NMC - Foundations of Statistical Modelling

Lorenzo Baiardi

28 Febbraio 2023

Contents

1 Introduzione

In questo elaborato andremo a studiare l'effeto delle attività personale 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 Dataset

Per lo studio di questo fenomeno utilizzeremo il dataset fornito:" Sjolander et al. (2009)" di cui visualizzeremo le caratteristiche.
Il Dataset fornisce un campione di numerosità: n = 33327 osservazioni.

```
load("nmc.RData")
#dicotomizzazione della variabile BMI
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" ...
```

Variabili:

- CVD: variabile d'interesse, presenza o meno di almeno un problema cardiovascolare.
  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: quantità dell'attività fisica 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.
  1. NO
  2. Former
  3. Current
- Alchol: frequenza dell'uso di alchol dell'individuo.
  1. Never
  2. Low
  3. Medium
  4. High

Per semplicificazione e maggior comprensione del problema, convertiremo alcune variabili categoriali in variabili ordinali facilitando l'utilizzo durante l'uso delle opportune funzioni.
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
#Smoke: 1-NO, 2-FORMER, 3-CURRENT
#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.smoke = c('NO', 'Former', 'Current')
c.alc =  c('Never', 'Low', 'Medium', 'High')

#Variabili ordinali
```

```r
nmc$fitness = as.numeric(ordered(nmc$fitness, c.fit))
nmc$smoke = as.numeric(ordered(nmc$smoke, c.smoke))
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  : num  1 1 2 2 2 2 1 2 1 1 ...
##  $ alc    : num  3 2 1 1 3 2 1 3 3 1 ...
```

3 Analisi Variabili

Dato che stiamo analizzando un problema che presenta come variabile di risposta una varibile binaria (CVD)utilizzeremo la regressione logistica, implementata in R tramite la funzione GLM.
Per prima cosa analizzeremo le singole regressioni logistiche semplici per ogni variabili all'interno del dataset.

```
#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
```

```
#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
```

```
#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
```

```r
#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
```

```r
#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
```

```
#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
## -0.3281  -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 ***
## nmc$smoke    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
```

```
#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
```

- 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 risultano non significative sono: PA, Smoke e Alchol.

Da notare che nella visualizzazione della regressione logistica semplice del sesso risulta che il sesso Maschile aumenta notevolmente la possibilità di incorrere in un CVD rispetto al sesso Femminale con valore: $SexMale \sim 0.91$.

4 Regressione multipla

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

```
#Regressioni logistiche multiple
#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
## -1.5830  -0.3387  -0.1935  -0.0949   3.6481
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.951164   0.212013 -37.503  < 2e-16 ***
## nmc$sexMale  0.787958   0.054412  14.481  < 2e-16 ***
## nmc$age      0.092492   0.002445  37.835  < 2e-16 ***
## nmc$bmi      0.225567   0.096862   2.329   0.0199 *
## nmc$fitness -0.183605   0.031060  -5.911  3.4e-09 ***
## nmc$pa       0.037140   0.108463   0.342   0.7320
## nmc$smoke    0.125146   0.045177   2.770   0.0056 **
## nmc$alc     -0.062944   0.035517  -1.772   0.0764 .
## ---
## 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: 10887  on 33319  degrees of freedom
## AIC: 10903
##
## 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 e gli errori standard nella regressione logistica multipla differescono da quelli presenti nelle regressioni logistiche semplici precedentemente analizzate.
- La variabile Sex mostra ancora come il sesso Maschile influisca positivamente nella presenza di CVD con valore: $SexMale \sim 0.788$.
- Anche le variabili BMI e Smoke mostrano un aumento nelle possibilità di insorgenza di un CVD.
- Viceversa, la variabile Fitness sembra ridurre la presenza di CVD con valore: $Fitness \sim -0,184$.

Selezioniamo ora i modelli eliminando gradualmente nella formula le variabili di significatività non superiori al 5%.

```
#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$alc, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5841  -0.3388  -0.1937  -0.0950   3.6467
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.938561   0.208732 -38.032  < 2e-16 ***
## nmc$sexMale  0.788736   0.054368  14.507  < 2e-16 ***
## nmc$age      0.092451   0.002441  37.869  < 2e-16 ***
## nmc$bmi      0.226349   0.096833   2.338  0.01941 *
## nmc$fitness -0.185841   0.030364  -6.120 9.33e-10 ***
## nmc$smoke    0.125336   0.045174   2.775  0.00553 **
## nmc$alc     -0.063108   0.035516  -1.777  0.07559 .
## ---
## 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: 10887  on 33320  degrees of freedom
## AIC: 10901
##
## Number of Fisher Scoring iterations: 7
```

La variabile Alchol risulta ancora non significativa.
Verifichiamo ora il modello senza la variabile Pa e Alchol.

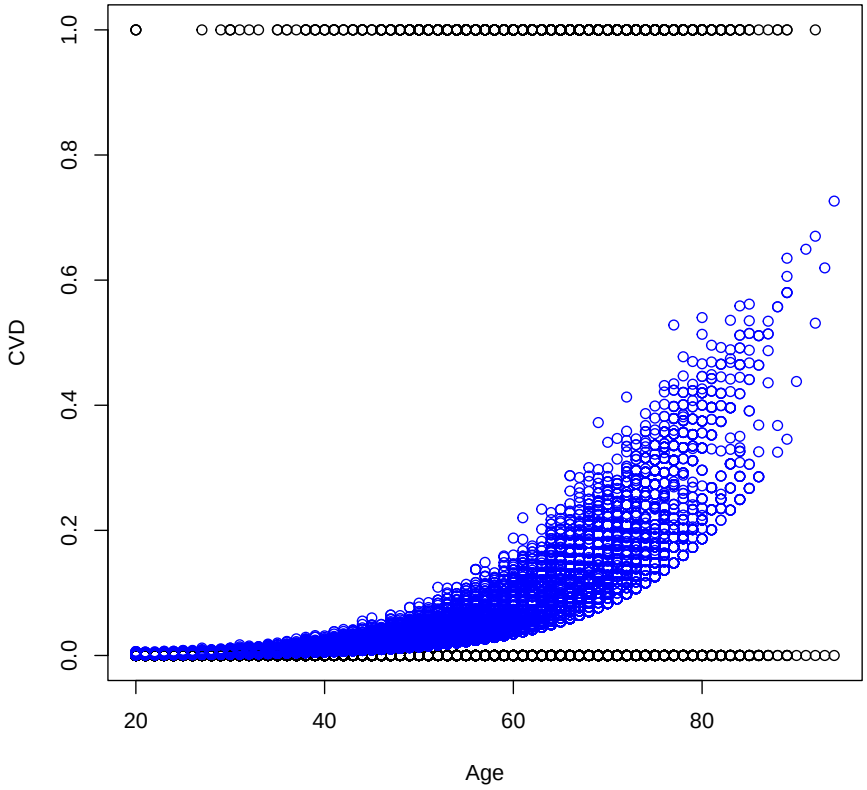
```
#Variabili: Sex, Age, BMI, Fitness, Smoke
fit <- glm(nmc$cvd ~ nmc$sex+nmc$age+nmc$bmi+nmc$fitness+
                     nmc$smoke, family=binomial)
summary(fit)

##
## Call:
## glm(formula = nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness +
##     nmc$smoke, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6098  -0.3381  -0.1941  -0.0936   3.6516
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.063977   0.196867 -40.962  < 2e-16 ***
## nmc$sexMale  0.772699   0.053590  14.419  < 2e-16 ***
## nmc$age      0.092827   0.002437  38.091  < 2e-16 ***
## nmc$bmi      0.230557   0.096798   2.382   0.0172 *
## nmc$fitness -0.188620   0.030326  -6.220 4.98e-10 ***
## nmc$smoke    0.107297   0.044058   2.435   0.0149 *
## ---
## 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: 10890  on 33321  degrees of freedom
## AIC: 10902
##
## Number of Fisher Scoring iterations: 7
```

- Le variabili risultato essere tutte significative secondo il *p-value*.

- I parametri stimati e gli errori standard non differiscono molto dal modello con tutte le variabili.

Visualizziamo ora il grafico dell'andamento del modello stimato con solo variabili significative. Il grafico riporterà i valori ordinati in base all'età.

```
pstima <- fit$fitted.values

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



Secondo il grafico e secondo il modello con solo variabili significative, all'aumentare dell'età di un individuo aumenta esponenzialmente la probabilità di avere un problema cardiovascolare.

4.1 Dati di esempio

Effettuiamo una valutazione della probabilità su uno specifico individuo in base ai modelli precedentemente valutati.

5 Interazioni fra Variabili

Valutiamo ora se è possibile che ci possano essere delle interazioni fra le variabili. Consideriamo il caso nel quale il fumo, alchol o il genere possano interagire con le altre variabili.

```
#Modello con interazione: Smoke e Alcol
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
## -1.5880  -0.3400  -0.1934  -0.0949   3.6478
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -8.015240   0.292281 -27.423  < 2e-16 ***
## nmc$sexMale        0.788591   0.054365  14.505  < 2e-16 ***
## nmc$age            0.092512   0.002447  37.802  < 2e-16 ***
## nmc$bmi            0.226148   0.096840   2.335   0.0195 *
## nmc$fitness       -0.185734   0.030365  -6.117 9.55e-10 ***
## nmc$smoke          0.180349   0.153282   1.177   0.2394
## nmc$alc           -0.032541   0.088925  -0.366   0.7144
## nmc$smoke:nmc$alc -0.022446   0.059852  -0.375   0.7076
## ---
## 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: 10887  on 33319  degrees of freedom
## AIC: 10903
##
## Number of Fisher Scoring iterations: 7
```

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

```r
#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
## -1.6033  -0.3382  -0.1923  -0.0939   3.6557
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -8.112470   0.197792 -41.015  < 2e-16 ***
## nmc$sexMale        0.777961   0.053680  14.492  < 2e-16 ***
## nmc$age            0.092762   0.002437  38.066  < 2e-16 ***
## nmc$bmi            0.886306   0.253661   3.494 0.000476 ***
## nmc$fitness       -0.188499   0.030341  -6.213 5.21e-10 ***
## nmc$smoke          0.141855   0.045493   3.118 0.001820 **
## nmc$bmi:nmc$smoke -0.463096   0.171427  -2.701 0.006904 **
## ---
## 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 33320  degrees of freedom
## AIC: 10897
##
## Number of Fisher Scoring iterations: 7
```

A differenza di Smoke e Alchol, l'interazione tra Smoke e BMI sembrano mostrare un'interazione significativa.
Proseguiamo con l'analisi di altre interazioni

```r
#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
## -1.5804  -0.3379  -0.1940  -0.0951   3.6456
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -7.912564   0.210117 -37.658  < 2e-16 ***
## nmc$sexMale      0.786844   0.054377  14.470  < 2e-16 ***
## nmc$age          0.092458   0.002441  37.872  < 2e-16 ***
## nmc$bmi         -0.044368   0.282346  -0.157  0.87513
## nmc$fitness     -0.186165   0.030365  -6.131 8.74e-10 ***
## nmc$smoke        0.125571   0.045180   2.779  0.00545 **
## nmc$alc         -0.073829   0.036998  -1.995  0.04599 *
## nmc$bmi:nmc$alc  0.121928   0.118248   1.031  0.30248
## ---
## 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 33319  degrees of freedom
## AIC: 10902
##
## Number of Fisher Scoring iterations: 7
```

L'interazione tra Alchol e BMI sembrerebbe essere non supportata.

```
#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
```

```
## -1.5117  -0.3451  -0.1919  -0.0919   3.7412
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -8.559984   0.259836 -32.944  < 2e-16 ***
## nmc$sexMale          1.700834   0.307213   5.536 3.09e-08 ***
## nmc$age              0.100391   0.003532  28.424  < 2e-16 ***
## nmc$bmi              0.224017   0.096838   2.313  0.02070 *
## nmc$fitness         -0.187792   0.030229  -6.212 5.22e-10 ***
## nmc$smoke            0.118232   0.044138   2.679  0.00739 **
## nmc$sexMale:nmc$age -0.014618   0.004762  -3.070  0.00214 **
## ---
## 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: 10881  on 33320  degrees of freedom
## AIC: 10895
##
## Number of Fisher Scoring iterations: 7
```

Anche qui abbiamo un interazione tra la variabile Sex e Age.

Analizziamo ora il modello con solo variabili significative integrandolo con le interazioni significative.
Il modello da valutare sarà:
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
## -1.5054  -0.3439  -0.1903  -0.0924   3.7451
##
## Coefficients:
```

```
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -8.608225   0.260647 -33.026  < 2e-16 ***
## nmc$sexMale          1.705155   0.307343   5.548 2.89e-08 ***
## nmc$age              0.100326   0.003534  28.386  < 2e-16 ***
## nmc$bmi              0.879152   0.253816   3.464 0.000533 ***
## nmc$fitness         -0.187646   0.030244  -6.204 5.49e-10 ***
## nmc$smoke            0.152533   0.045549   3.349 0.000812 ***
## nmc$sexMale:nmc$age -0.014604   0.004764  -3.066 0.002173 **
## nmc$bmi:nmc$smoke   -0.462549   0.171461  -2.698 0.006982 **
## ---
## 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: 10873  on 33319  degrees of freedom
## AIC: 10889
##
## Number of Fisher Scoring iterations: 7
```

Il modello risultante risulta essere significativo.
Nonostante ciò, per avere un modello il più semplice possibile valuteremo nel corso dell'elaborato il modello indipendente.

6 Visualizzazione Differenze

Precedentemente, durante l'analisi dei vari modelli e delle variabili, abbiamo notato come ci siano delle differenze all'interno dei dati.
Abbiamo notato come il sesso maschile abbia una maggior possibilità di CVD rispetto al sesso femmile.
Valutiamo all'interno di un grafico se questa nostra ipotesi sia verificata nel modello.

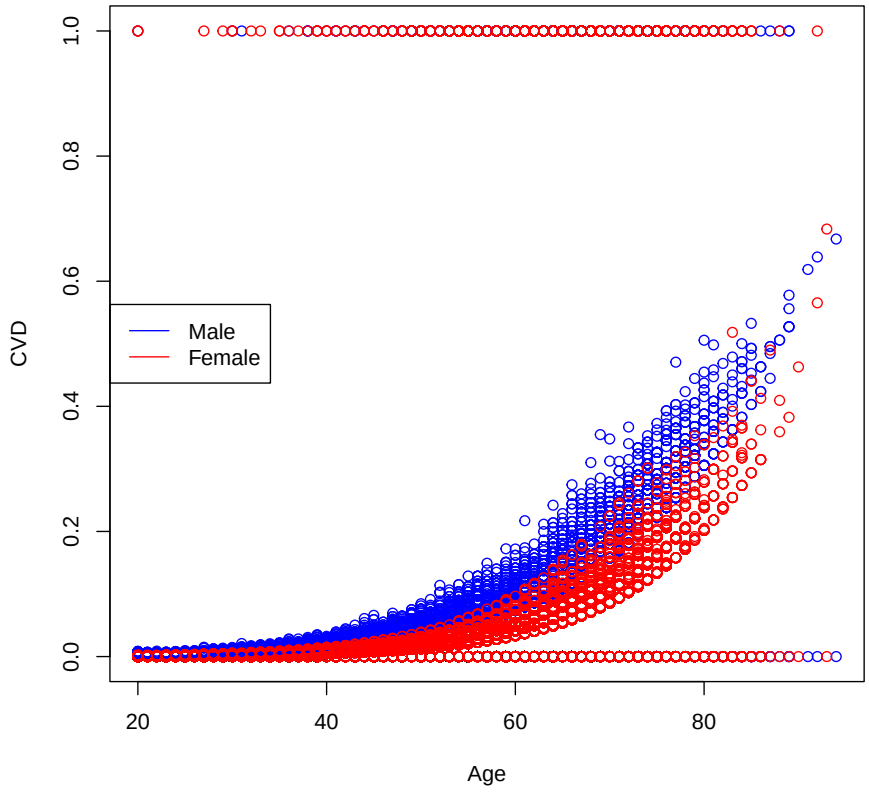
```
#Modello nel caso di Sesso Maschile e Femminile
#Sex: Male
fit.male <- glm(nmc$cvd[nmc$sex=="Male"] ~
                  nmc$age[nmc$sex=="Male"]+
                  nmc$bmi[nmc$sex=="Male"]+
                  nmc$fitness[nmc$sex=="Male"]+
                  nmc$smoke[nmc$sex=="Male"],
                  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
## -1.4838  -0.4607  -0.2786  -0.1121   3.4515
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -7.070001   0.272214 -25.972  < 2e-16 ***
## nmc$age[nmc$sex == "Male"]      0.085549   0.003304  25.896  < 2e-16 ***
## nmc$bmi[nmc$sex == "Male"]      0.116204   0.156961   0.740 0.459095
## nmc$fitness[nmc$sex == "Male"] -0.159803   0.042160  -3.790 0.000150 ***
## nmc$smoke[nmc$sex == "Male"]    0.204226   0.059384   3.439 0.000584 ***
## ---
## 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: 5301.1  on 11125  degrees of freedom
## AIC: 5311.1
##
## Number of Fisher Scoring iterations: 6
```

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

#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  -0.2824  -0.1629  -0.0841   3.7212
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -8.294401   0.291381 -28.466  < 2e-16 ***
## nmc$age[nmc$sex == "Female"]      0.099545   0.003598  27.664  < 2e-16 ***
## nmc$bmi[nmc$sex == "Female"]      0.280513   0.123385   2.273    0.023 *
## nmc$fitness[nmc$sex == "Female"] -0.212546   0.043505  -4.886 1.03e-06 ***
## nmc$smoke[nmc$sex == "Female"]    0.018569   0.067375   0.276    0.783
## ---
## 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: 5574.1  on 22192  degrees of freedom
## AIC: 5584.1
##
## Number of Fisher Scoring iterations: 7
```

```r
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)
```



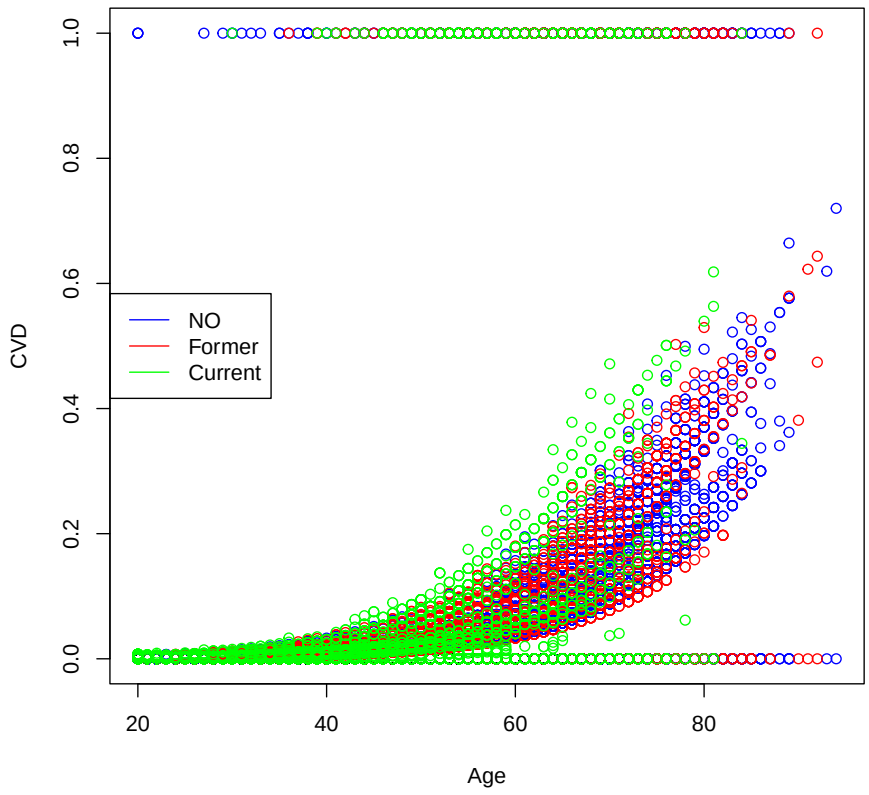
Visualizzando il grafico, si può verificare direttamente come il genere maschile sia più a soggetto a rischi di CVD rispetto al genere femminile, validando quindi l'ipotesi precedentemente descritta.

Analizziamo ora se ci siano delle differenze anche tra le varie categorie di fumatori.

```r
#Modello nel caso di Smoke: NO, Former e Current.
#Smoke: NO
fit.smokeNO <- glm(nmc$cvd[nmc$smoke==1] ~
                     nmc$sex[nmc$smoke==1]+
                     nmc$age[nmc$smoke==1]+
                     nmc$bmi[nmc$smoke==1]+
                     nmc$fitness[nmc$smoke==1],
                     family=binomial)
summary(fit.smokeNO)

##
## Call:
## glm(formula = nmc$cvd[nmc$smoke == 1] ~ nmc$sex[nmc$smoke ==
##     1] + nmc$age[nmc$smoke == 1] + nmc$bmi[nmc$smoke == 1] +
##     nmc$fitness[nmc$smoke == 1], family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5957  -0.3401  -0.1835  -0.0799   3.6504
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -8.016035   0.223722 -35.830  < 2e-16 ***
## nmc$sex[nmc$smoke == 1]Male   0.704569   0.067084  10.503  < 2e-16 ***
## nmc$age[nmc$smoke == 1]       0.093265   0.002991  31.178  < 2e-16 ***
## nmc$bmi[nmc$smoke == 1]       0.375275   0.119900   3.130  0.00175 **
## nmc$fitness[nmc$smoke == 1] -0.170238   0.038071  -4.472 7.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8638.5  on 21924  degrees of freedom
## Residual deviance: 6969.1  on 21920  degrees of freedom
## AIC: 6979.1
##
## Number of Fisher Scoring iterations: 7

pstima.smokeNO <- fit.smokeNO$fitted.values

#Smoke: Former
fit.smokeFormer <- glm(nmc$cvd[nmc$smoke==2] ~
                         nmc$sex[nmc$smoke==2]+
                         nmc$age[nmc$smoke==2]+
                         nmc$bmi[nmc$smoke==2]+
                         nmc$fitness[nmc$smoke==2],
```

```
                          family=binomial)
summary(fit.smokeFormer)

##
## Call:
## glm(formula = nmc$cvd[nmc$smoke == 2] ~ nmc$sex[nmc$smoke ==
##     2] + nmc$age[nmc$smoke == 2] + nmc$bmi[nmc$smoke == 2] +
##     nmc$fitness[nmc$smoke == 2], family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4363  -0.3446  -0.2127  -0.1346   3.1190
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -7.750105   0.336333 -23.043  < 2e-16 ***
## nmc$sex[nmc$smoke == 2]Male   0.895537   0.106114   8.439  < 2e-16 ***
## nmc$age[nmc$smoke == 2]       0.089673   0.004835  18.548  < 2e-16 ***
## nmc$bmi[nmc$smoke == 2]       0.160680   0.172033   0.934 0.350302
## nmc$fitness[nmc$smoke == 2] -0.201153   0.056651  -3.551 0.000384 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3780.1  on 8729  degrees of freedom
## Residual deviance: 3136.3  on 8725  degrees of freedom
## AIC: 3146.3
##
## Number of Fisher Scoring iterations: 7

pstima.smokeFormer <- fit.smokeFormer$fitted.values

#Smoke: Current
fit.smokeCurrent <- glm(nmc$cvd[nmc$smoke==3] ~
                          nmc$sex[nmc$smoke==3]+
                          nmc$age[nmc$smoke==3]+
                          nmc$bmi[nmc$smoke==3]+
                          nmc$fitness[nmc$smoke==3],
                          family=binomial)
summary(fit.smokeCurrent)

##
## Call:
## glm(formula = nmc$cvd[nmc$smoke == 3] ~ nmc$sex[nmc$smoke ==
##     3] + nmc$age[nmc$smoke == 3] + nmc$bmi[nmc$smoke == 3] +
```

```
##     nmc$fitness[nmc$smoke == 3], family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3881  -0.2836  -0.1685  -0.0969   3.2755
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -7.776926   0.596957 -13.028  < 2e-16 ***
## nmc$sex[nmc$smoke == 3]Male   1.186402   0.202823   5.849 4.93e-09 ***
## nmc$age[nmc$smoke == 3]       0.095775   0.009278  10.323  < 2e-16 ***
## nmc$bmi[nmc$smoke == 3]      -1.501985   0.734212  -2.046   0.0408 *
## nmc$fitness[nmc$smoke == 3] -0.228125   0.111269  -2.050   0.0403 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 973.15  on 2671  degrees of freedom
## Residual deviance: 761.72  on 2667  degrees of freedom
## AIC: 771.72
##
## Number of Fisher Scoring iterations: 7

pstima.smokeCurrent <- fit.smokeCurrent$fitted.values

#Plot
plot(nmc$age[nmc$smoke==1], nmc$cvd[nmc$smoke==1],
     xlab="Age", ylab="CVD", col="blue")
points(nmc$age[nmc$smoke==2], nmc$cvd[nmc$smoke==2], col="red")
points(nmc$age[nmc$smoke==3], nmc$cvd[nmc$smoke==3], col="green")
points(sort(nmc$age[nmc$smoke==1]),
       pstima.smokeNO[order(nmc$age[nmc$smoke==1])], col="blue")
points(sort(nmc$age[nmc$smoke==2]),
       pstima.smokeFormer[order(nmc$age[nmc$smoke==2])], col="red")
points(sort(nmc$age[nmc$smoke==3]),
       pstima.smokeCurrent[order(nmc$age[nmc$smoke==3])], col="green")
legend(x = "left", legend = c("NO", "Former", "Current"), lty=c(1, 1, 1),
       col=c("blue","red", "green"), lwd = 1)
```



Come immaginabile essere, anche in questo caso i fumatori "Current" rispetto alle altre categoria e alla solità età, mostra una maggior probabilità nell' insorgenza di un problema cardiovascolare.

Analizziamo ora il modello nel caso di differenti PA.

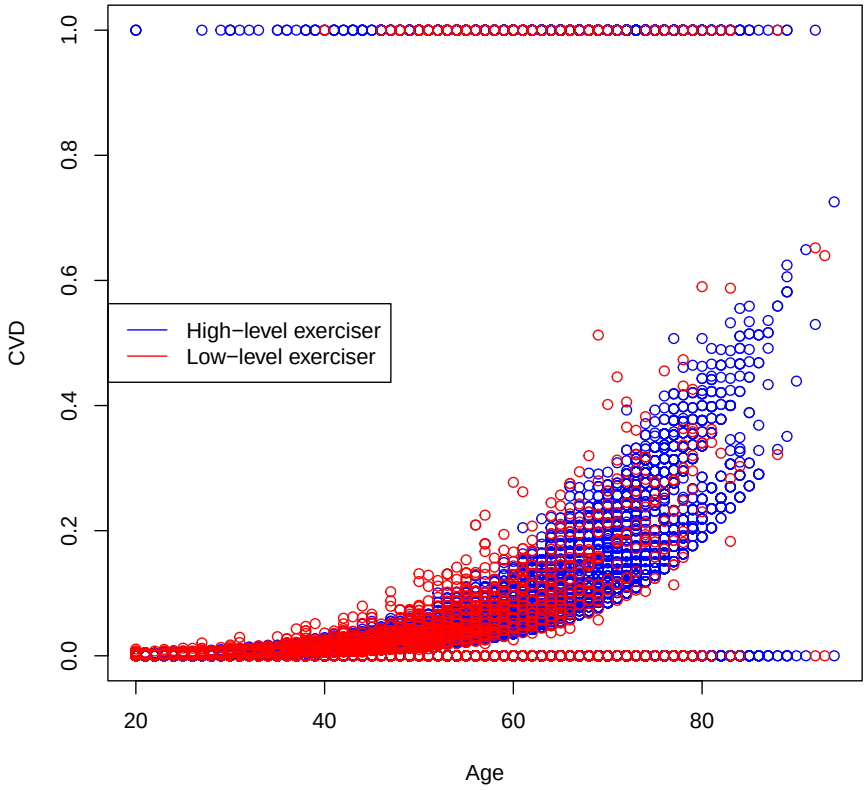
```
#Modello nel caso di PA: 0, 1.
#PA: 0
fit.pa0 <- glm(nmc$cvd[nmc$pa==0] ~
                  nmc$sex[nmc$pa==0]+
                  nmc$age[nmc$pa==0]+
                  nmc$bmi[nmc$pa==0]+
                  nmc$fitness[nmc$pa==0]+
                  nmc$smoke[nmc$pa==0],
                  family=binomial)
summary(fit.pa0)
```

```
##
## 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
## -1.6081  -0.3392  -0.1938  -0.0929   3.6514
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -8.087372   0.208487 -38.791  < 2e-16 ***
## nmc$sex[nmc$pa == 0]Male   0.767611   0.055542  13.820  < 2e-16 ***
## nmc$age[nmc$pa == 0]       0.092832   0.002546  36.460  < 2e-16 ***
## nmc$bmi[nmc$pa == 0]       0.178714   0.104816   1.705   0.0882 .
## nmc$fitness[nmc$pa == 0] -0.178048   0.032332  -5.507 3.65e-08 ***
## nmc$smoke[nmc$pa == 0]     0.099821   0.046057   2.167   0.0302 *
## ---
## 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: 10146  on 30902  degrees of freedom
## AIC: 10158
##
## Number of Fisher Scoring iterations: 7
```

```
pstima.pa0 <- fit.pa0$fitted.values

#PA: 1
fit.pa1 <- 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.pa1)
```

```
##
## 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
## -1.4532  -0.3133  -0.1902  -0.1062   2.9593
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -7.979866   0.669446 -11.920  < 2e-16 ***
## nmc$sex[nmc$pa == 1]Male   0.831774   0.208221   3.995 6.48e-05 ***
## nmc$age[nmc$pa == 1]       0.093210   0.008631  10.799  < 2e-16 ***
## nmc$bmi[nmc$pa == 1]       0.542069   0.262255   2.067   0.0387 *
## nmc$fitness[nmc$pa == 1] -0.284707   0.111606  -2.551   0.0107 *
## nmc$smoke[nmc$pa == 1]     0.170012   0.153308   1.109   0.2674
## ---
## 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 2413  degrees of freedom
## AIC: 752.78
##
## Number of Fisher Scoring iterations: 7

pstima.pa1 <- fit.pa1$fitted.values

#Plot
plot(nmc$age[nmc$pa==0], nmc$cvd[nmc$pa==0],
     xlab="Age", ylab="CVD", col="blue")
points(nmc$age[nmc$pa==1], nmc$cvd[nmc$pa==1], col="red")
points(sort(nmc$age[nmc$pa==0]), pstima.pa0[order(nmc$age[nmc$pa==0])], col="blue")
points(sort(nmc$age[nmc$pa==1]), pstima.pa1[order(nmc$age[nmc$pa==1])], col="red")
legend(x="left",
       legend=c("High-level exerciser", "Low-level exerciser"),
       lty=c(1, 1), col=c("blue","red"), lwd = 1)
```



Il grafico ci mostra come ci sia una concentrazione maggiore di casi negli individui più anziani che praticano esercizi di alto livello rispetto a quelli che praticano esercizi di basso livello.

7 Selezione del Modello

Per la selezione del modello utilizziamo i metodi Backward, Forward e Both basati sui criteri di penalizzazione AIC e BIC.
Utilizziamo la formula base con intercetta e il modello che comprende tutte le variabili.

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

7.1 Forward

Verifichiamo le formule della procedura Forward con AIC e BIC.

7.1.1 AIC

```
#Forward: AIC
forward.AIC <- step(fit.0, scope=formula(fit.all),
                    direction="forward", trace=0, k=2)
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$alc, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5841  -0.3388  -0.1937  -0.0950   3.6467
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.938561   0.208732 -38.032  < 2e-16 ***
## nmc$age      0.092451   0.002441  37.869  < 2e-16 ***
## nmc$sexMale  0.788736   0.054368  14.507  < 2e-16 ***
## nmc$fitness -0.185841   0.030364  -6.120 9.33e-10 ***
## nmc$smoke    0.125336   0.045174   2.775  0.00553 **
## nmc$bmi      0.226349   0.096833   2.338  0.01941 *
## nmc$alc     -0.063108   0.035516  -1.777  0.07559 .
## ---
```

```
## 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: 10887  on 33320  degrees of freedom
## AIC: 10901
##
## Number of Fisher Scoring iterations: 7
```

7.1.2 BIC

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

7.2 Backward

Verifichiamo le formule della procedura Backward con AIC e BIC.

7.2.1 AIC

```
#Backward: AIC
backward.AIC <- step(fit.all, direction="backward", trace=0, k=2)
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$alc, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5841  -0.3388  -0.1937  -0.0950   3.6467
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.938561   0.208732 -38.032  < 2e-16 ***
## nmc$sexMale  0.788736   0.054368  14.507  < 2e-16 ***
## nmc$age      0.092451   0.002441  37.869  < 2e-16 ***
## nmc$bmi      0.226349   0.096833   2.338  0.01941 *
## nmc$fitness -0.185841   0.030364  -6.120 9.33e-10 ***
## nmc$smoke    0.125336   0.045174   2.775  0.00553 **
## nmc$alc     -0.063108   0.035516  -1.777  0.07559 .
## ---
## 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: 10887  on 33320  degrees of freedom
## AIC: 10901
##
## Number of Fisher Scoring iterations: 7
```

7.2.2 BIC

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

## Start:  AIC=10970.51
## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness + nmc$pa +
##     nmc$smoke + nmc$alc
##
##               Df Deviance   AIC
## - nmc$pa       1    10887 10960
## - nmc$alc      1    10890 10963
## - nmc$bmi      1    10892 10965
## - nmc$smoke    1    10895 10968
## <none>              10887 10970
## - nmc$fitness  1    10922 10995
## - nmc$sex      1    11097 11170
## - nmc$age      1    13047 13120
##
## Step:  AIC=10960.21
## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness + nmc$smoke +
##     nmc$alc
##
##               Df Deviance   AIC
## - nmc$alc      1    10890 10953
## - nmc$bmi      1    10893 10955
## - nmc$smoke    1    10895 10957
## <none>              10887 10960
## - nmc$fitness  1    10925 10987
## - nmc$sex      1    11098 11160
## - nmc$age      1    13049 13111
##
## Step:  AIC=10952.95
## nmc$cvd ~ nmc$sex + nmc$age + nmc$bmi + nmc$fitness + nmc$smoke
##
##               Df Deviance   AIC
## - nmc$bmi      1    10896 10948
## - nmc$smoke    1    10896 10948
## <none>              10890 10953
## - nmc$fitness  1    10929 10981
## - nmc$sex      1    11098 11150
## - nmc$age      1    13059 13111
##
## Step:  AIC=10947.98
```

```
## nmc$cvd ~ nmc$sex + nmc$age + nmc$fitness + nmc$smoke
##
##               Df Deviance   AIC
## - nmc$smoke    1    10902 10943
## <none>              10896 10948
## - nmc$fitness  1    10943 10984
## - nmc$sex      1    11101 11143
## - nmc$age      1    13072 13114
##
## Step:  AIC=10943.43
## nmc$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
```

7.3 Both

Verifichiamo 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",
                 trace=0, k=2)
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$alc, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5841  -0.3388  -0.1937  -0.0950   3.6467
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.938561   0.208732 -38.032  < 2e-16 ***
## nmc$age      0.092451   0.002441  37.869  < 2e-16 ***
## nmc$sexMale  0.788736   0.054368  14.507  < 2e-16 ***
## nmc$fitness -0.185841   0.030364  -6.120 9.33e-10 ***
## nmc$smoke    0.125336   0.045174   2.775  0.00553 **
## nmc$bmi      0.226349   0.096833   2.338  0.01941 *
## nmc$alc     -0.063108   0.035516  -1.777  0.07559 .
## ---
## 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: 10887  on 33320  degrees of freedom
```

```
## AIC: 10901
##
## Number of Fisher Scoring iterations: 7
```

7.3.2 BIC

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

- la procedura FORWARD, BACKWARD e BOTH AIC:
  CVD $\sim$ Age + Sex + Fitness + Smoke + BMI + Alchol
- la procedura FORWARD, BACKWARD e BOTH BIC:
  CVD $\sim$ Age + Sex + Fitness

8 Modelli Grafici

8.1 Tabella Frequenze

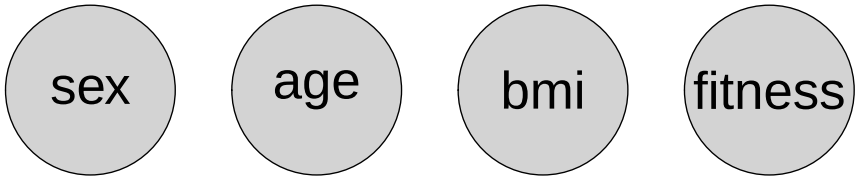
```
#Tabella
ftable(sex+bmi+pa ~ smoke+alc+fitness, nmc)
```

```
##                   sex Female                Male
##                   bmi      0         1        0         1
##                   pa       0    1    0    1    0    1    0    1
## smoke alc fitness
## 1     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
## 2     1   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
## 3     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
```

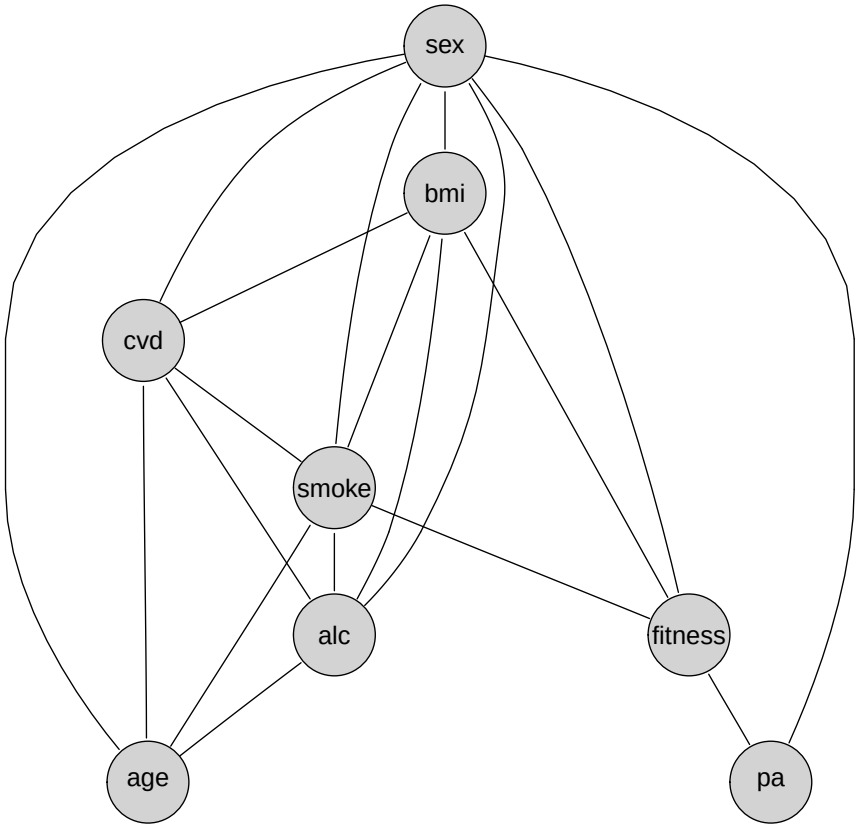
8.2 Grafici

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

## Model: A dModel with 4 variables
##  -2logL    :      411657.71 mdim :   80 aic :    411817.71
##  ideviance :          -0.00 idf  :    0 bic :    412490.84
##  deviance  :        5912.02 df   : 1419

plot(mInit)
```

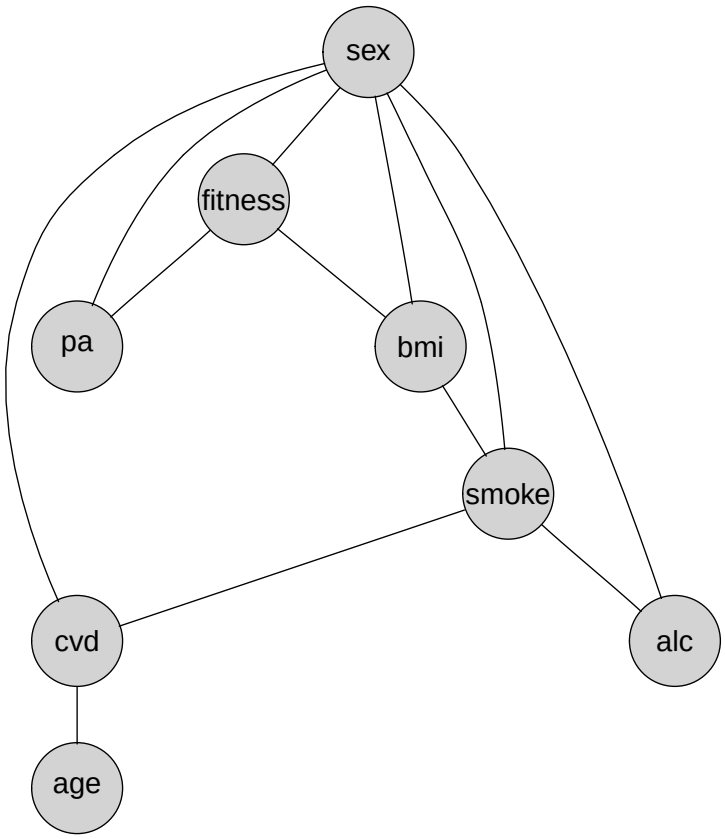



```
#Formula modello saturo
sat.nmc <- dmod(~.^., data=nmc)

#Formula modello indipendenza completa
ind.nmc <- dmod(~.^1, data=nmc)
```

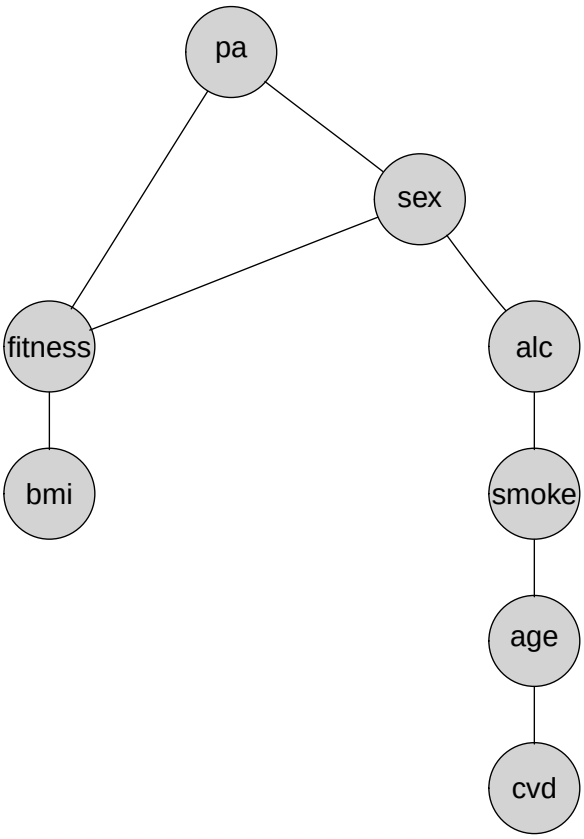
```
#AIC
m.sat.AIC <- stepwise(sat.nmc)
plot(m.sat.AIC)
```



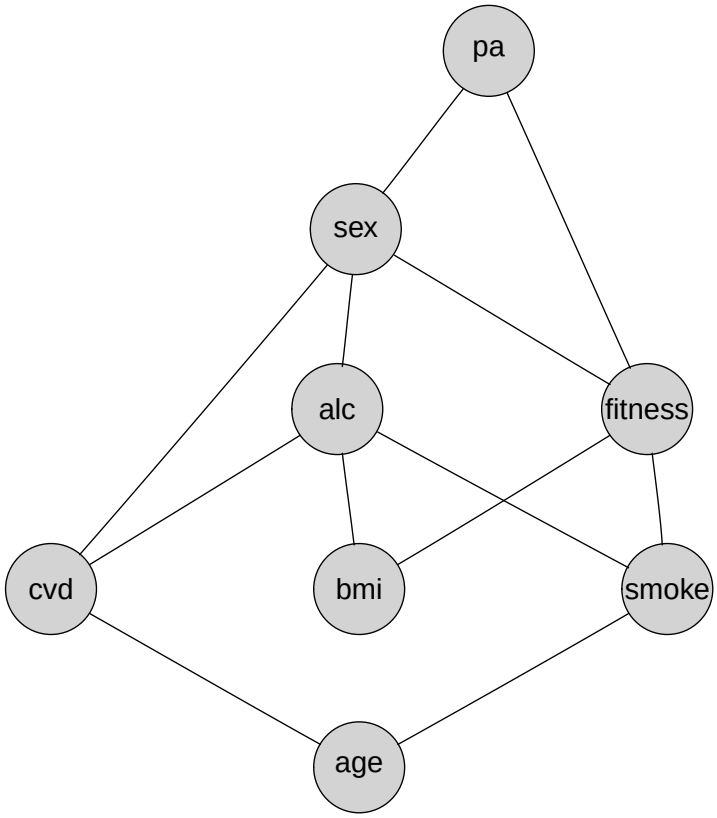

```
#BIC
m.sat.BIC <- stepwise(sat.nmc, k=log(length(nmc$cvd)))
plot(m.sat.BIC)
```




```
m.decomposable <- stepwise(ind.nmc, k=log(length(nmc$cvd)),
                           type="decomposable",
                           direction = "forward",
                           details=0)
m.unrestricted <- stepwise(ind.nmc, k=log(length(nmc$cvd)),
                           type="unrestricted",
                           direction = "forward",
                           details=0)
plot(m.decomposable)
```

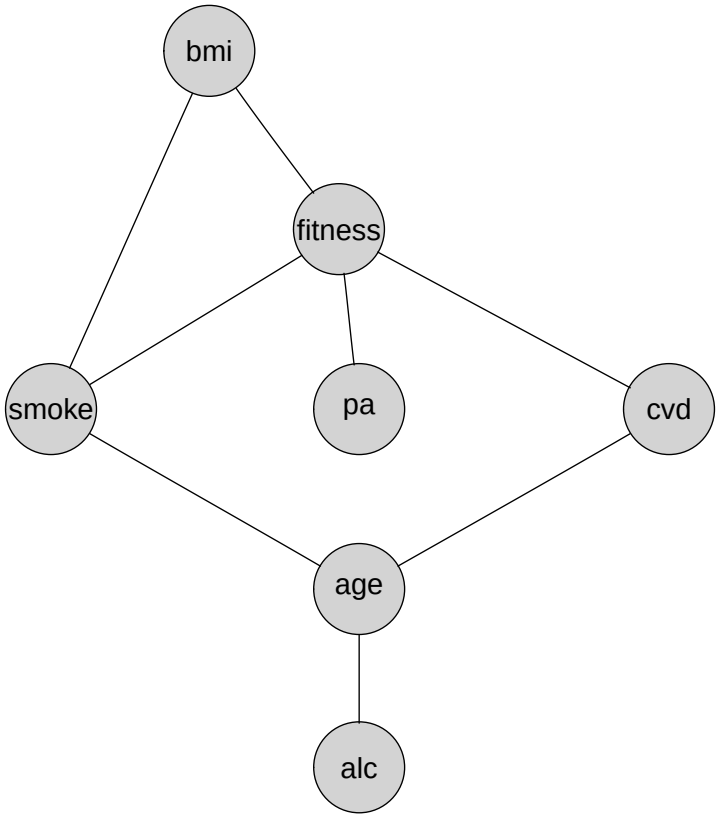



```
plot(m.unrestricted)
```

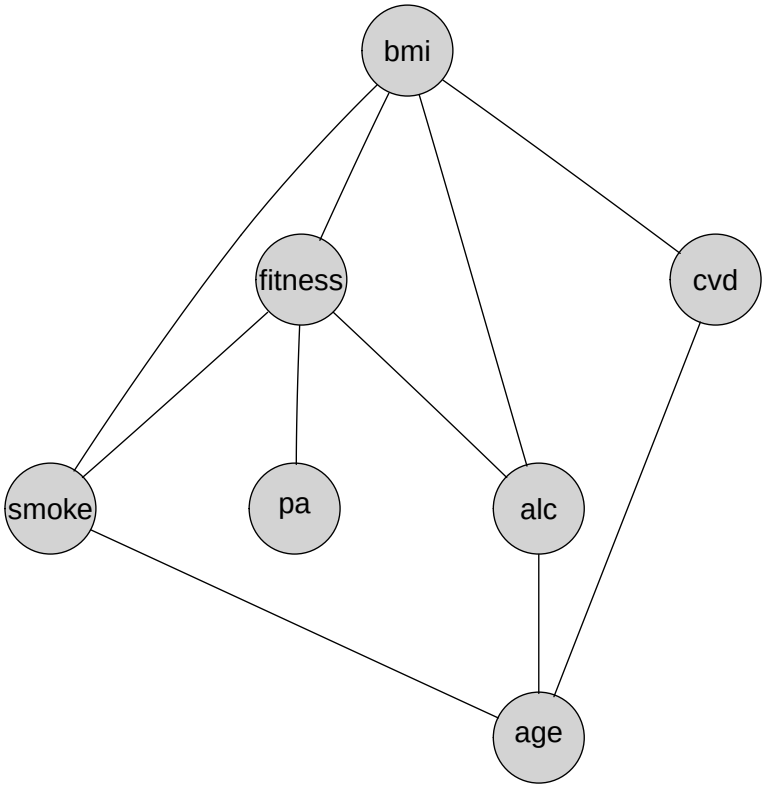



```
#ci.nmc <- ciTest(nmc, set=c("Sex"))
```

```
#Considero grafici in base al genere
#Male
male.data <- nmc[(nmc$sex=='Male'), c(2:8)]
nmc.male <- dmod(~.^., data=male.data)
m.male <- stepwise(nmc.male, type='unrestricted')
plot(m.male)
```




```
#Female
female.data <- nmc[(nmc$sex=='Female'), c(2:8)]
nmc.female <- dmod(~.^., data=female.data)
m.female <- stepwise(nmc.female, type='unrestricted')
plot(m.female)
```


bmi
fitness
cvd
smoke
pa
alc
age