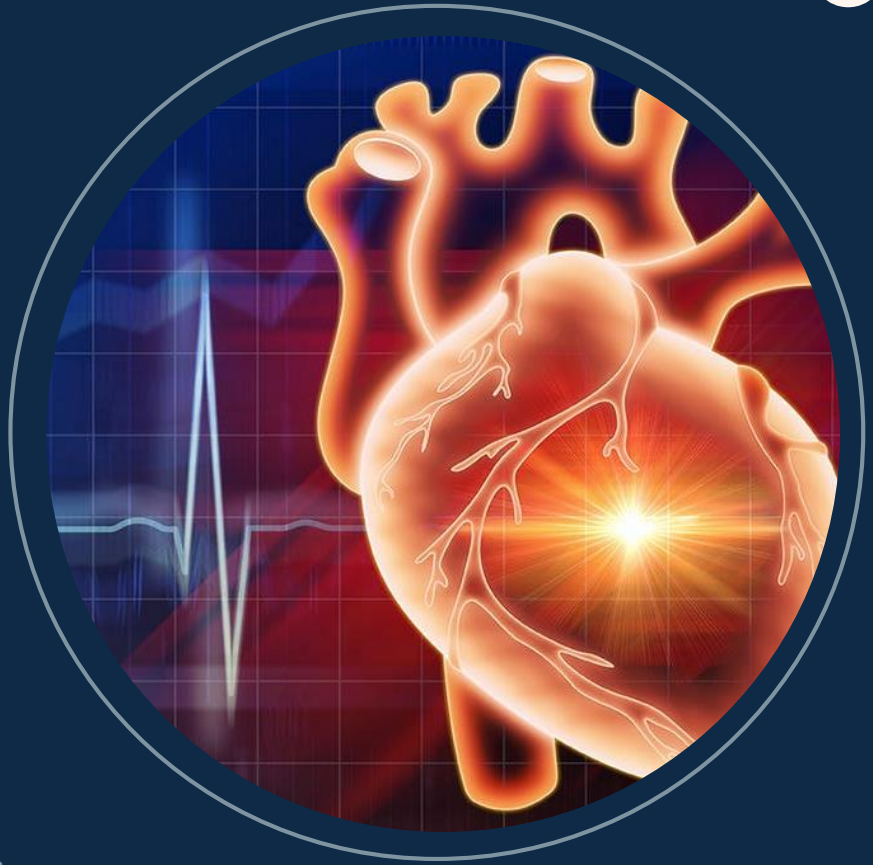
Malattie Cardiovascolari

Previsione attraverso un modello di
Regressione Logistica Multipla



ANGELICA IACOVELLI
HELENA DELL'ANNA
LORENZO CARLASSARA
MARTINA DONZELLI

Contesto:

Analisi di un campione di uomini in una regione ad alto rischio di malattie cardiache del Western Cape, Sudafrica.

Dati:

Sbp	Tobacco	Ldl	Adiposity	Famhist	TypeA	Obesity	Alcohol	Age	Chd
Pressione Arteriosa Sistolica (mmHg)	Quantità di Tabacco attuale (Kg)	Lipoproteine a bassa densità (mmol/L)	Valori di Adiposità (%)	Storico familiare di malattie cardiache (Absent, Present)	Personalità di tipo A (%)	Obesità valutata con IMC (kg/m^2)	Quantità di Alcol attuale (%)	Età del paziente	Assenza o presenza della malattia coronarica { 0,1 }

Obiettivo:

Costruzione di un modello in grado di prevedere l'insorgenza di malattie coronariche, tramite lo studio dei seguenti dati, per ogni paziente analizzato.

VISUALIZZAZIONE **VARIABILI QUANTITATIVE** & **COVARIATA CATEGORICA**

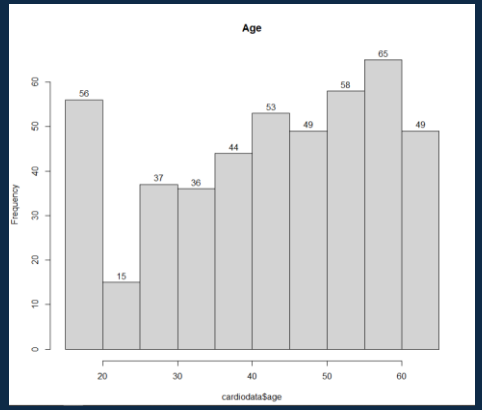
```
      ind         sbp     tobacco         ldl   adiposity     famhist       typea     obesity
"integer"   "integer"    "double"    "double"    "double" "character"   "integer"    "double"
  alcohol         age         chd
 "double"   "integer"   "integer"
```

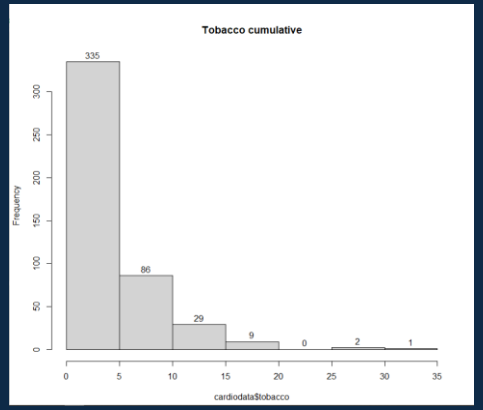
```
'data.frame':   462 obs. of  11 variables:
 $ ind      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ sbp      : int  160 144 118 170 134 132 142 114 114 132 ...
 $ tobacco  : num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
 $ ldl      : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
 $ adiposity: num  23.1 28.6 32.3 38 27.8 ...
 $ famhist  : Factor w/ 2 levels "Absent","Present": 2 1 2 2 2 2 1 2 2 2 ...
 $ typea    : int  49 55 52 51 60 62 59 62 49 69 ...
 $ obesity  : num  25.3 28.9 29.1 32 26 ...
 $ alcohol  : num  97.2 2.06 3.81 24.26 57.34 ...
 $ age      : int  52 63 46 58 49 45 38 58 29 53 ...
 $ chd      : int  1 1 0 1 1 0 0 1 0 1 ...
```

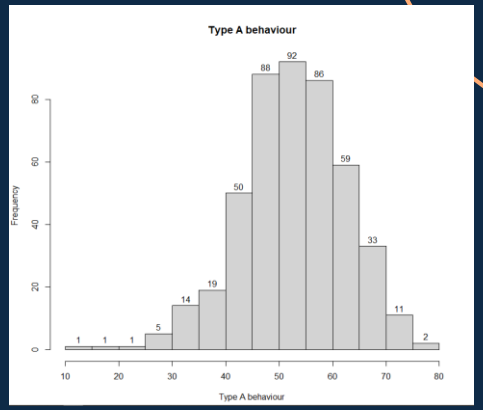
```
Absent Present
   270     192
```

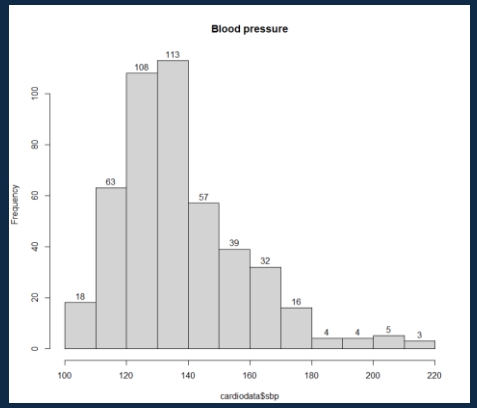
La **covariata famhist** indica per ogni paziente la presenza o meno di problemi cardiaci in famiglia, quindi specifichiamo a R di trattarla come una **discreta su due livelli**.

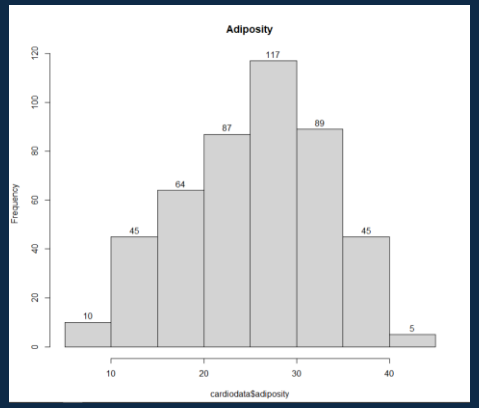
VISUALIZZAZIONE **VARIABILI QUANTITATIVE**

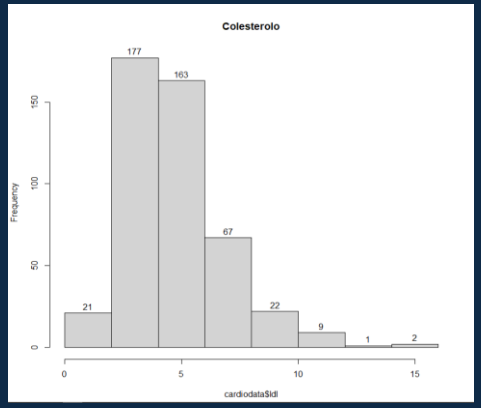








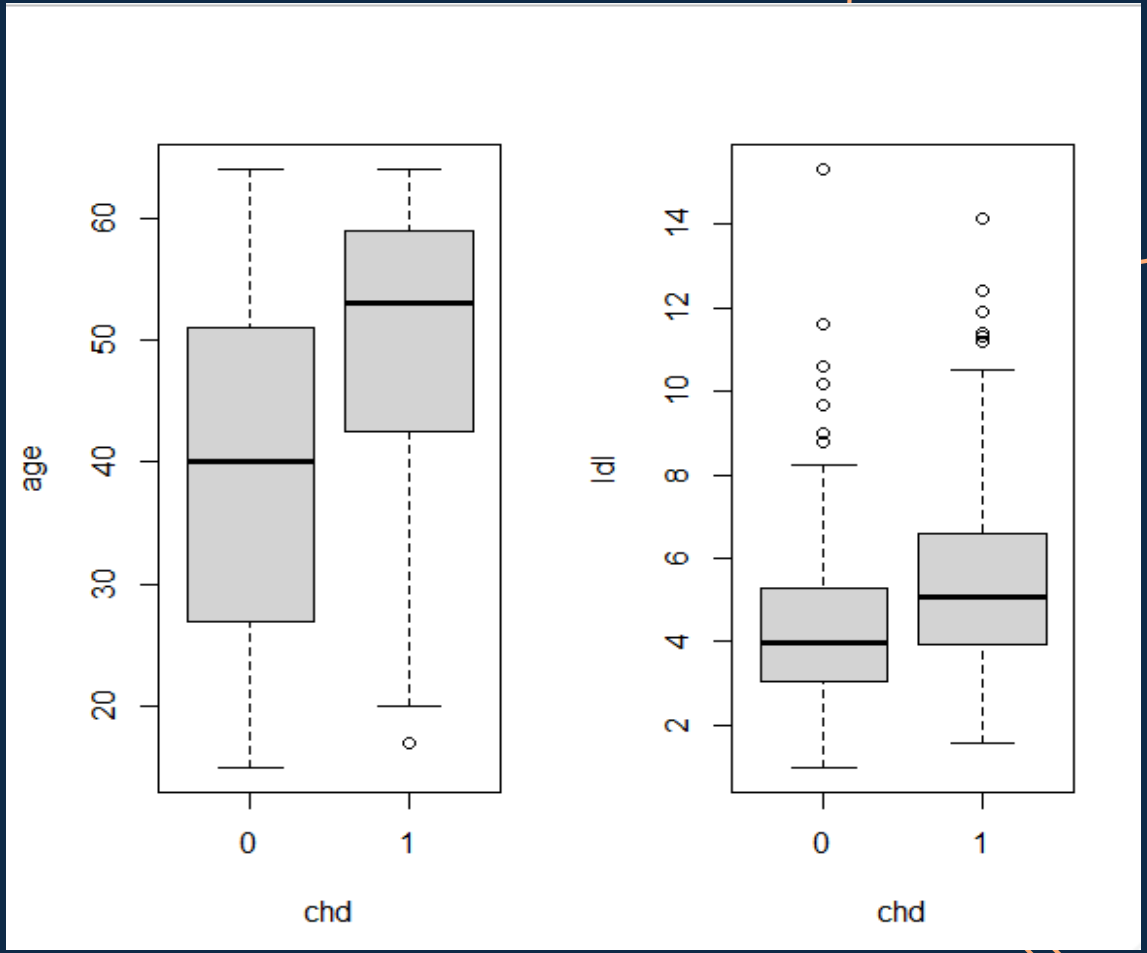


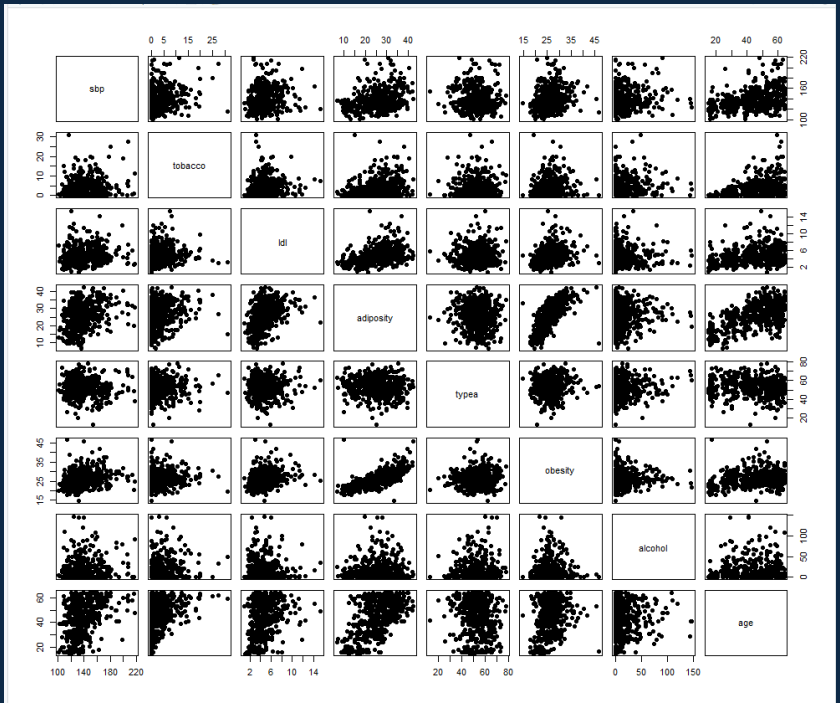


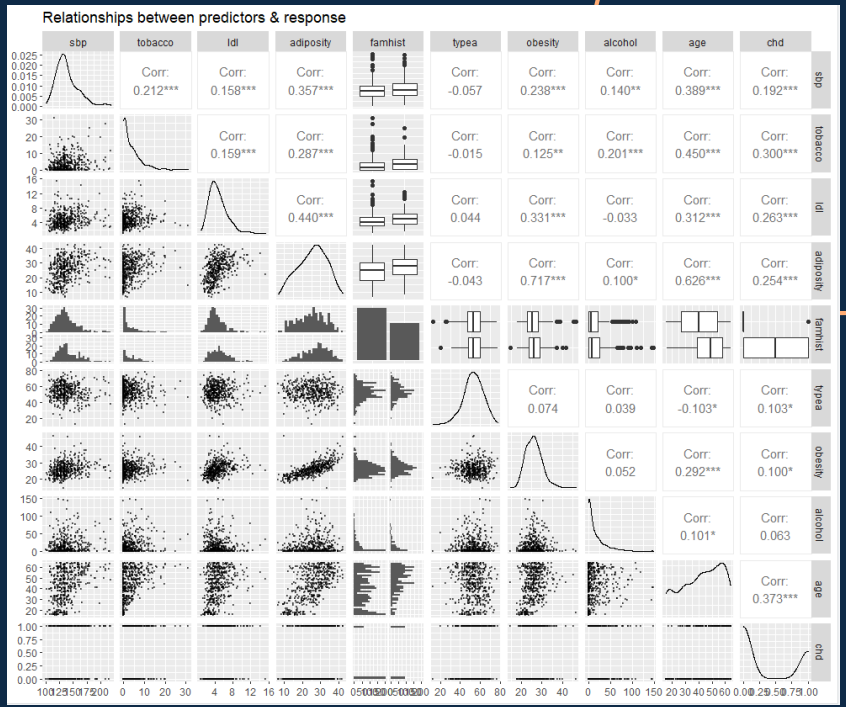
ANALISI GRAFICA

1. BOXPLOT

- **CHD vs AGE**
  Grande differenza
  tra popolazioni
- **CHD vs LDL**
  Buona differenza
  tra popolazioni,
  ma presenza di outliers







2. PAIRS e GGPAIRS

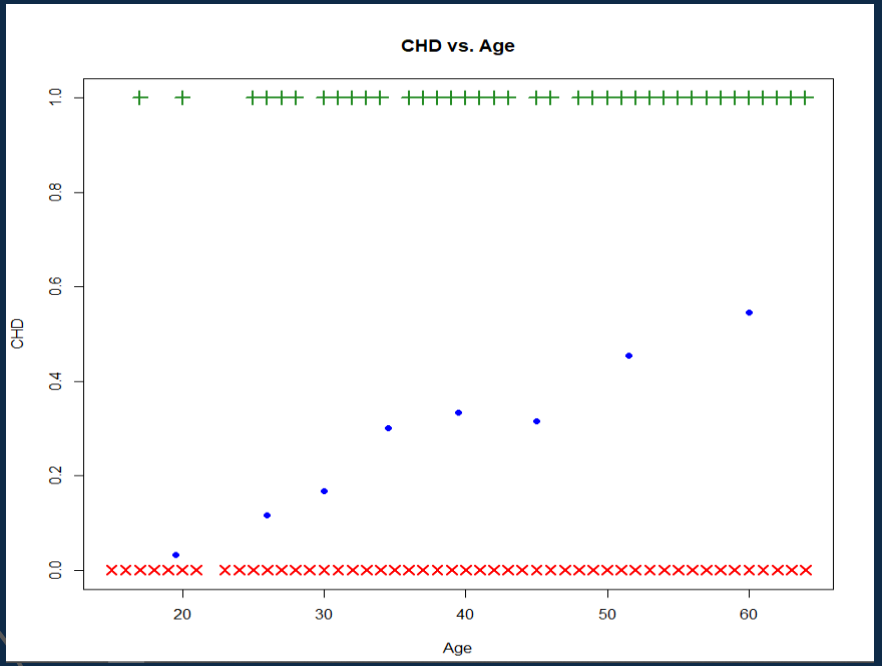
- Andamenti riconoscibili
- Valori alti di correlazione

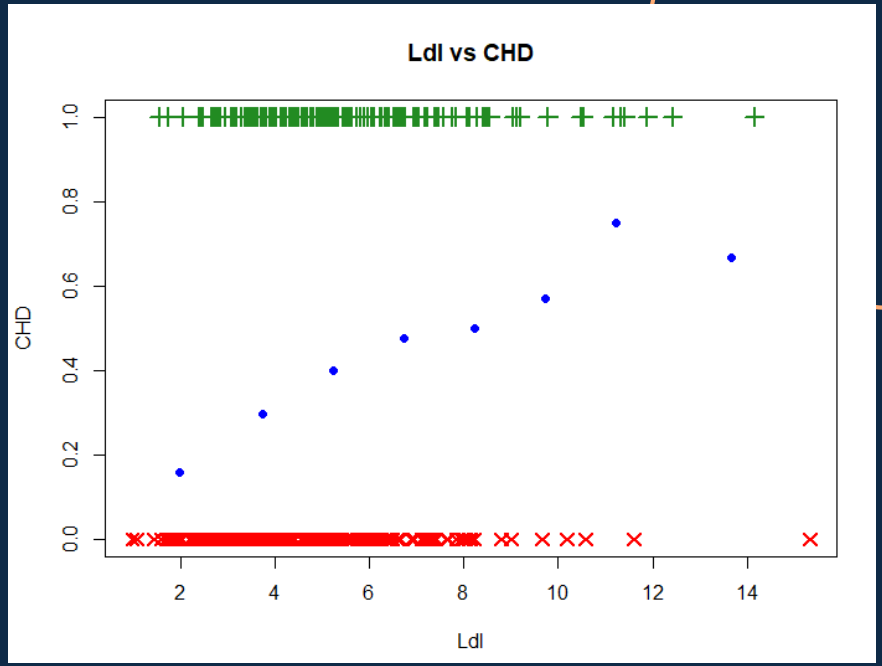
2

Una nuova domanda

Modellino finale

- ADIPOSITY
  vs AGE + ALCOHOL+ OBESITY





3. ANALISI GRAFICA

AGE vs CHD

Andamento piuttosto regolare, quasi lineare

LDL vs CHD

Andamento abbastanza regolare, fino a circa 9;
Poche osservazioni sui valori superiori

Fittiamo il **modello di regressione logistica:**

```
Call:
glm(formula = chd ~ -ind + sbp + tobacco + ldl + adiposity +
    famhist + typea + obesity + alcohol + age, family = binomial(link = logit),
    data = cardiodata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7781  -0.8213  -0.4387   0.8889   2.5435

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.1507209  1.3082600  -4.701 2.58e-06 ***
sbp             0.0065040  0.0057304   1.135 0.256374
tobacco         0.0793764  0.0266028   2.984 0.002847 **
ldl             0.1739239  0.0596617   2.915 0.003555 **
adiposity       0.0185866  0.0292894   0.635 0.525700
famhistPresent  0.9253704  0.2278940   4.061 4.90e-05 ***
typea           0.0395950  0.0123202   3.214 0.001310 **
obesity        -0.0629099  0.0442477  -1.422 0.155095
alcohol         0.0001217  0.0044832   0.027 0.978350
age             0.0452253  0.0121298   3.728 0.000193 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.14  on 452  degrees of freedom
AIC: 492.14
```

Significatività bassa di:
- Alcohol
- Adiposity
- Sbp
- Obesity

Residual deviance< Null variance

AIC: 492.14

Cerchiamo di migliorare il modello eliminando la variabile meno influente : **Alcohol**

```
Call:
glm(formula = chd ~ -ind - alcohol + sbp + tobacco + ldl + adiposity +
    famhist + typea + obesity + age, family = binomial(link = logit),
    data = cardiodata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7795  -0.8207  -0.4391   0.8882   2.5427

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.150058   1.308008  -4.702 2.58e-06 ***
sbp             0.006524   0.005685   1.148 0.251149
tobacco         0.079515   0.026114   3.045 0.002327 **
ldl             0.173770   0.059393   2.926 0.003436 **
adiposity       0.018631   0.029245   0.637 0.524079
famhistPresent  0.925831   0.227266   4.074 4.63e-05 ***
typea           0.039604   0.012316   3.216 0.001302 **
obesity        -0.062957   0.044216  -1.424 0.154489
age             0.045191   0.012061   3.747 0.000179 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.14  on 453  degrees of freedom
AIC: 490.14
```

Significatività bassa di:
- Adiposity
- Sbp
- Obesity

Residual deviance< Null variance

AIC: 490.14

Cerchiamo di migliorare il modello eliminando la variabile meno influente : **Adiposity**

```
Call:
glm(formula = chd ~ -ind - alcohol - adiposity + sbp + tobacco +
    ldl + famhist + typea + obesity + age, family = binomial(link = logit),
    data = cardiodata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8245  -0.8189  -0.4415   0.8892   2.5530

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.416927   1.240101  -5.175 2.28e-07 ***
sbp             0.006780   0.005683   1.193  0.23286
tobacco         0.079886   0.026157   3.054  0.00226 **
ldl             0.182102   0.058077   3.136  0.00172 **
famhistPresent  0.924464   0.227061   4.071 4.67e-05 ***
typea           0.038966   0.012266   3.177  0.00149 **
obesity        -0.042200   0.029437  -1.434  0.15169
age             0.048927   0.010556   4.635 3.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.55  on 454  degrees of freedom
AIC: 488.55
```

Significatività bassa di:
- Sbp
- Obesity

Residual deviance< Null variance

AIC: 488.55

Cerchiamo di migliorare il modello eliminando la variabile meno influente : **Sbp**

```
Call:
glm(formula = chd ~ -ind - alcohol - adiposity - sbp + tobacco +
    ldl + famhist + typea + obesity + age, family = binomial(link = logit),
    data = cardiodata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8941  -0.8164  -0.4329   0.8966   2.5442

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.70273    1.07640  -5.298 1.17e-07 ***
tobacco      0.07999    0.02598   3.079  0.00208 **
ldl          0.18372    0.05818   3.158  0.00159 **
famhistPresent  0.91610    0.22645   4.046 5.22e-05 ***
typea        0.03827    0.01222   3.133  0.00173 **
obesity     -0.03760    0.02910  -1.292  0.19638
age          0.05211    0.01024   5.087 3.63e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 473.98  on 455  degrees of freedom
AIC: 487.98
```

Significatività bassa di:
- Obesity

Residual deviance< Null variance

AIC: 487.98

Cerchiamo di migliorare il modello eliminando la variabile meno influente : **Obesity**

```
Call:
glm(formula = chd ~ -ind - alcohol - adiposity - sbp - obesity +
    tobacco + ldl + famhist + typea + age, family = binomial(link = logit),
    data = cardiodata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9165  -0.8054  -0.4430   0.9329   2.6139

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.44644    0.92087  -7.000 2.55e-12 ***
tobacco       0.08038    0.02588   3.106  0.00190 **
ldl           0.16199    0.05497   2.947  0.00321 **
famhistPresent  0.90818    0.22576   4.023 5.75e-05 ***
typea         0.03712    0.01217   3.051  0.00228 **
age           0.05046    0.01021   4.944 7.65e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 475.69  on 456  degrees of freedom
AIC: 487.69
```

Tutte le covariate sono significative

Residual deviance< Null variance

AIC: 487.69

AIC diminuisce da 492,14 a 487,69 → Il modello ridotto non è meno informativo del completo

Conferma con **Stepwise**:

```
Start:  AIC=492.14
chd ~ -ind + sbp + tobacco + ldl + adiposity + famhist + typea +
    obesity + alcohol + age

            Df Deviance    AIC
- alcohol    1   472.14 490.14
- adiposity  1   472.55 490.55
- sbp        1   473.44 491.44
<none>           472.14 492.14
- obesity    1   474.23 492.23
- ldl        1   481.07 499.07
- tobacco    1   481.67 499.67
- typea      1   483.05 501.05
- age        1   486.53 504.53
- famhist    1   488.89 506.89

Step:  AIC=490.14
chd ~ sbp + tobacco + ldl + adiposity + famhist + typea + obesity +
    age

            Df Deviance    AIC
- adiposity  1   472.55 488.55
- sbp        1   473.47 489.47
<none>           472.14 490.14
- obesity    1   474.24 490.24
- ldl        1   481.15 497.15
- tobacco    1   482.06 498.06
- typea      1   483.06 499.06
- age        1   486.64 502.64
- famhist    1   488.99 504.99

Step:  AIC=488.55
chd ~ sbp + tobacco + ldl + famhist + typea + obesity + age

          Df Deviance    AIC
- sbp      1   473.98 487.98
<none>           472.55 488.55
- obesity  1   474.65 488.65
- tobacco  1   482.54 496.54
- ldl      1   482.95 496.95
- typea    1   483.19 497.19
- famhist  1   489.38 503.38
- age      1   495.48 509.48

Step:  AIC=487.98
chd ~ tobacco + ldl + famhist + typea + obesity + age

          Df Deviance    AIC
- obesity  1   475.69 487.69
<none>           473.98 487.98
- tobacco  1   484.18 496.18
- typea    1   484.30 496.30
- ldl      1   484.53 496.53
- famhist  1   490.58 502.58
- age      1   502.11 514.11

Step:  AIC=487.69
chd ~ tobacco + ldl + famhist + typea + age

          Df Deviance    AIC
<none>           475.69 487.69
- ldl      1   484.71 494.71
- typea    1   485.44 495.44
- tobacco  1   486.03 496.03
- famhist  1   492.09 502.09
- age      1   502.38 512.38
```

Modello finale identico

```
Call:  glm(formula = chd ~ tobacco + ldl + famhist + typea + age, family = binomial(link = logit),
    data = cardiodata)

Coefficients:
   (Intercept)         tobacco             ldl  famhistPresent           typea             age
      -6.44644         0.08038         0.16199         0.90818         0.03712         0.05046

Degrees of Freedom: 461 Total (i.e. Null);  456 Residual
Null Deviance:      596.1
Residual Deviance: 475.7        AIC: 487.7
```

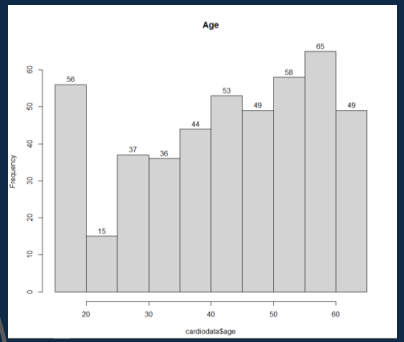
Confronto tramite **ANOVA**:

```
Analysis of Deviance Table

Model 1: chd ~ -ind + sbp + tobacco + ldl + adiposity + famhist + typea +
    obesity + alcohol + age
Model 2: chd ~ -ind - alcohol - adiposity - sbp - obesity + tobacco +
    ldl + famhist + typea + age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       452     472.14
2       456     475.69 -4  -3.5455    0.471
```

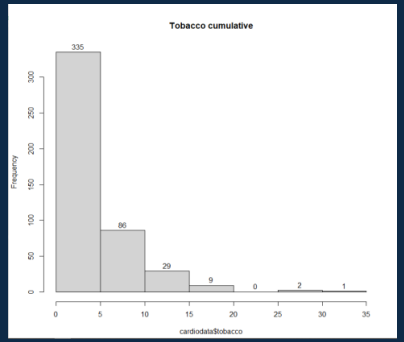
Differenza non statisticamente significativa→ Il modello ridotto non è meno informativo del completo

Variabili utilizzate nel **modello** più **performante**

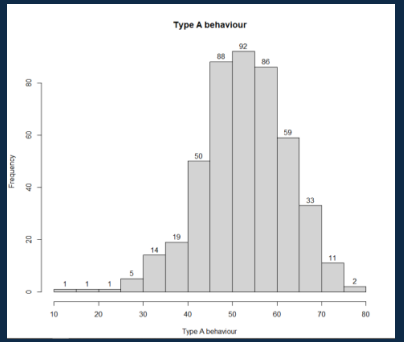
```
     age
1.286985
```
ODDS RATIO



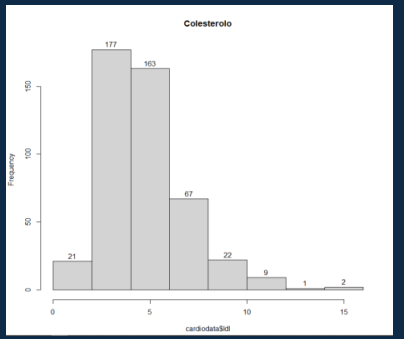

```
 tobacco
1.494627
```
ODDS RATIO




```
   typea
1.449404
```
ODDS RATIO




```
    ldl
1.382624
```
ODDS RATIO



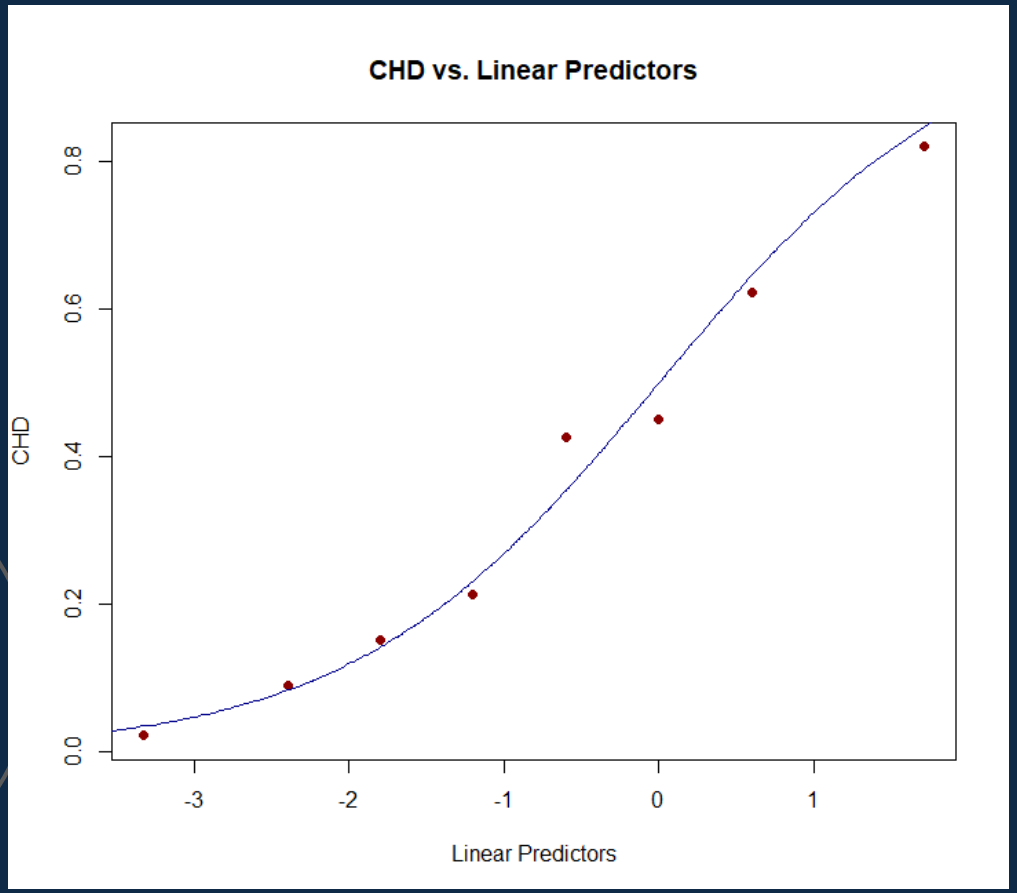

```
famhistPresent
      2.479793
```
ODDS RATIO

Si può osservare che per i pazienti con un parente positivo in famiglia il rischio di avere un problema cardiaco è quasi 2.5 superiore rispetto a un paziente che non ha nessun parente con questo disturbo!

VERIFICA DELLA **BONTÀ DEL MODELLO**

```
Logistic Regression Model

 lrm(formula = cardiodata$chd ~ fitted(mod5), x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                             Ratio Test           Indexes          Indexes
 Obs           462    LR chi2     116.82    R2       0.308    C        0.792
  0            302    d.f.             1    g        1.352    Dxy      0.584
  1            160    Pr(> chi2) <0.0001    gr       3.867    gamma    0.584
 max |deriv| 4e-07                          gp       0.262    tau-a    0.265
                                            Brier    0.173

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -2.5571 0.2427 -10.54 <0.0001
 mod5       5.0776 0.5361   9.47 <0.0001
```

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  mod$y, fitted(mod)
X-squared = 2.2243, df = 8, p-value = 0.9734
```



- Divisione intervallo
- Visualizzazione grafica
- Coerenza della **Link Function**

MATRICI DI CONFUSIONE

	Valori PREDETTI	
Valori REALI	TP	FP
	FN	TN

valori.reali	valori.predetti 0	1
0	256	46
1	73	87

SOGLIA = 0,5

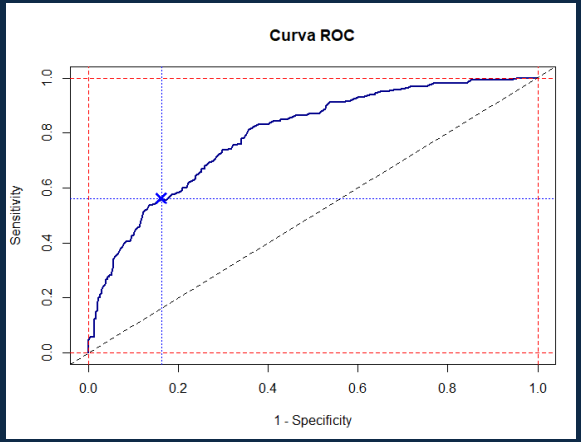
Accuracy ≃ 0,74 (Misclassification ≃ 0,26)
Sensitivity ≃ 0,56
Specificity ≃ 0,84

valori.reali	valori.predetti 0	1
0	211	91
1	42	118

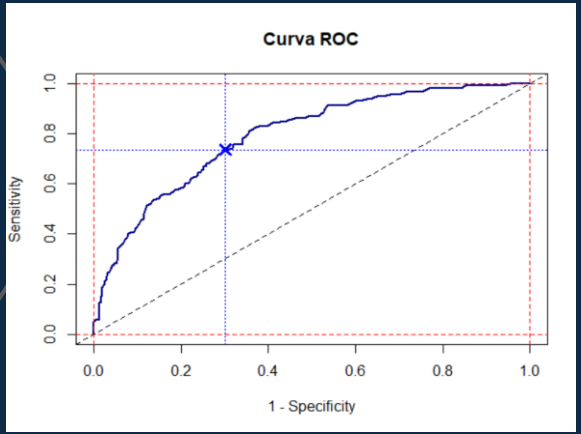
SOGLIA = 0,35 (media di Y)

Accuracy ≃ 0,71 (Misclassification ≃ 0,29)
Sensitivity ≃ 0,74
Specificity ≃ 0,70

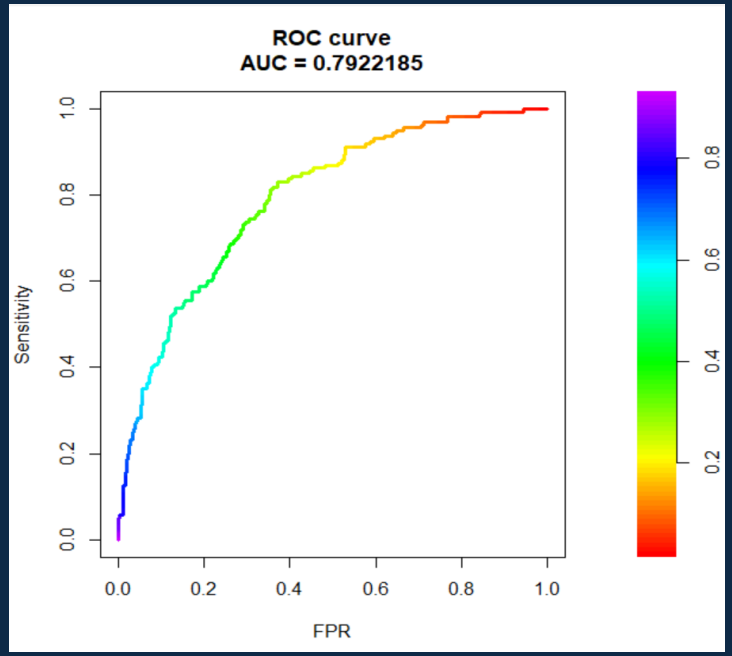
CURVA ROC



SOGLIA = 0,5
Sensitivity ≃ 0,55
Specificity ≃ 0,80



SOGLIA = 0,35 (media di Y)
Sensitivity ≃ 0,75
Specificity ≃ 0,70



Nei **Test diagnostici di screening** :
Preferibile avere più falsi positivi che falsi negativi:
- Prediligo Sensitivity a Specificity

Nei **Test diagnostici di controllo** :
Preferibile avere più falsi negativi che falsi positivi:
- Prediligo Specificity a Sensitivity

Modellino

1. Modello Lineare

ADIPOSITY vs AGE + ALCOHOL + OBESITY

Alcohol risulta non significativa

```
Residuals:
     Min       1Q   Median       3Q      Max
-31.4088  -2.5993   0.1418   2.8061  19.5791

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.101901   1.261353 -10.387   <2e-16 ***
age           0.241469   0.014214  16.989   <2e-16 ***
alcohol       0.007751   0.008124   0.954    0.341
obesity       1.076550   0.049092  21.929   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.247 on 458 degrees of freedom
Multiple R-squared:  0.704,     Adjusted R-squared:  0.7021
F-statistic: 363.1 on 3 and 458 DF,  p-value: < 2.2e-16
```

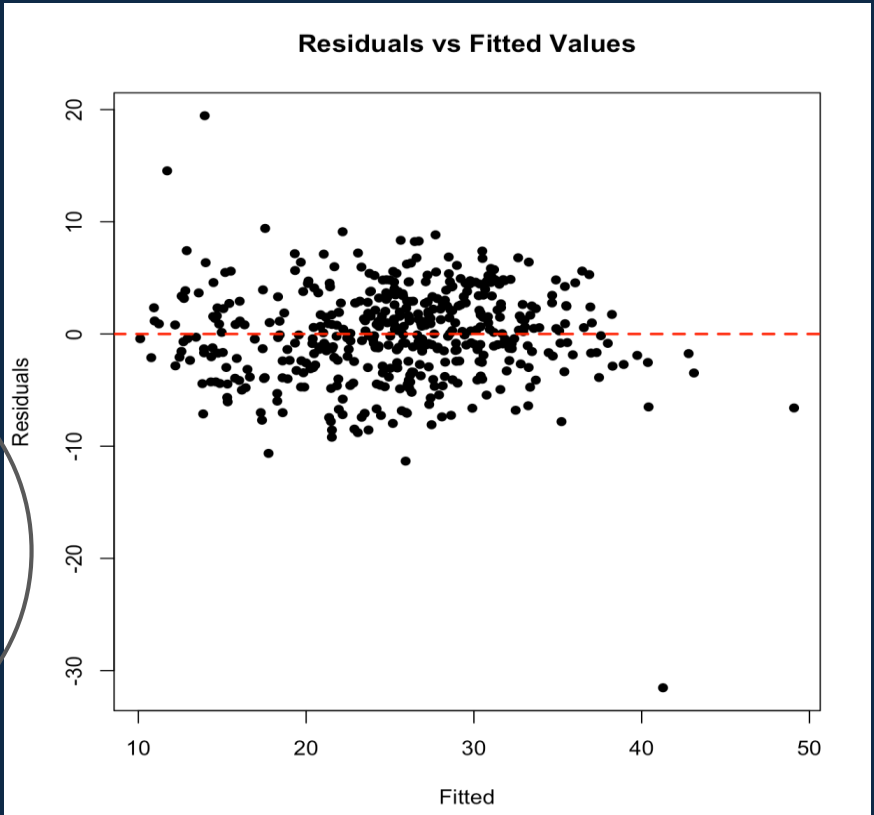
ADIPOSITY vs AGE + OBESITY

Stessi R^2 e p-value

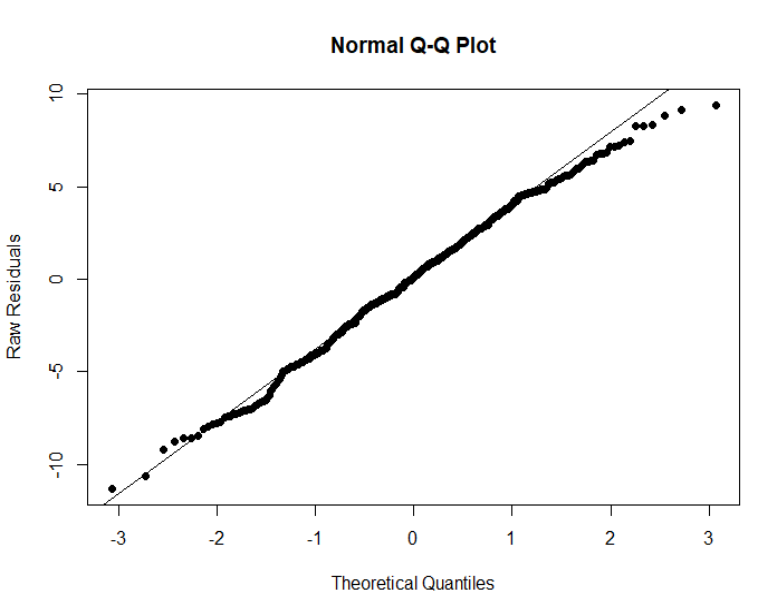
```
Residuals:
     Min       1Q   Median       3Q      Max
-31.5317  -2.5228   0.1244   2.7604  19.4554

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.05046    1.26008  -10.36   <2e-16 ***
age           0.24269    0.01415   17.15   <2e-16 ***
obesity       1.07764    0.04907   21.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.247 on 459 degrees of freedom
Multiple R-squared:  0.7034,    Adjusted R-squared:  0.7021
F-statistic: 544.3 on 2 and 459 DF,  p-value: < 2.2e-16
```

2. Veridicità del modello

Omoschedasticità



Normalità




```
        Shapiro-Wilk normality test

data:  rid2$res
W = 0.99439, p-value = 0.09106
```

CONCLUSIONI

Modello principale

- Modello veritiero per il nostro dataset

Modellino

- Evidenza per affermare la correlazione

Thanks!