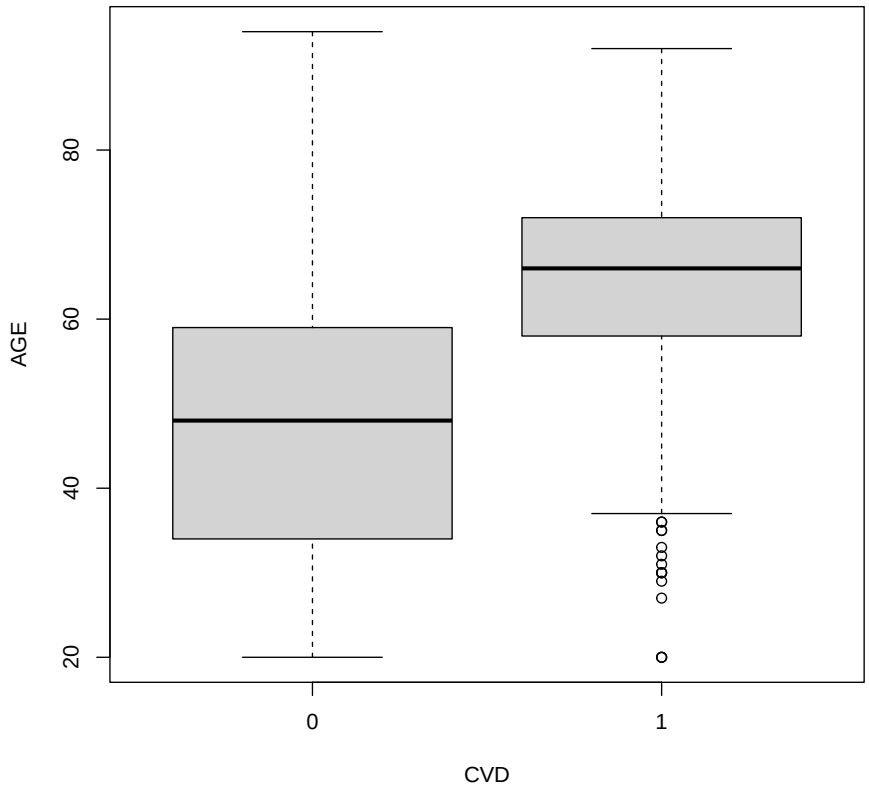
3.1 Age

```
#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: Age ~ 0.092.
- La variabile Age è molto significativa secondo il *p-value*.

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

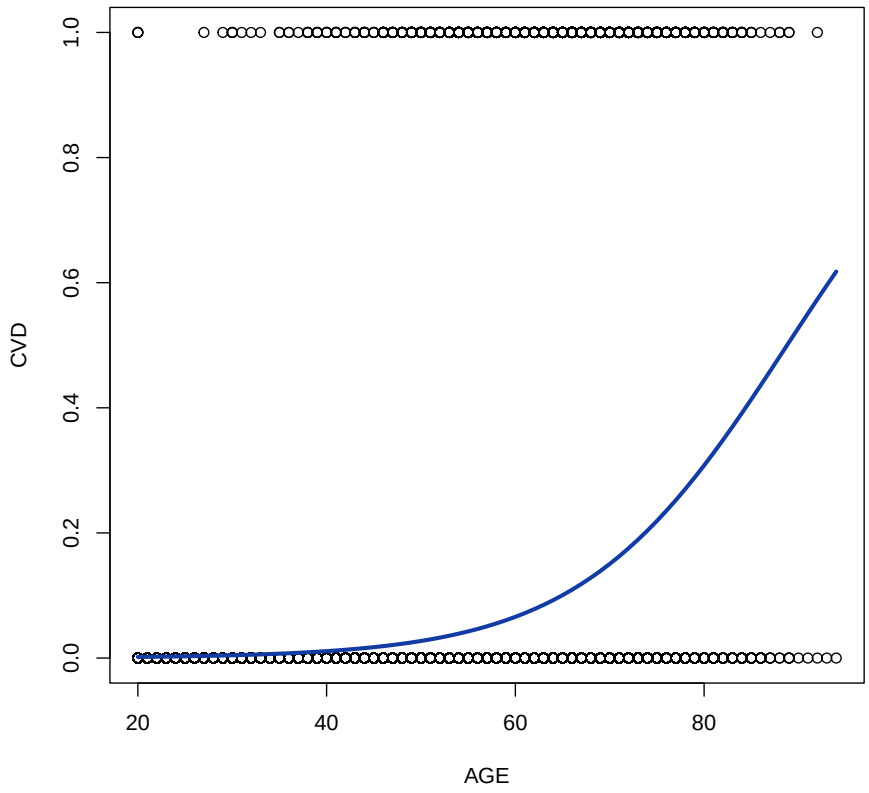
```r
#Boxplot
boxplot(nmc$age~nmc$cvd, xlab="CVD", ylab="AGE")
```



- 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.

```r
#Plot
pstima.age <- fit.age$fitted.values
plot(nmc$age, nmc$cvd, xlab="AGE", ylab="CVD")
lines(sort(nmc$age), pstima.age[order(nmc$age)], lwd=3, col="#123ba3")
```



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

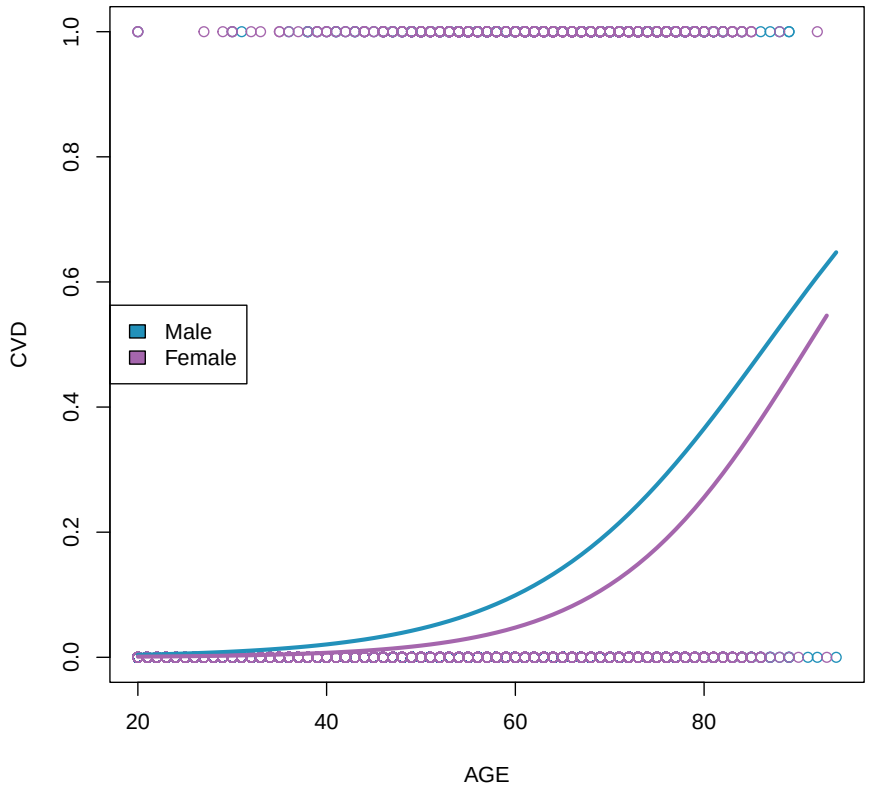
```
#Regressioni logistiche semplici
#Sex
fit.sex <- glm(nmc$cvd ~ nmc$sex, family=binomial)
summary(fit.sex)

##
## 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: SEX:MALE ~ 0.910.
- La variabile SEX risulta molto 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)
pstima.sex.male <- fit.sex.male$fitted.values
#Modello per Femmina
```

```r
fit.sex.female <- glm(nmc$cvd[nmc$sex=="Female"] ~
                        nmc$age[nmc$sex=="Female"],
                        family=binomial)
pstima.sex.female <- fit.sex.female$fitted.values
#Plot
plot(nmc$age[nmc$sex=="Male"], nmc$cvd[nmc$sex=="Male"],
     xlab="AGE", ylab="CVD", col="#2291ba")
points(nmc$age[nmc$sex=="Female"], nmc$cvd[nmc$sex=="Female"],
       col="#a566ad")
lines(nmc$age[nmc$sex=="Male"],pstima.sex.male,lwd=3,col="#2291ba")
lines(nmc$age[nmc$sex=="Female"],pstima.sex.female,lwd=3,col="#a566ad")
legend(x="left",legend=c("Male","Female"),fill=c("#2291ba","#a566ad"))
```



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

3.3 BMI

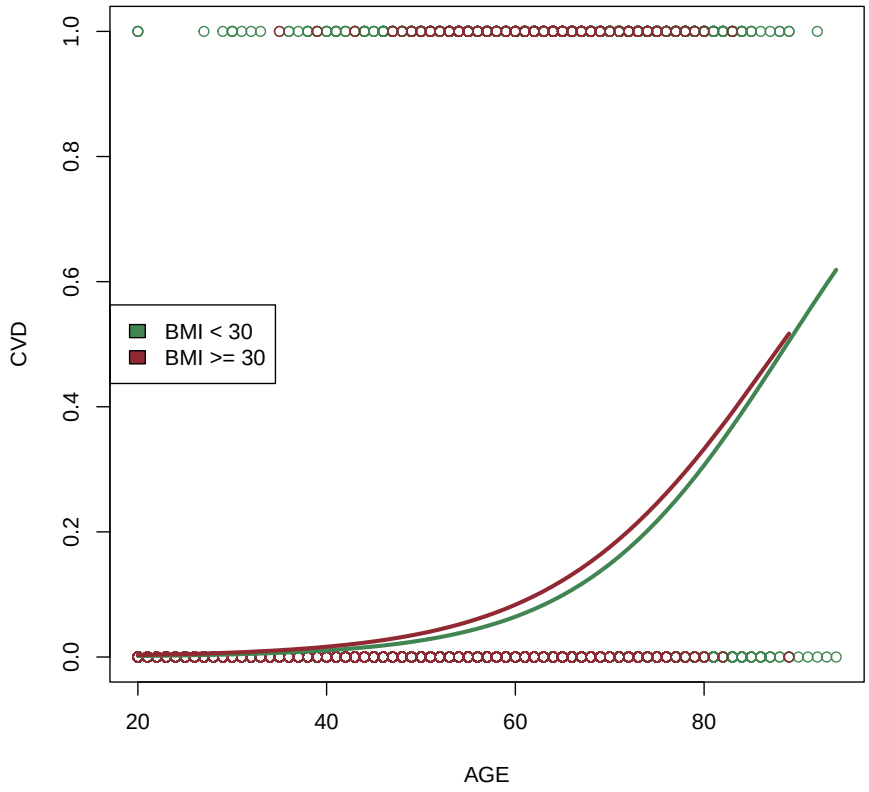
```
#BMI
fit.bmi <- glm(nmc$cvd ~ nmc$bmi, family=binomial)
summary(fit.bmi)

##
## 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 ~ 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)
pstima.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)
pstima.bmi.1 <- fit.bmi.1$fitted.values
```

```r
#Plot
plot(nmc$age[nmc$bmi==0], nmc$cvd[nmc$bmi==0],
     xlab="AGE", ylab="CVD", col="#408552")
points(nmc$age[nmc$bmi==1], nmc$cvd[nmc$bmi==1], col="#912933")
lines(nmc$age[nmc$bmi==0], pstima.bmi.0, lwd=3, col="#408552")
lines(nmc$age[nmc$bmi==1], pstima.bmi.1, lwd=3, col="#912933")
legend(x="left", legend=c("BMI < 30", "BMI >= 30"),
       fill=c("#408552","#912933"))
```



Le due curve sono molto simili tra di loro, 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 e significativa 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: Categoriale
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
```

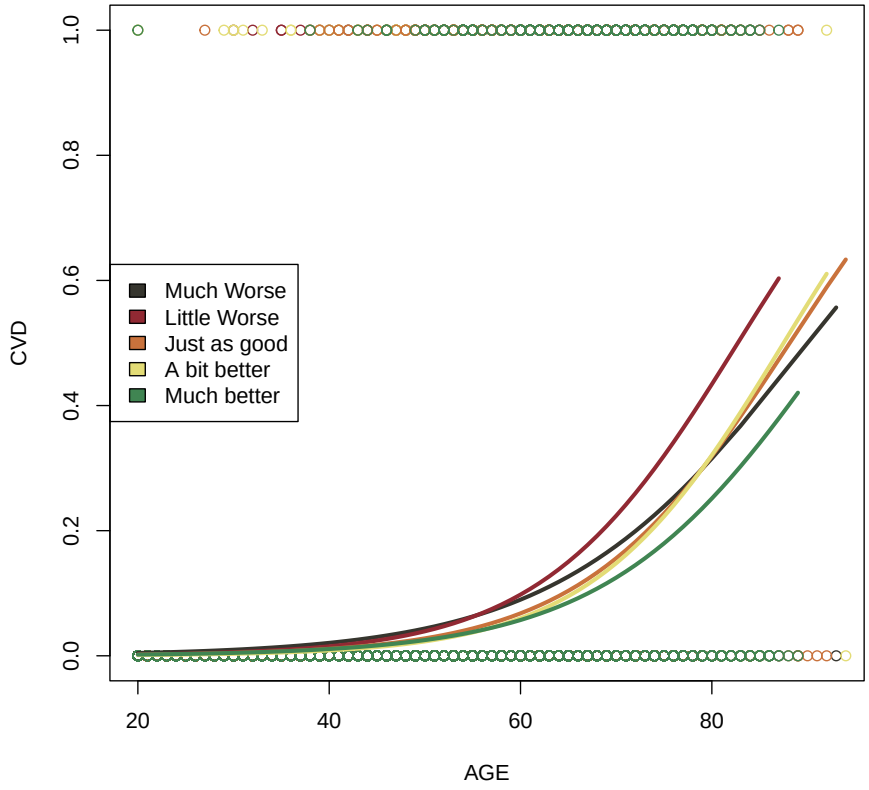
```
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.81935    0.04066 -69.344  < 2e-16 ***
## fitnessJust as good -0.20312    0.05783  -3.512 0.000444 ***
## fitnessLittle Worse -0.23313    0.09169  -2.542 0.011009 *
## fitnessMuch better  -0.03049    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:JUSTASGOOD e FITNESS:LITTLEWORSE risultano significative.

Visualizziamo il comportamento della variabile FITNESS all'aumentare dell'età.

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

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

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

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

#Plot
plot(nmc$age[fitness=="Much Worse"],nmc$cvd[fitness=="Much Worse"],
     xlab="AGE", ylab="CVD", col="#36352e")
points(nmc$age[fitness=="Little Worse"],
       nmc$cvd[fitness=="Little Worse"], col="#912933")
points(nmc$age[fitness=="Just as good"],
       nmc$cvd[fitness=="Just as good"], col="#e3dc76")
points(nmc$age[fitness=="A bit better"],
       nmc$cvd[fitness=="A bit better"], col="#c9723c")
points(nmc$age[fitness=="Much better"],
       nmc$cvd[fitness=="Much better"], col="#408552")
lines(nmc$age[fitness=="Much Worse"], pstima.fitness.muchworse,
      lwd=3, col="#36352e")
lines(nmc$age[fitness=="Little Worse"], pstima.fitness.littleworse,
      lwd=3, col="#912933")
lines(nmc$age[fitness=="Just as good"], pstima.fitness.justasgood,
      lwd=3, col="#c9723c")
lines(nmc$age[fitness=="A bit better"], pstima.fitness.abitbetter,
      lwd=3, col="#e3dc76")
lines(nmc$age[fitness=="Much better"], pstima.fitness.muchbetter,
      lwd=3, col="#408552")
legend(x="left",
       legend=c("Much Worse", "Little Worse", "Just as good",
                "A bit better", "Much better"),
       fill=c("#36352e","#912933","#c9723c","#e3dc76","#408552"))
```



Attraverso il grafico notiamo che la categoria FITNESS:MUCHBETTER è quella meno soggetta rispetto a tutte le altre. Viceversa la categoria FITNESS:LITTLEWORSE ha più probabilità di incorrere in un problema cardiovascolare.

Chi è della categoria FITNESS:MUCHWORSE ha meno probabilità rispetto alla categoria FITNESS:LITTLEWORSE 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 FITNESS come significativa.

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
```

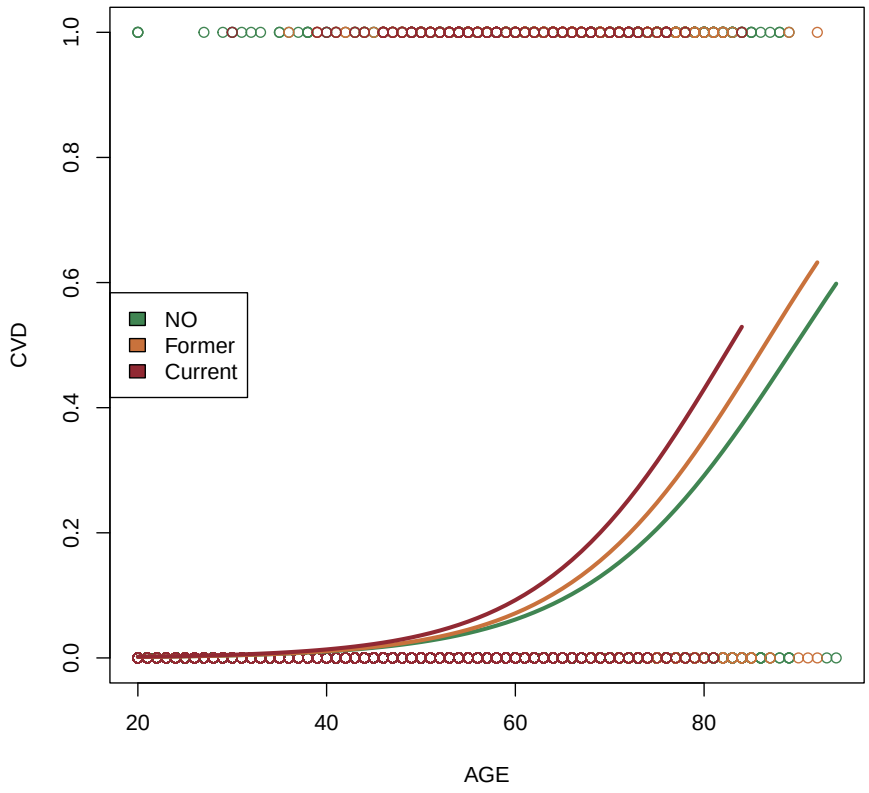
```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13400  on 33326  degrees of freedom
## Residual deviance: 13400  on 33325  degrees of freedom
## AIC: 13404
##
## Number of Fisher Scoring iterations: 5
```

- 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)
pstima.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)
pstima.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)
pstima.smoke.current <- fit.smoke.current$fitted.values

#Plot
plot(nmc$age[nmc$smoke=="NO"], nmc$cvd[nmc$smoke=="NO"],
     xlab="AGE", ylab="CVD", col="#408552")
points(nmc$age[nmc$smoke=="Former"], nmc$cvd[nmc$smoke=="Former"],
       col="#c9723c")
points(nmc$age[nmc$smoke=="Current"], nmc$cvd[nmc$smoke=="Current"],
       col="#912933")
lines(nmc$age[nmc$smoke=="NO"], pstima.smoke.no,
      lwd=3, col="#408552")
lines(nmc$age[nmc$smoke=="Former"], pstima.smoke.former,
```

```
      lwd=3, col="#c9723c")
lines(nmc$age[nmc$smoke=="Current"], pstima.smoke.current,
      lwd=3, col="#912933")
legend(x="left", legend=c("NO", "Former", "Current"),
       fill=c("#408552","#c9723c", "#912933"))
```



Possiamo notare come 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.

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.
- Un 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, Alchol
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 + ALCHOL

- 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 ALCHOL.
- 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 $\sim$ 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 $\sim$ -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, Alchol
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
```

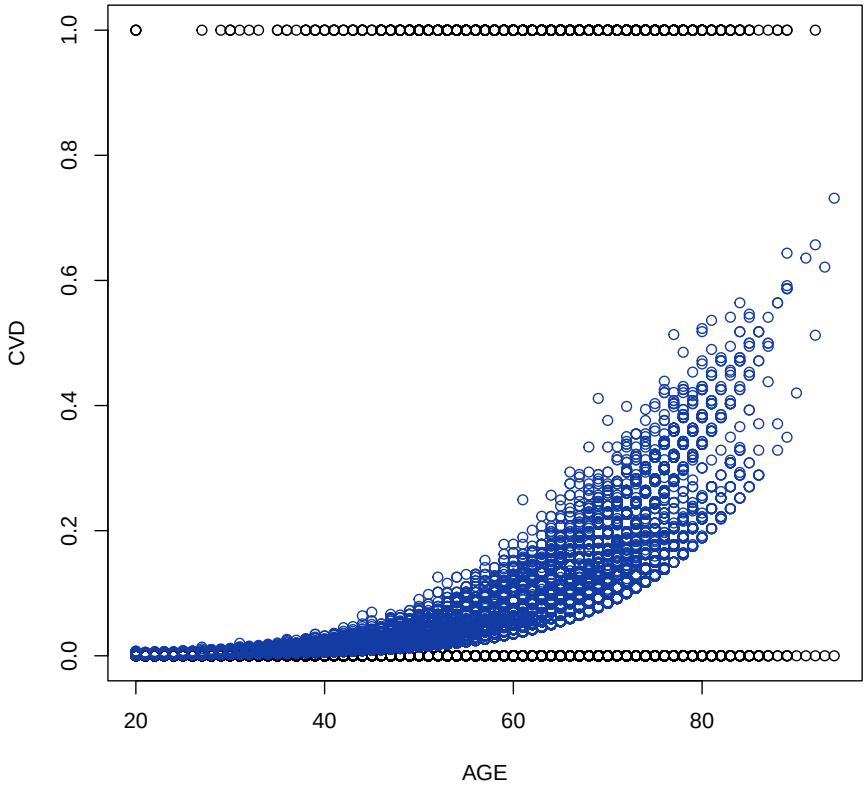
Il modello risultate è:
Modello: CVD $\sim$ SEX + AGE + BMI + FITNES + SMOKE

- Le variabili risultato 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="#123ba3")
```



4.3 Commento

- Il modello risulta essere:
  Modello: CVD$\sim$ SEX + AGE + BMI + FITNESS + SMOKE
- Come visto nelle regressioni logistiche semplici, le variabili SEX:MALE, AGE e BMI continuano ad influenzare positivamente la comparsa di problemi cardiovascolari.
- Al contrario, le variabili significative FITNESS, SMOKE:FORMER e SMOKE:NO riducono la possibilità di avere un CVD.
- Di conseguenza la categoria SMOKE:CURRENT ha una probabilità maggiore nell'insorgenza di CVD.

5 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.

5.1 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.

5.2 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 mostra un'interazione significativa, variando il valore stimato e diminuendo la significatività della variabile BMI. In questo caso la variabile BMI assume valore stimato negativo, influendo negativamente nella comparsa di CVD.

5.3 Alchol e BMI

Come per il caso di SMOKE, verifichiamo se il consumo di ALCHOL associato ad un maggior indice di massa corporea influisca 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 ALCHOL e BMI non è supportata.

5.4 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.

5.5 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.

5.6 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.

5.7 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.

5.8 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
```

5.9 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 (SMOKE x BMI) non risultando veritiero dato che esistono molti studi che affermano come l'utilizzo della sigaretta possa far rallentare il metabolismo.

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.

Decido quindi di non considerare questo modello perché non fornisce alcun contributo decisivo per il nostro problema, andando contro anche alle analisi che fino a qui abbiamo valutato.

6 Selezione del Modello

Utilizziamo adesso i metodi Backward, Forward e Both basati sui criteri di penalizzazione AIC e BIC per un'ulteriore 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.

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

6.1 Backward

Verifichiamo le formule della procedura Backward con AIC e BIC.

6.1.1 AIC

```
#Backward: AIC
backward.AIC <- step(fit.all, direction="backward",
                     k=2, trace=FALSE)
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
```

6.1.2 BIC

```
#Backward: BIC
backward.BIC <- step(fit.all, direction="backward",
                     k=log(length(nmc$cvd)), trace=FALSE)
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
```

6.2 Forward

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

6.2.1 AIC

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

6.2.2 BIC

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

6.3 Both

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

6.3.1 AIC

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

6.3.2 BIC

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

6.4 Commento

Le formule ottenute dalle tre procedure sono:

- Le procedure FORWARD, BACKWARD e BOTH AIC:
  CVD $\sim$ AGE + SEX + FITNESS + SMOKE + BMI + ALCHOL
- Le procedure FORWARD, BACKWARD e BOTH BIC:
  CVD $\sim$ AGE + SEX + FITNESS

7 Grafi non orientati

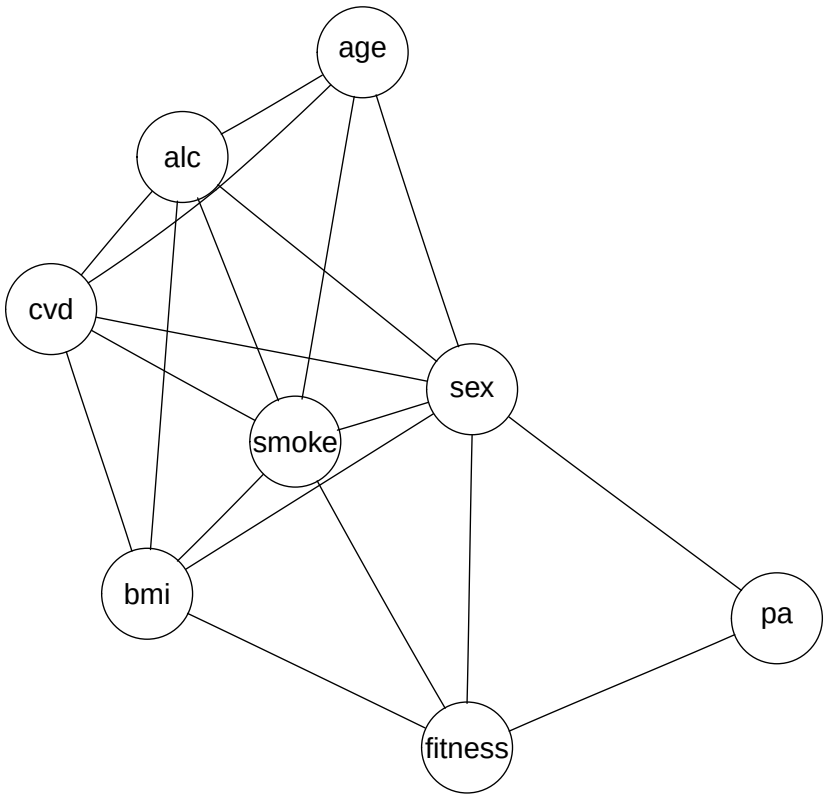
Analizziamo adesso lo spazio dei modelli da un punto di vista grafico attraverso la visualizzazione di grafi associati al Dataset. Utilizzeremo anche qui procedure di Backward e Forward con metodi di penalizzazione AIC e BIC.

```
#Formula modello saturo e indipendente
sat <- dmod(~.^., data=nmc)
ind <- dmod(~.^1, data=nmc)
```

7.1 Backward

7.1.1 AIC

```
#Backward:AIC
m.aic.backward <- stepwise(sat, direction="backward")
m.aic.backward

## Model: A dModel with 8 variables
##  -2logL    :      548496.14 mdim : 3705 aic :    555906.14
##  ideviance :       19336.00 idf  : 3618 bic :    587080.46
##  deviance  :       16793.15 df   : 68294

plot(as(m.aic.backward, "graphNEL"), "fdp")
```

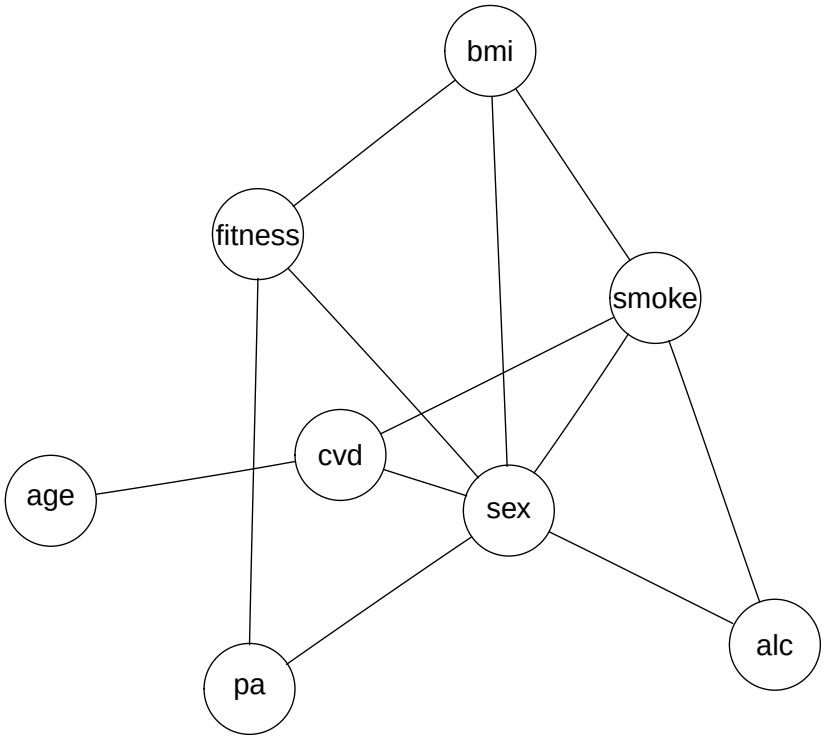


In questo primo grafo, la variabile di risposta CVD risulta essere direttamente connessa con le variabili BMI, SMOKE, AGE, ALC e SEX mentre risulta indipendente dalle variabili FITNESS e PA condizionatamente alle altre.

7.1.2 BIC

```
#Backward:BIC
m.bic.backward <- stepwise(sat, k=log(length(nmc$cvd)),
                  direction="backward")
m.bic.backward

## Model: A dModel with 8 variables
##  -2logL    :      557579.16 mdim :  209 aic :    557997.16
##  ideviance :       10252.97 idf  :  122 bic :    559755.72
##  deviance  :       25876.18 df   : 71790
```

```
plot(as(m.bic.backward, "graphNEL"), "fdp")
```



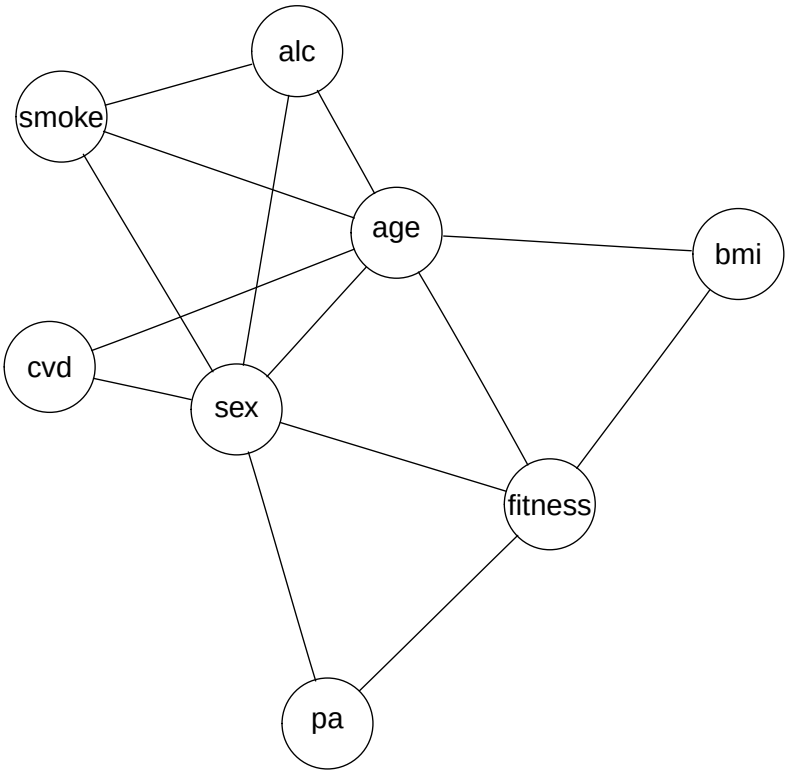
Con il criterio BIC invece la variabile CVD rimane direttamente connesse con le variabili SMOKE, SEX e AGE e non direttamente connesse con le altre.

7.2 Forward

7.2.1 AIC

```
#AIC Forward
m.aic.forward <- stepwise(ind, direction="forward")
m.aic.forward

## Model: A dModel with 8 variables
##  -2logL    :      547503.08 mdim : 2934 aic :    553371.08
```

```
##  ideviance :       20329.06 idf  : 2847 bic :    578058.12
##  deviance  :       15800.09 df   : 69065

plot(as(m.aic.forward, "graphNEL"), "fdp")
```



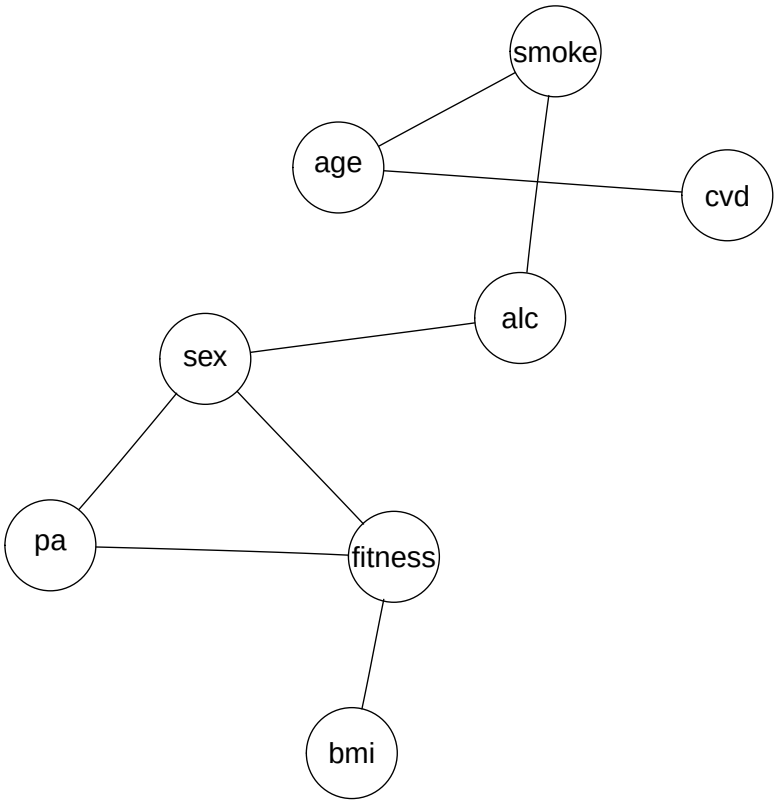
Nella procedura AIC con criterio di selezione AIC, la variabile CVD è connessa con le sole variabili SEX e AGE.

7.2.2 BIC

```
#BIC Forward
m.bic.forward <- stepwise(ind, k=log(length(nmc$cvd)),
                  direction="forward")
m.bic.forward

## Model: A dModel with 8 variables
```

```
##  -2logL    :      555553.98 mdim :  335 aic :    556223.98
##  ideviance :       12278.15 idf  :  248 bic :    559042.71
##  deviance  :       23851.00 df   : 71664
```

```
plot(as(m.bic.forward, "graphNEL"), "fdp")
```



Con il criterio di selezione BIC, la variabile CVD è direttamente connessa solo con la variabile AGE.

7.3 Commento

- In tutte le procedure, la variabile di risposta CVD risulta sempre direttamente connessa con la variabile AGE e in modo molto forte con la variabile SEX.
- In tutte le procedure, le variabili PA e FITNESS risultano direttamente connesse e legate alla variabile SEX.

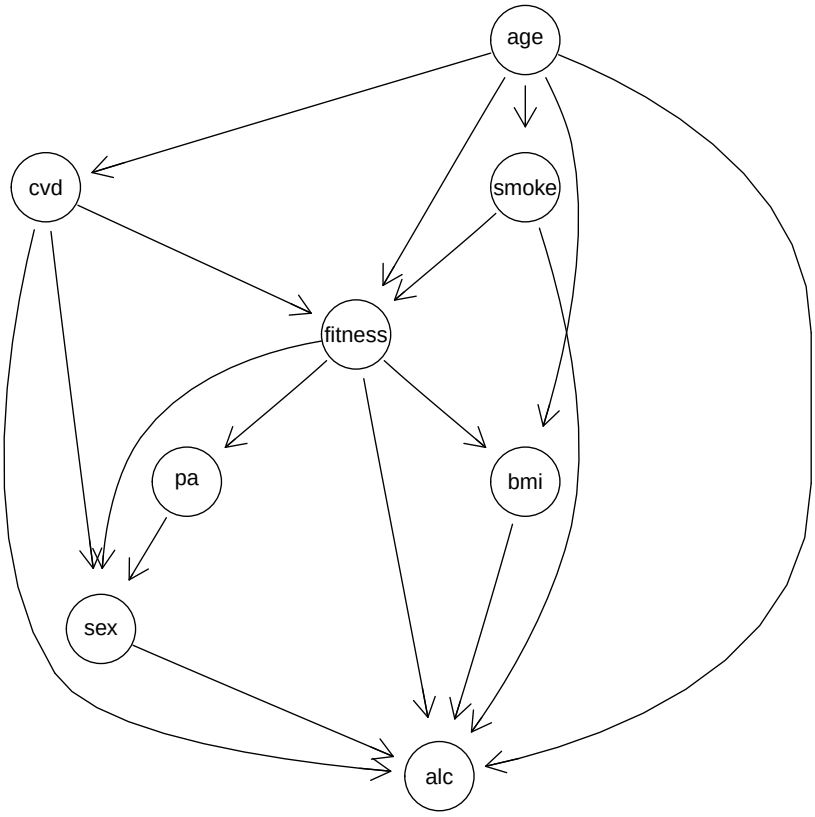
8 Reti Bayesiane

Prima di poter individuare una prima rete bayesiana dobbiamo adattare i numeri affinché la funzione hc possa essere eseguita.

```
#Adattamento del dataset
nmc.bn <- nmc
nmc.bn$sex = as.numeric(nmc.bn$sex == "Male")
nmc.bn$age = as.numeric(nmc.bn$age)
nmc.bn$cvd = as.numeric(nmc.bn$cvd)
nmc.bn$pa = as.numeric(nmc.bn$pa)
nmc.bn$smoke = smoke.ord
str(nmc.bn)

## 'data.frame': 33327 obs. of  8 variables:
##  $ sex    : num  1 0 1 0 1 0 1 1 1 1 ...
##  $ age    : num  94 93 92 92 91 90 89 89 89 89 ...
##  $ bmi    : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ cvd    : num  0 0 0 1 0 0 0 1 0 1 ...
##  $ fitness: num  3 1 4 3 4 4 4 4 4 4 ...
##  $ pa     : num  0 1 1 0 0 0 0 0 0 0 ...
##  $ smoke  : num  1 1 2 2 2 2 1 2 1 1 ...
##  $ alc    : num  3 2 1 1 3 2 1 3 3 1 ...
```

Visualizziamo la prima rete bayesiana tramite la funzione hc.

```
#Rete Bayesiana
bn <- hc(nmc.bn)
plot(as(amat(bn), "graphNEL"))
```



Questa prima rete mostra delle dipendenze non realistiche, come ad esempio l'influenza che ha il FITNESS e il PA (Attività Fisica) nella determinazione del SEX. Per questo motivo dobbiamo dare un ordinamento alle variabili permettendo di non avere incoerenze tra i vari archi.

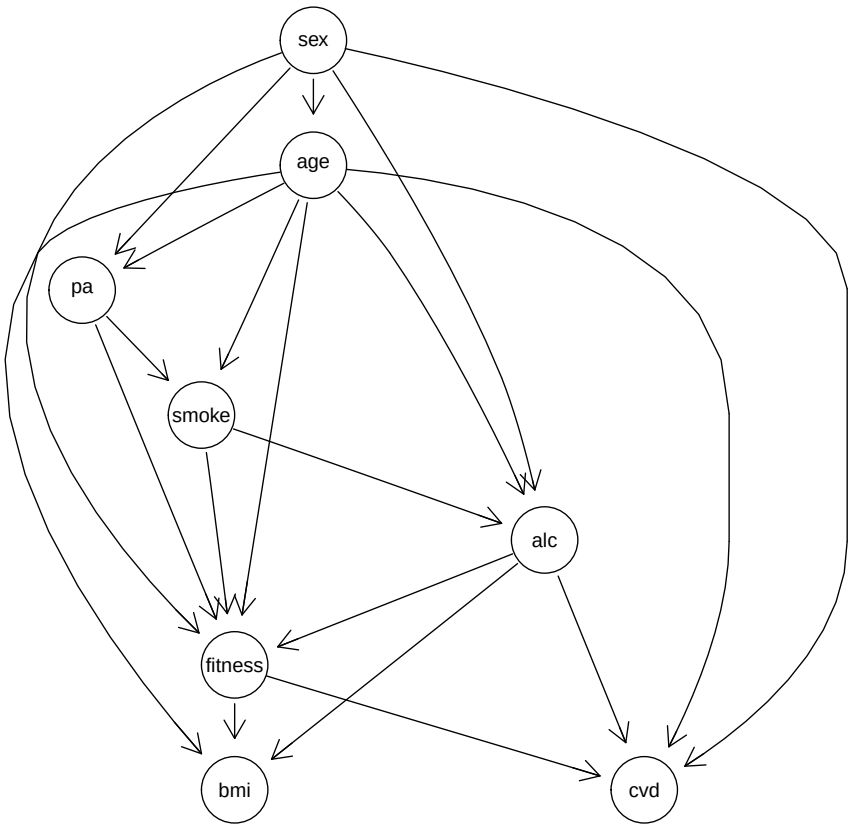
8.1 Ordinamento delle Variabili

L'ordinamento che andrò ad utilizzare sarà:

- Variabili di background: SEX, AGE
- Attività che influenzano il CVD: ALCHOL, SMOKE, PA
- Condizione fisica del paziente: BMI, FITNESS
- CVD

```r
#Ordinamento delle variabili
#1-SEX, 1-AGE, 3-BMI, 4-CVD, 3-FITNESS, 2-PA, 2-SMOKE, 2-ALC
block<-c(1, 1, 3, 4, 3, 2, 2, 2)
blnmc.bn <- matrix(0, nrow=8, ncol=8)
rownames(blnmc.bn) <- colnames(blnmc.bn) <- names(nmc.bn)
for (b in 2:4) blnmc.bn[block==b, block<b] <- 1
blackL <- data.frame(get.edgelist(as(blnmc.bn, "igraph")))
names(blackL) <- c("from", "to")
```

```r
#Rete Bayesiana con ordinamento
bn.o <- hc(nmc.bn, blacklist=blackL)
plot(as(amat(bn.o), "graphNEL"))
```

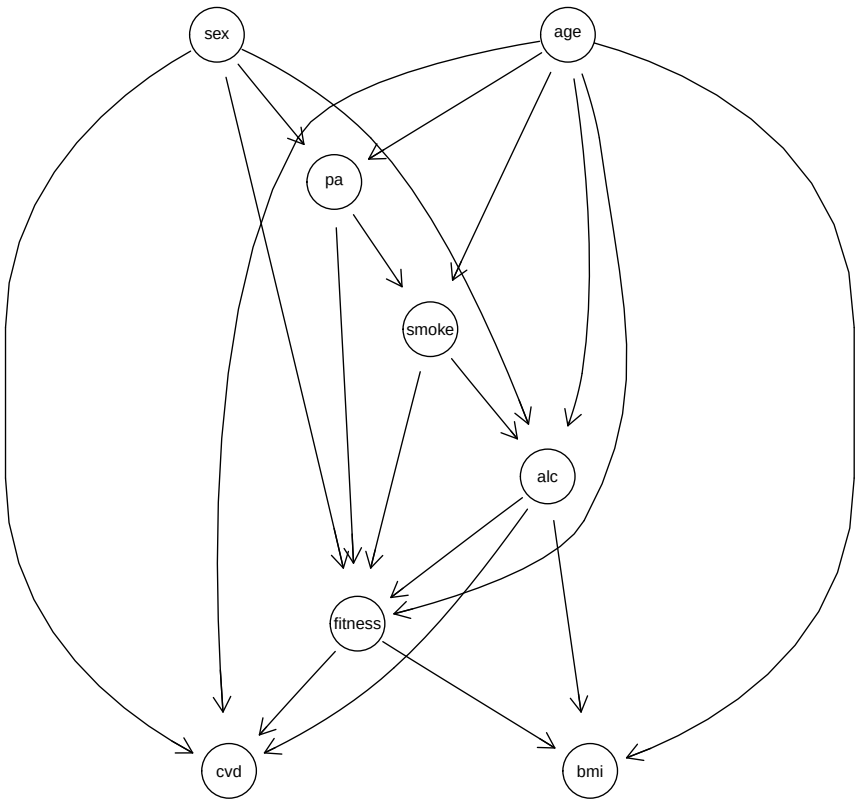


Anche in questo caso la rete risultante mostra un incongruenza nell'arco tra SEX e AGE (il Sesso non può essere condizionata dall'età della persona).

Rimuoviamo allora l'arco e rieseguiamo la funzione hc.

```
#Rimozione arco tra SEX e AGE
block<-c(1, 1, 3, 4, 3, 2, 2, 2)
blnmc.bn <- matrix(0, nrow=8, ncol=8)
rownames(blnmc.bn) <- colnames(blnmc.bn) <- names(nmc.bn)
for (b in 2:4) blnmc.bn[block==b, block<b] <- 1
blnmc.bn[1,2] = 1
blnmc.bn[2,1] = 1
blackL <- data.frame(get.edgelist(as(blnmc.bn, "igraph")))
names(blackL) <- c("from", "to")
```

```
#Bayesian Network finale
m.bn <- hc(nmc.bn, blacklist=blackL)
plot(as(amat(m.bn), "graphNEL"))
```



In questo ultimo grafo, possiamo notare come ci sia una relazione diretta tra le variabili AGE, SEX, FITNESS e ALCHOL. Viceversa la variabile di risposta CVD risulta indipendente dalle variabili SMOKE, PA e BMI.

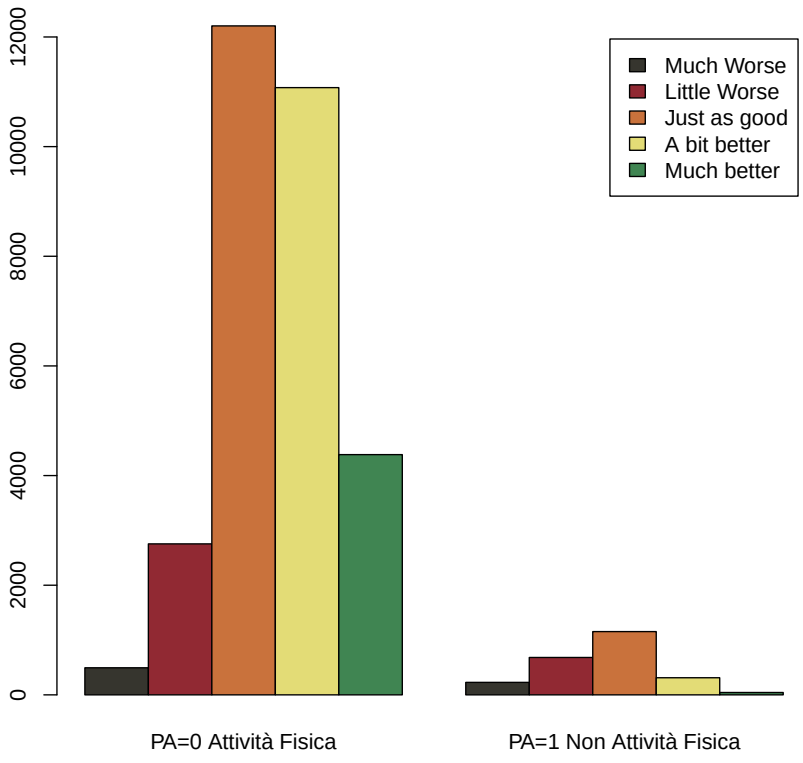
9 Considerazioni sul Modello

Effuiamo qualche considerazione su un possibile modello finale.

9.1 Fitness e PA

Durante l'analisi abbiamo visto come le variabili FITNESS e PA fossero direttamente connessa tra di loro. Valutiamo questa connessione tramite un Barplot per la visualizzazione tra le varie categorie di FITNESS e la loro attività fisica.

```
#Barplot Fitness e PA
barplot(table(nmc$fitness, nmc$pa),
        names.arg=c("PA=0 Attività Fisica", "PA=1 Non Attività Fisica"),
        legend.text=c("Much Worse", "Little Worse","Just as good",
                "A bit better", "Much better"),
        col=c("#36352e","#912933", "#c9723c", "#e3dc76", "#408552"), beside=TRUE)
```



Come possiamo vedere, la variabile FITNESS e PA risultano connesse anche all'interno dell'istogramma, mostrandoci come l'attività fisica induca maggiormente ad una miglior condizione di salute rispetto a chi non la pratica.

Dicotomizzo la variabile FITNESS per effettuare una regressione e valutare meglio all'interno di un Barplot.

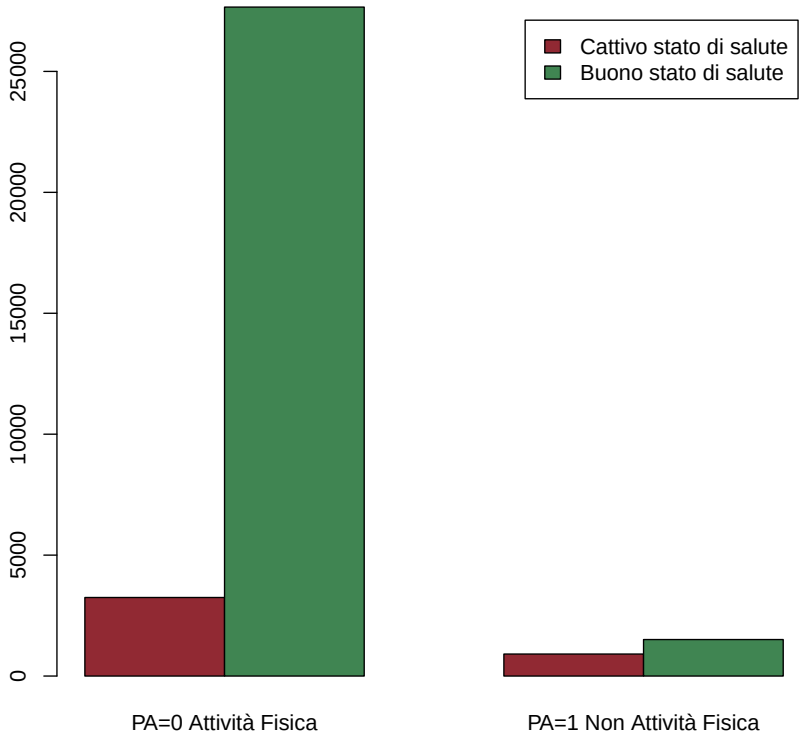
La dicotomizzazione che ho utilizzato è:

0. FITNESS:Little Worse, Much Worse
1. FITNESS:Just as Good, A bit Better, Much Better

```
#Dicotomizzazione Fitness
fitness.dic <- as.numeric(fitness == "Just as good"|
                          fitness == "A bit better"|
                          fitness == "Much better")
```

```
#Regressione tra Fitness e PA
fit.fitness <- glm(fitness.dic~nmc$pa, family=binomial)
summary(fit.fitness)

##
## Call:
## glm(formula = fitness.dic ~ nmc$pa, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1227   0.4712   0.4712   0.4712   0.9715
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.14195    0.01855  115.49   <2e-16 ***
## nmc$pa      -1.63619    0.04589  -35.66   <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: 25083  on 33326  degrees of freedom
## Residual deviance: 23981  on 33325  degrees of freedom
## AIC: 23985
##
## Number of Fisher Scoring iterations: 4
```

```r
#Barplot Fitness Dicotomizzata e PA
barplot(table(fitness.dic, nmc$pa),
        names.arg=c("PA=0 Attività Fisica", "PA=1 Non Attività Fisica"),
        legend.text=c("Cattivo stato di salute","Buono stato di salute"),
        col=c("#912933", "#408552"), beside=TRUE)
```

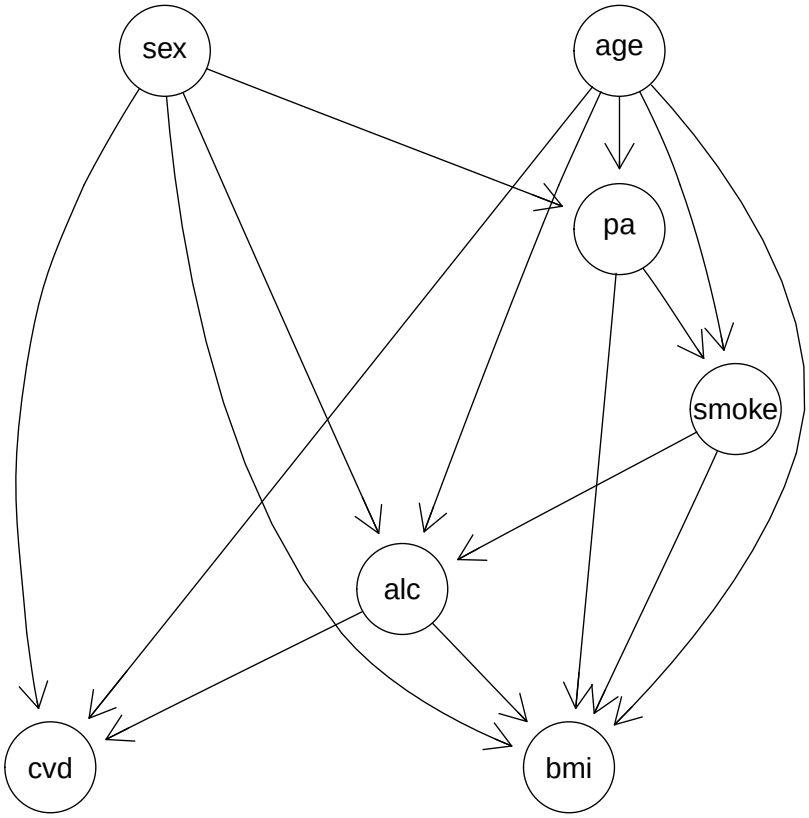


Come è possibile vedere anche dalla semplice regressione logistica, l'attività fisica influisce positivamente nello stato di salute dell'individuo.

Dato che la variabile PA risulta a carattere più oggettivo rispetto a FITNESS, ritengo più valido dar importanza alla variabile PA rispetto a FITNESS per la valutazione di insorgenza di CVD.

Valutiamo ora la rete bayesiana associato al Dataset senza la presenza della variabile FITNESS, tenendo conto dell'ordinamento delle variabili fatto precedentemente.

```
#Rete Bayesiana senza Fitness
#1-SEX, 1-AGE, 3-BMI, 4-CVD, 2-PA, 2-SMOKE, 2-ALC
nmc.bn = subset(nmc.bn, select=-c(fitness))
block<-c(1, 1, 3, 4, 2, 2, 2)
blnmc.bn <- matrix(0, nrow=7, ncol=7)
rownames(blnmc.bn) <- colnames(blnmc.bn) <- names(nmc.bn)
for (b in 2:4) blnmc.bn[block==b, block<b] <- 1
blnmc.bn[1,2] = 1
blnmc.bn[2,1] = 1
blackL <- data.frame(get.edgelist(as(blnmc.bn, "igraph")))
names(blackL) <- c("from", "to")
m.bn <- hc(nmc.bn, blacklist=blackL)
plot(as(amat(m.bn), "graphNEL"))
```



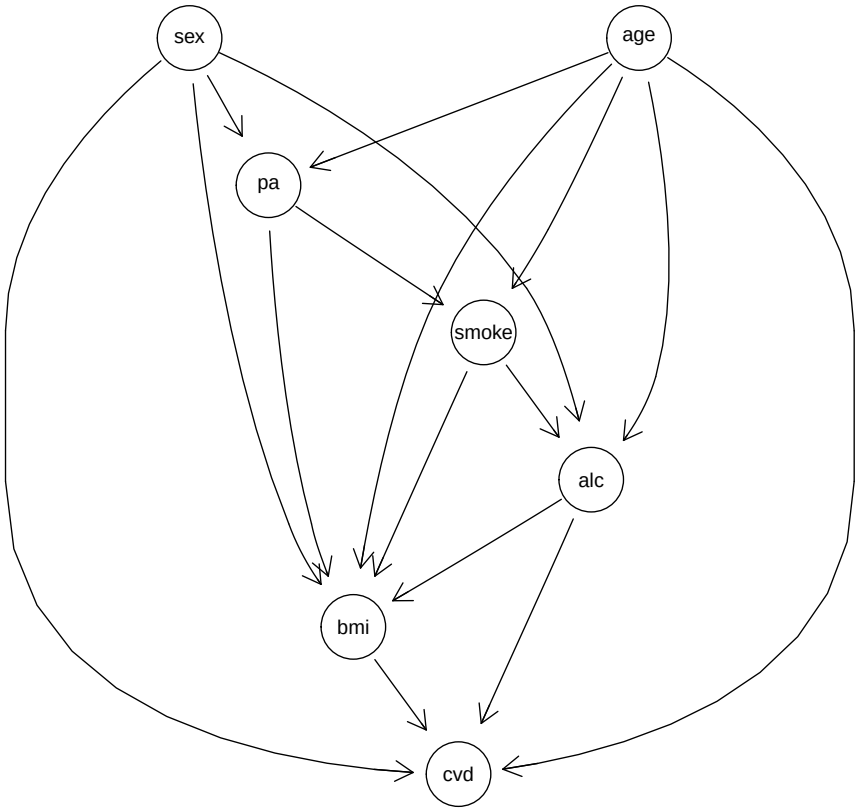
Possiamo vedere come in questa rete la variabile di risposta CVD sia sempre direttamente connessa con le variabili SEX, AGE e ALCHOL. Da notare come in

questo ultimo grafo la variabile BMI sia influenzata direttamente dalle variabili SEX, AGE, ALCHOL, PA e SMOKE.

9.2 BMI

Precedentemente, durante l'analisi delle regressioni logistiche, abbiamo verificato come un aumento dell'indice di massa corporea sia direttamente connessa all'insorgenza di malattie cardiovascolari. Per questo motivo, all'interno del grafico andremo ad aggiungere l'arco tra BMI e CVD.

```
#Rete Bayesiana con arco da BMI a CVD
m.bn.bmicvd <- DAG(cvd~sex:age:alc:bmi,alc~sex:age:smoke,smoke~pa:age,
                   pa~sex:age, bmi~sex:alc:pa:age:smoke)
plot(as(m.bn.bmicvd, "graphNEL"))
```



Verifichiamo ora l'influenza delle variabili SEX, AGE, PA, ALC e SMOKE con la variabile BMI.

```r
fit.bmi <- glm(nmc$bmi~nmc$sex+nmc$age+nmc$pa+smoke.ord+nmc$alc,
               family=binomial)
summary(fit.bmi)
```

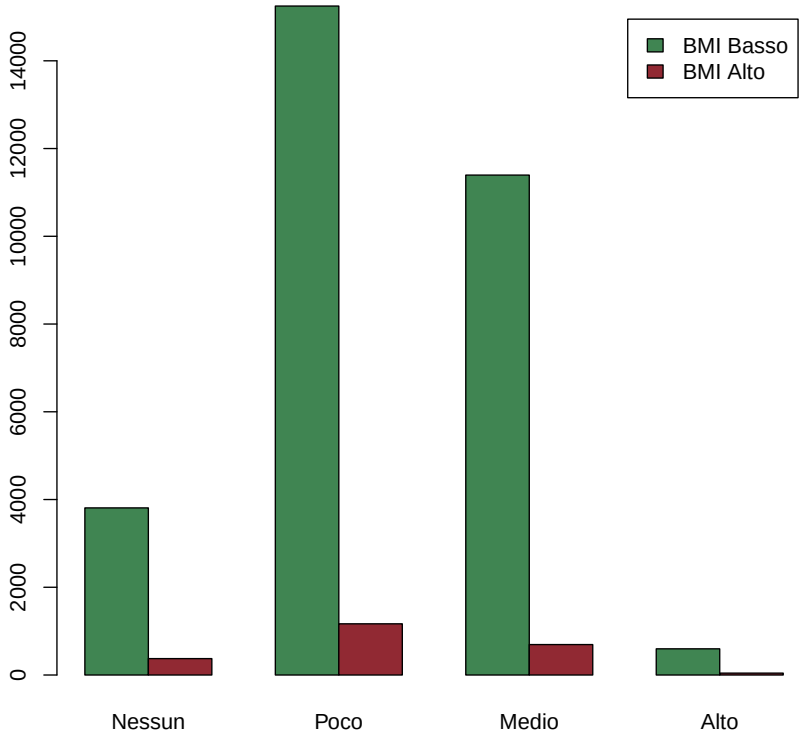
```
##
## Call:
## glm(formula = nmc$bmi ~ nmc$sex + nmc$age + nmc$pa + smoke.ord +
##     nmc$alc, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8170  -0.4024  -0.3567  -0.3181   2.7488
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.926869   0.104871 -27.909  < 2e-16 ***
## nmc$sexMale -0.324341   0.049706  -6.525 6.79e-11 ***
## nmc$age      0.011322   0.001377   8.220  < 2e-16 ***
## nmc$pa       0.758050   0.066422  11.413  < 2e-16 ***
## smoke.ord    0.213703   0.033352   6.407 1.48e-10 ***
## nmc$alc     -0.235959   0.032115  -7.347 2.02e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16621  on 33326  degrees of freedom
## Residual deviance: 16319  on 33321  degrees of freedom
## AIC: 16331
##
## Number of Fisher Scoring iterations: 5
```

Dalla regressione logistica per BMI risulta che:

- Tutte le variabili SEX, AGE, PA, ALC e SMOKE risultano significative.
- Fare attività fisica riduce il BMI.
- L'età aumenta l'indice BMI.
- Essere un fumatore o un ex-fumatore aumenta l'indice BMI.
- Il sesso maschile ha un indice BMI inferiore rispetto al sesso femminile.
- Il consumo di alchol sembra diminuire l'indice BMI.

Verifichiamo tramite Barplot come si distribuiscono le persone sia in base all'indice di BMI e al consumo di alchol.

```
barplot(table(nmc$bmi, nmc$alc),
        names.arg=c("Nessun","Poco",
                    "Medio", "Alto"),
        legend.text=c("BMI Basso", "BMI Alto"),
        col=c("#408552", "#912933"), beside=TRUE)
```



Dal Barplot possiamo vedere come ci sia una prevalenza di persone con basso indice BMI per ogni categoria di consumatori di alchol rispetto alle persone con un alto indice di BMI.

Analizziamo allora la percentuale tra queste due categorie di persone.

```
#Percentuale persone con alto BMI che bevono molto alchol
n.alc.high = nrow(nmc.bn[nmc.bn$alc==4,])
nrow(nmc.bn[nmc.bn$bmi==1&nmc.bn$alc==4,])/n.alc.high

## [1] 0.0671875
```

```r
#Percentuale persone con alto BMI che bevono alchol nella media
n.alc.med = nrow(nmc.bn[nmc.bn$alc==3,])
nrow(nmc.bn[nmc.bn$bmi==1&nmc.bn$alc==3,])/n.alc.med

## [1] 0.05740756

#Percentuale persone con basso BMI che bevono molto alchol
n.alc.high = nrow(nmc.bn[nmc.bn$alc==4,])
nrow(nmc.bn[nmc.bn$bmi==0&nmc.bn$alc==4,])/n.alc.high

## [1] 0.9328125

#Percentuale persone con bevono BMI che bevono alchol nella media
n.alc.med = nrow(nmc.bn[nmc.bn$alc==3,])
nrow(nmc.bn[nmc.bn$bmi==0&nmc.bn$alc==3,])/n.alc.med

## [1] 0.9425924
```

Verifichiamo se togliendo la variabile ALCHOL per BMI risulti ancora significativo.

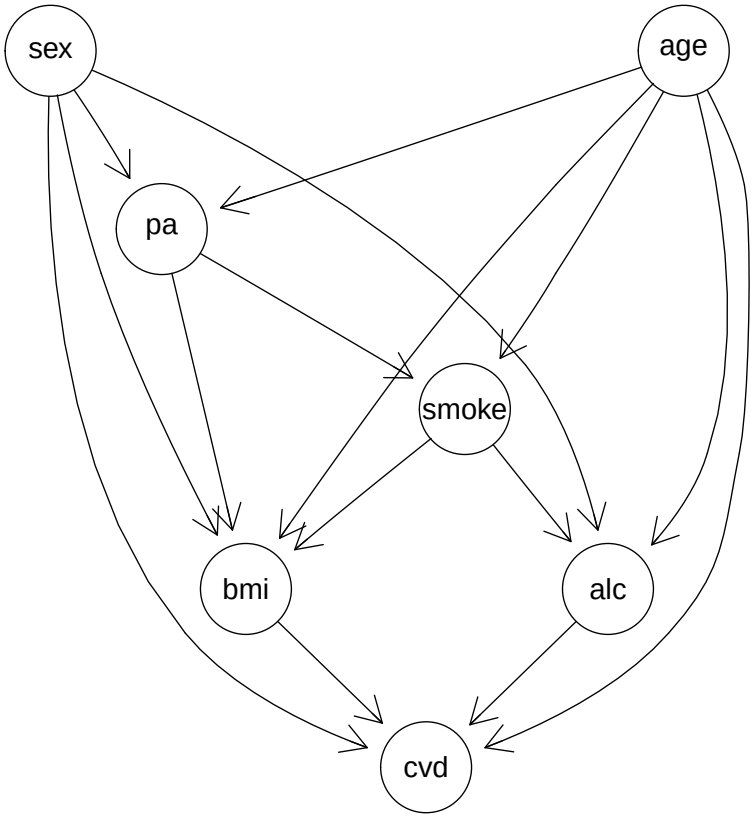
```r
fit.bmi.nalc <- glm(nmc$bmi~nmc$sex+nmc$age+nmc$pa+smoke.ord,
               family=binomial)
summary(fit.bmi.nalc)

##
## Call:
## glm(formula = nmc$bmi ~ nmc$sex + nmc$age + nmc$pa + smoke.ord,
##     family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6986  -0.4001  -0.3627  -0.3209   2.5988
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.341359   0.089561 -37.308  < 2e-16 ***
## nmc$sexMale -0.379425   0.049150  -7.720 1.17e-14 ***
## nmc$age      0.010972   0.001392   7.880 3.28e-15 ***
## nmc$pa       0.765789   0.066320  11.547  < 2e-16 ***
## smoke.ord    0.159274   0.032623   4.882 1.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16621  on 33326  degrees of freedom
```

```
## Residual deviance: 16373  on 33322  degrees of freedom
## AIC: 16383
##
## Number of Fisher Scoring iterations: 5
```

Anche senza la presenza della variabile ALCHOL, questo modello per BMI risulta significativo non modificando di molto il comportamento e la significatività delle altre variabili.

Rimuoviamo quindi l'arco tra ALCHOL e BMI.

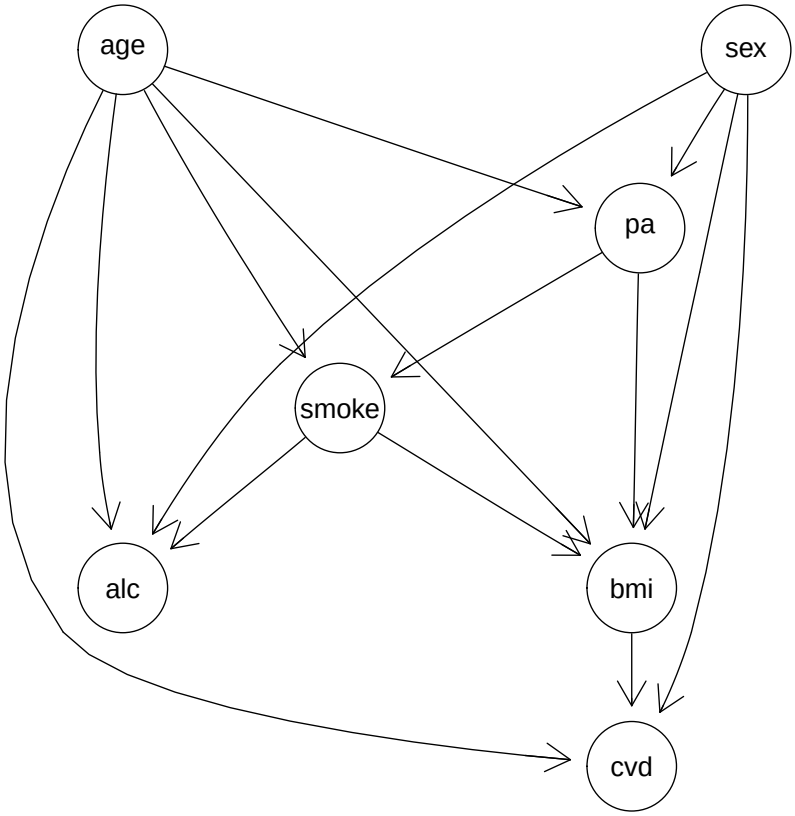
```
#Rete Bayesiana senza arco da Alchol a BMI
m.bn.bmicvd <- DAG(cvd~sex:age:alc:bmi, alc~sex:age:smoke, smoke~pa:age,
                   pa~sex:age, bmi~sex:pa:age:smoke)
plot(as(m.bn.bmicvd, "graphNEL"))
```



9.3 Alchol

Sempre durante l'analisi delle regressioni logistiche e durante la selezione del modello con le procedure Forward e Backward, abbiamo visto come la variabile Alchol non influenzi la variabile di risposta CVD. Per questo motivo, dato che è presente l'arco tra CVD e ALCHOL all'interno nella rete, andremo a rimuovere l'arco da ALCHOL a CVD.

```
#Rete Bayesiana senza arco da Alchol a CVD
m.bn.bmicvd <- DAG(cvd~sex:age:bmi, alc~sex:age:smoke, smoke~pa:age,
                   pa~sex:age, bmi~sex:pa:age:smoke)
plot(as(m.bn.bmicvd, "graphNEL"))
```



Da questo DAG capiamo che la variabile di risposta non solo è connessa con le variabili SEX, AGE ma anche dalla variabile BMI.

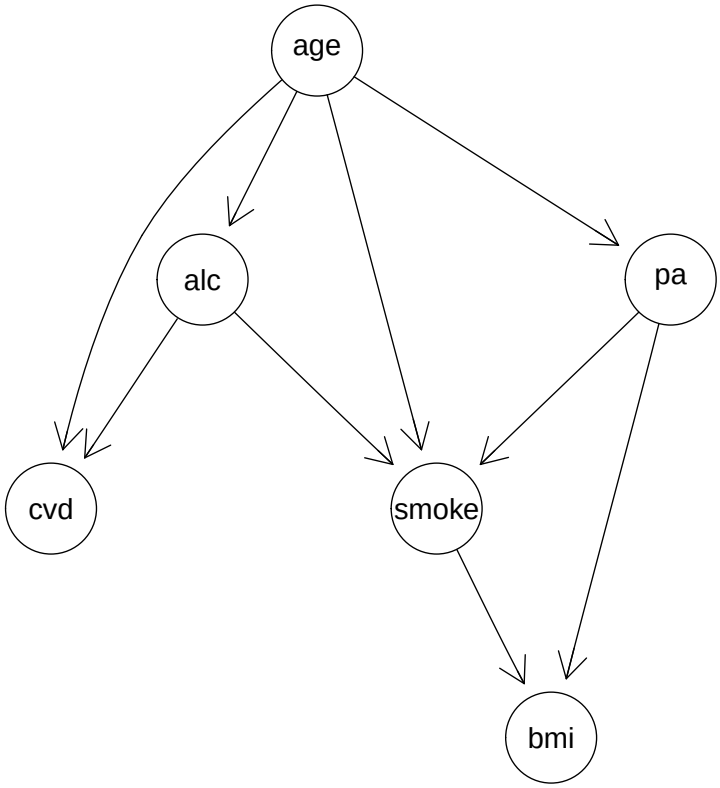
9.4 Maschio e Femmina

Durante le precedenti analisi, abbiamo notato come ci sia una probabilità maggiore del sesso maschile rispetto a quello femminile.

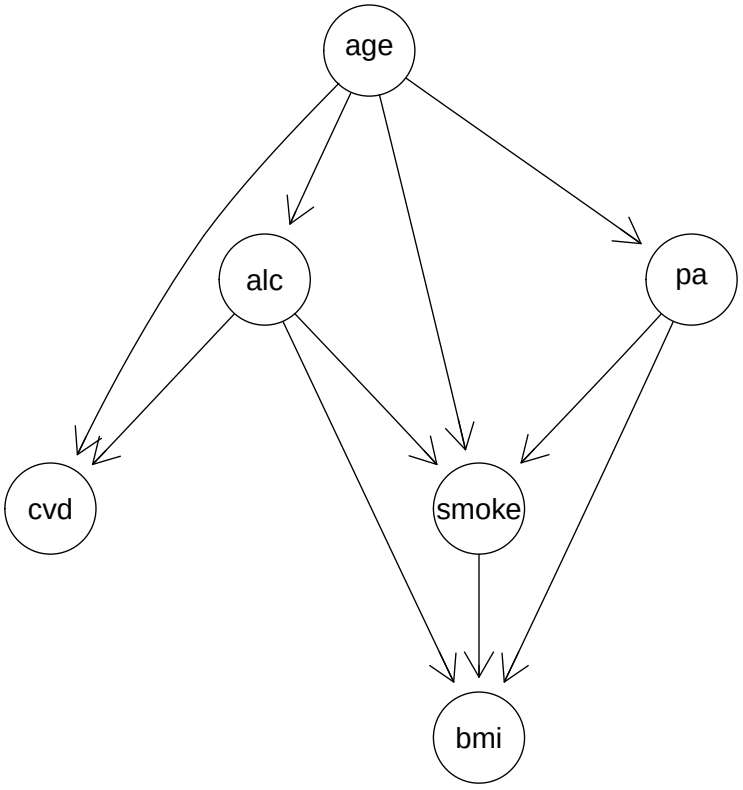
Valutiamo la differenza di due reti bayesiane che comprendo per una il genere maschile e l'altra quella femminile.

```
#Dataset: sotto-problema Sex
male <- (nmc.bn$sex==1)
female <- (nmc.bn$sex==0)
nmc.bn.male <- nmc.bn[male, c(2:7)]
nmc.bn.female <- nmc.bn[female, c(2:7)]
```

```
#Rete Bayesiana per Male
#1-AGE, 3-BMI, 4-CVD, 2-PA, 2-SMOKE, 2-ALC
block<-c(1, 3, 4, 2, 2, 2)
blnmc.bn <- matrix(0, nrow=6, ncol=6)
rownames(blnmc.bn) <- colnames(blnmc.bn) <- names(nmc.bn.male)
for (b in 2:4) blnmc.bn[block==b, block<b] <- 1
blnmc.bn[1,2] = 1
blnmc.bn[2,1] = 1
blackL <- data.frame(get.edgelist(as(blnmc.bn, "igraph")))
names(blackL) <- c("from", "to")
m.bn.male <- hc(nmc.bn.male, blacklist=blackL)
plot(as(amat(m.bn.male), "graphNEL"))
```




```
#Rete Bayesiana per Female
#1-AGE, 3-BMI, 4-CVD, 2-PA, 2-SMOKE, 2-ALC
block<-c(1, 3, 4, 2, 2, 2)
blnmc.bn <- matrix(0, nrow=6, ncol=6)
rownames(blnmc.bn) <- colnames(blnmc.bn) <- names(nmc.bn.female)
for (b in 2:4) blnmc.bn[block==b, block<b] <- 1
blnmc.bn[1,2] = 1
blnmc.bn[2,1] = 1
blackL <- data.frame(get.edgelist(as(blnmc.bn, "igraph")))
names(blackL) <- c("from", "to")
m.bn.female <- hc(nmc.bn.female, blacklist=blackL)
plot(as(amat(m.bn.female), "graphNEL"))
```



Le reti risultanti risultano uguali se non consideriamo l'arco presente tra la variabile ALCHOL E BMI e quello tra ALCHOL e CVD.

Verifichiamo adesso la percentuale tra uomini e donne nelle varie categorie.

```
n.male = nrow(nmc.bn[nmc.bn$sex==1,])
n.female = nrow(nmc.bn[nmc.bn$sex==0,])
```

```
#Percentuale degli uomini fumano
nrow(nmc.bn[nmc.bn$smoke==3&nmc.bn$sex==1,])/n.male

## [1] 0.0672956

#Percentuale degli uomini ex-fumatori
nrow(nmc.bn[nmc.bn$smoke==2&nmc.bn$sex==1,])/n.male

## [1] 0.2915544
```

```r
#Percentuale delle donne che fumano
nrow(nmc.bn[nmc.bn$smoke==3&nmc.bn$sex==0,])/n.female

## [1] 0.08663333

#Percentuale delle donne ex-fumatori
nrow(nmc.bn[nmc.bn$smoke==2&nmc.bn$sex==0,])/n.female

## [1] 0.2471055
```

```r
#Percentuale degli uomini fanno attività fisica
nrow(nmc.bn[nmc.bn$pa==0&nmc.bn$sex==1,])/n.male

## [1] 0.9184187
```

```r
nrow(nmc.bn[nmc.bn$pa==0&nmc.bn$sex==0,])/n.female

## [1] 0.9319277
```

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: CVD $\sim$ SEX + AGE + BMI

Infatti i fattori che aumentano la probabilità di CVD sono:

- L'aumento dell'età, con una maggior evidenza superata la soglia dei 40 anni.
- Essere maschio.
- Avere un alto indice di massa corporea.

Tuttavia dobbiamo tenere in considerazione che il fumare può aumentare l'indice di massa corporea, andando ad influenzare negativamente l'indice di massa corporea.

Sempre per ridurre l'indice di massa corporea e quindi prevenire in qualche modo la CVD è consigliabile fare attività fisica.