Statistica corso progredito: Prova pratica

Cesaro Giada

12 febbraio 2019

1 Descrizione dei dati

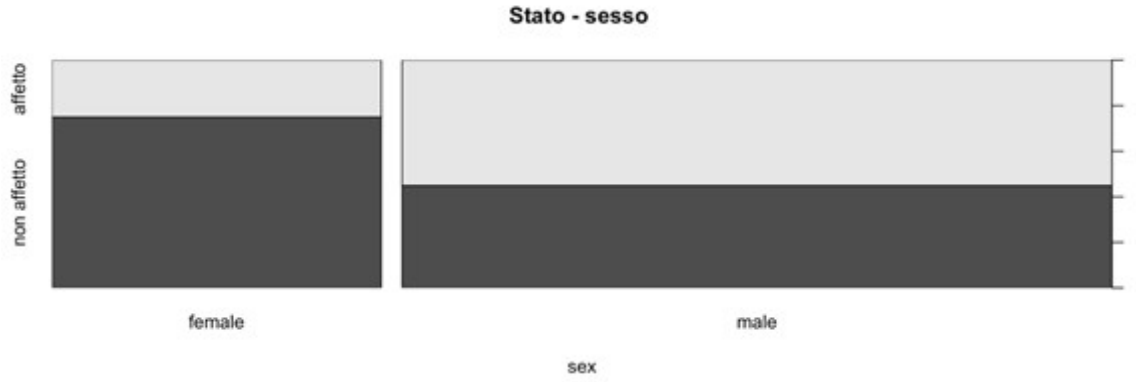
Si analizzino i dati con l'obiettivo di costruire un modello per la previsione della variabile **Target (1=affetto da malattia cardiaca)** in base alle altre 13 caratteristiche rilevate.

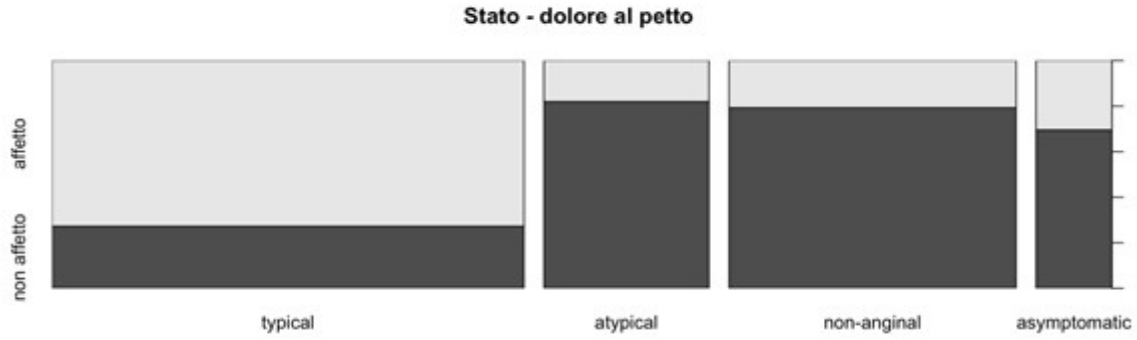
- **Target**: affetto da malattia cardiaca
  - 1= sì
  - 2= no
- **Age**: età in anni
- **Sex:**
  - 1 = maschio
  - 0 = femmina
- **Cp**: tipo di dolore toracico
  - 0 = “typical"
  - 1 = "atypical"
  - 2 = "non-anginal"
  - 3 = "asymptomatic”
- **Trestbps**: pressione sanguigna a riposo in mmHg al momento del ricovero ospedaliero
- **Chol**: livello di colesterolo sierico in mg/dl
- **Fbs**: livello di zuccheri nel sangue >120mg/dL

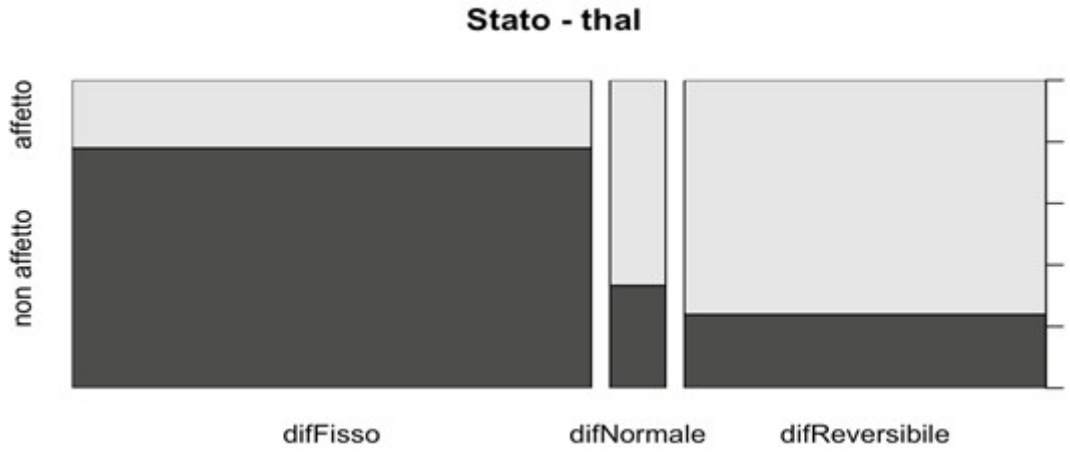
    - 1 = vero
    - 0 = falso
- **Restecg**: elettrocardiogramma a riposo
    - 0 = “normal”
    - 1 = having ST-T wave abnormality
    - 2 = left ventricular hypertrophy
- **Thalach**: massima frequenza cardiaca raggiunta
- **Exang**: l’esercizio fisico provoca angina da sforzo
    - 1 = si
    - 0 = no
- **Oldpeak**: diminuzione del ST in fase di recupero dopo attività fisica
- **Slope**: pendenza del picco del ST relativamente all’attività fisica
    - 0 = “upsloping”
    - 1 = “flat”
    - 2 = “downsloping”
- **Ca**: numero di arterie colorate con fluoroscopia (0-4)
- **Thal**: risultato del thallium stress test
    - 1 = normale
    - 2 = difetto fisso
    - 3 = difetto reversibile.

2 Analisi dei dati

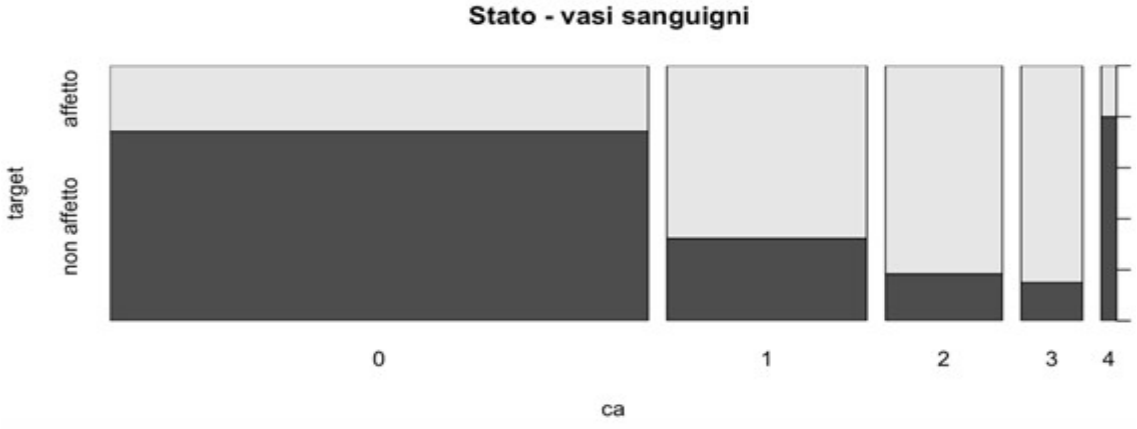
Analizzo le relazioni tra **Target** e le altre variabili rilevate prese singolarmente. Trasformo innanzitutto le opportune variabili in fattori ed assegno nomi interpretabili ai livelli.

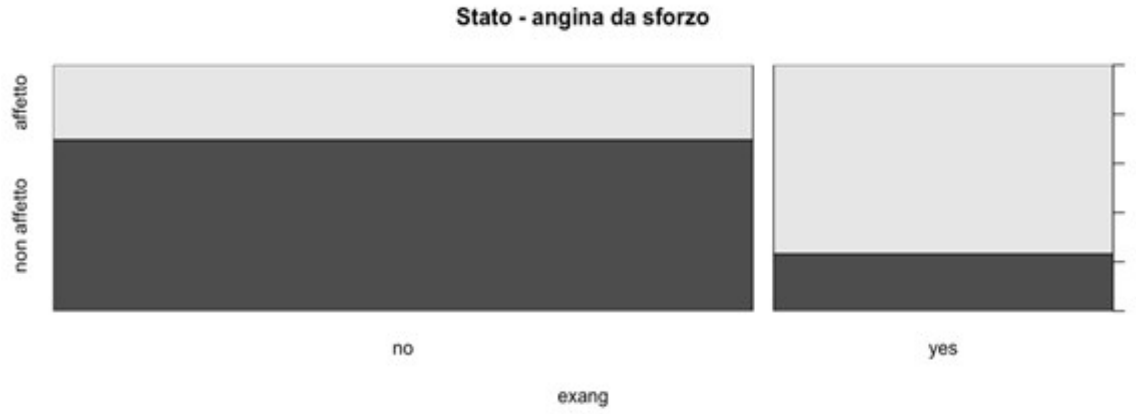

Stato - sesso
affetto
non affetto
female
male
sex

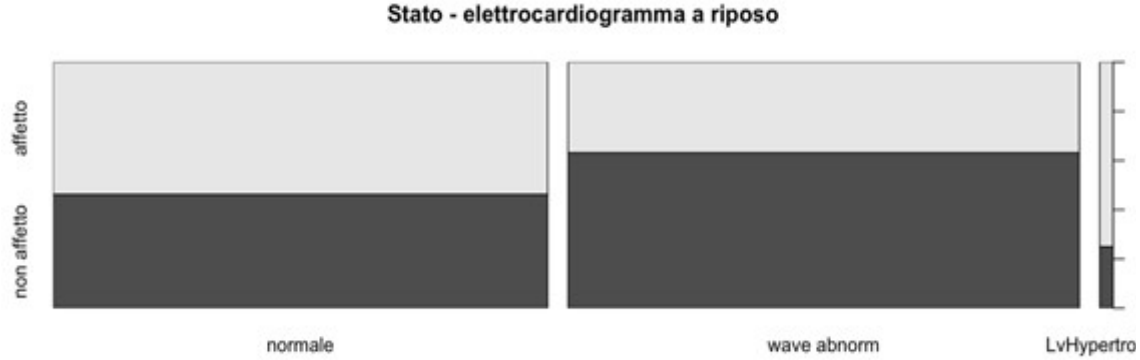

Stato - dolore al petto
affetto
non affetto
typical
atypical
non-anginal
asymptomatic

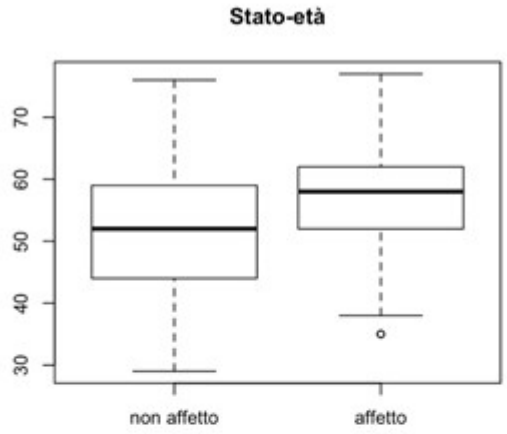

Stato - thal
affetto
non affetto
difFisso
difNormale
difReversibile

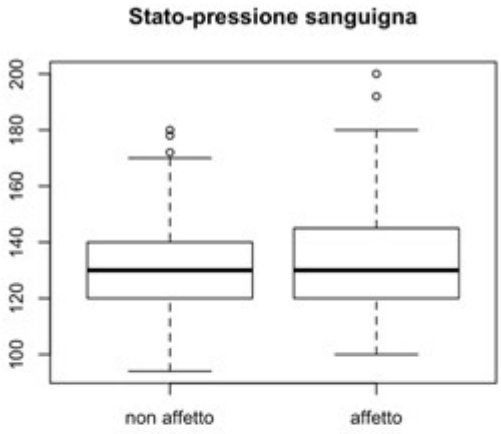

Stato - pendenza ST
target
affetto
non affetto
ascendente
piatto
discendente
slope

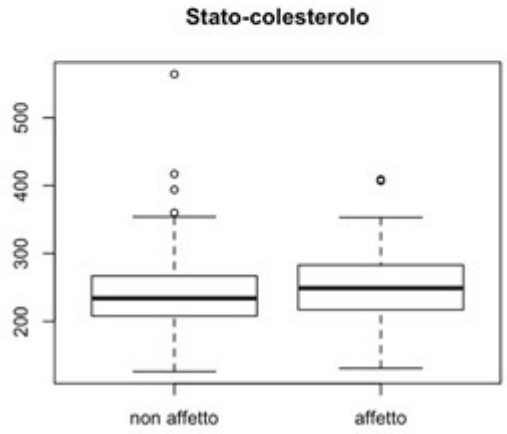

Stato - vasi sanguigni
target
affetto
non affetto
0
1
2
3
4
ca

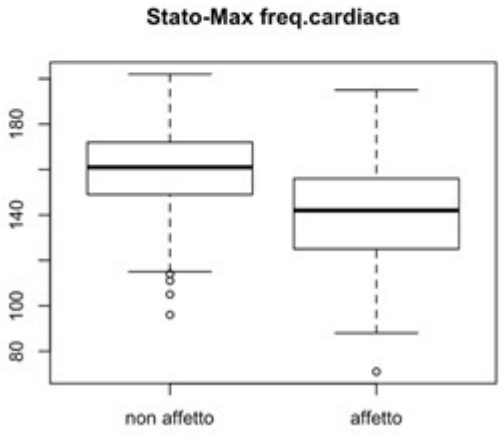

Stato - angina da sforzo
affetto
non affetto
no
yes
exang

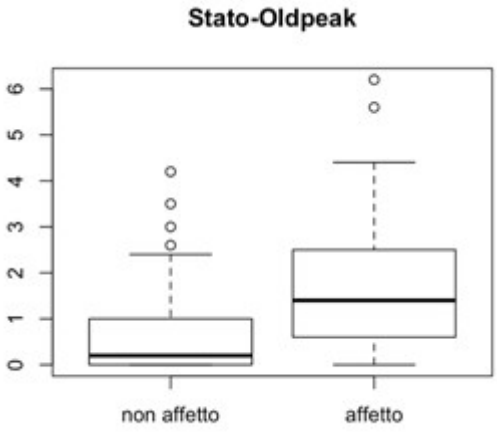

Stato - elettrocardiogramma a riposo
affetto
non affetto
normale
wave abnorm
LvHypertro


Stato-età
30
40
50
60
70
non affetto
affetto


Stato-pressione sanguigna
100
120
140
160
180
200
non affetto
affetto


Stato-colesterolo
200
300
400
500
non affetto
affetto


Stato-Max freq.cardiaca
80
100
140
180
non affetto
affetto


Stato-Oldpeak
0
1
2
3
4
5
6
non affetto
affetto

Dalle prime analisi descrittive, che analizzano l'effetto marginale delle singole variabili sulla variabile di interesse **target**, sembra che la probabilità di soffrire di malattie cardiache sia più alta per individui **maschi**, con dolore al petto (**cp**) di tipo **angina tipica**, elettrocardiogramma a riposo con valore **wave abnorm,** che provano dolore al petto durante l'esercizio fisico (**exang yes**). Inoltre, sono maggiormente presenti malattie cardiache se la **pendenza dell' ST** è piatta o crescente, se il numero di vasi sanguigni colorati da fluoroscopia (**ca**) cresce, se il risultato del thallium stress test (**thal**) è normale o reversibile, se la massima frequenza cardiaca raggiunta (**thalach**) è più bassa, se l'**età** è più elevata e il valore di **oldpeak** è alto.

Non sembrano avere effetto significativo la pressione sanguigna (**trestbps**) e il colesterolo nel sangue (**chol**).

2.1 Analisi congiunte

Si analizzano ora possibili relazioni/effetti congiunti tra le variabili.

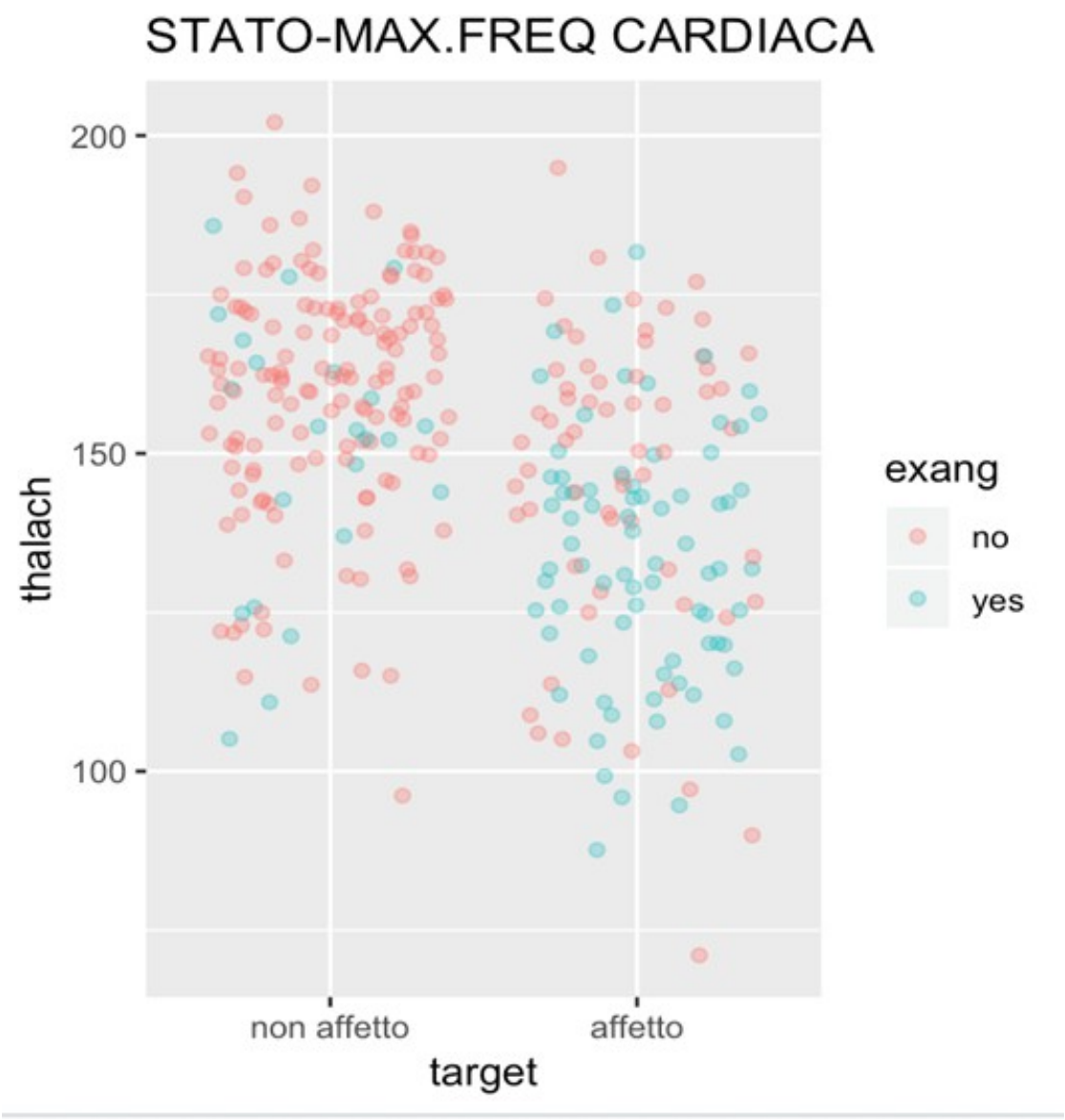


Fig.1 Si nota che individui che raggiungono frequenze cardiache più elevate tendono a non avere angina da sforzo durante l'esercizio. Entrambe sono caratteristiche che individuano una non predisposizione alle malattie cardiache. Il fatto che siano legate può far sì che l'inserimento di una nel modello potrebbe portare alla non significatività dell'altra.

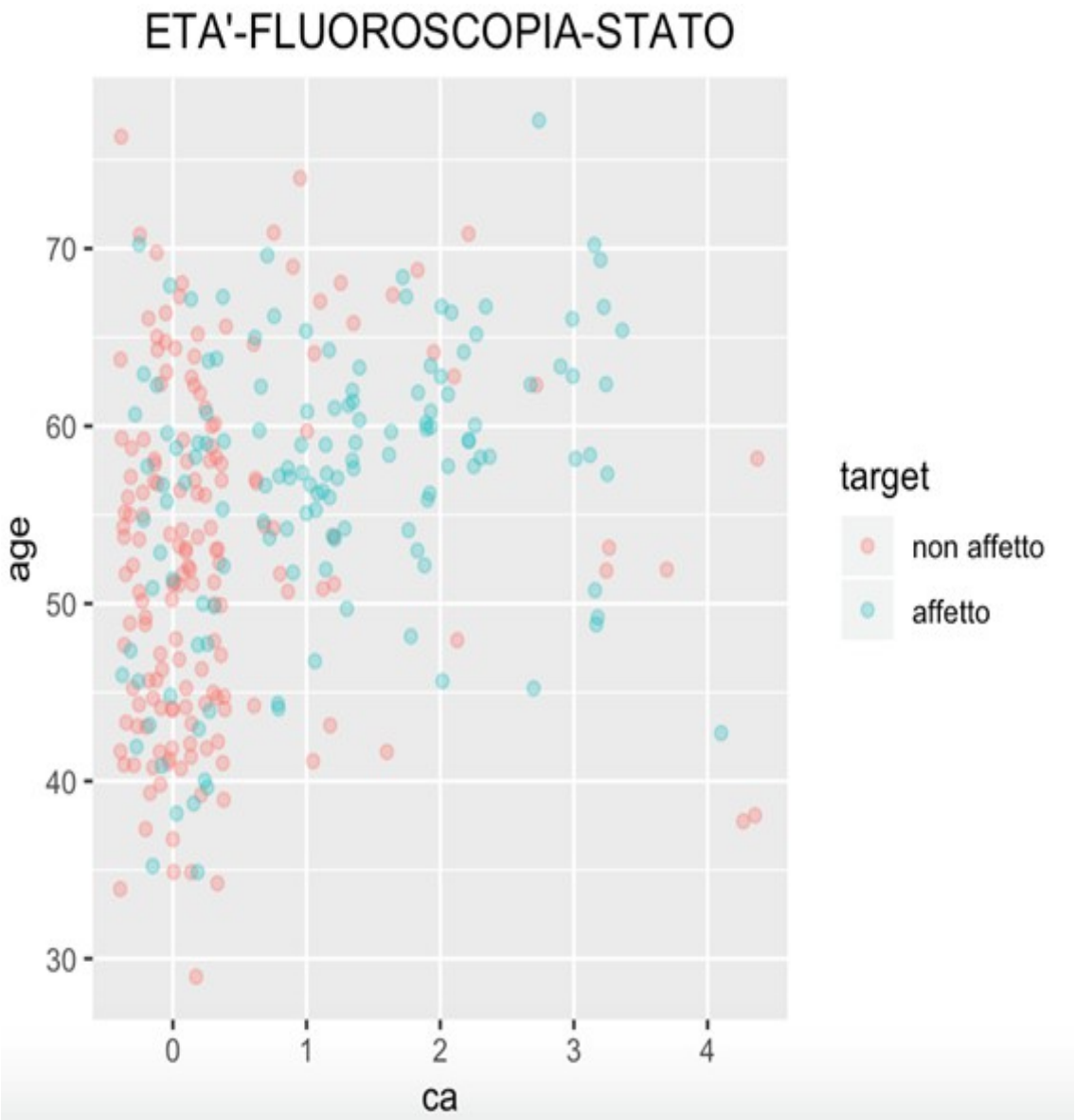


Fig.2 Il numero di vasi sanguigni colorati per fluoroscopia (**Ca**) tende a crescere con l'**età**, e ad essere maggiore per individui affetti rispetto ai non affetti (come già notato). Si vede anche che solo 5 unità presentano livello 4, si cui 4 non affetti e con età simili di quelle del gruppo 0. Per questa ragione il livello 4 verrà agglomerato al livello 0.

Fig.3 Le femmine sono affette da malattie cardiache a partire da **età** più elevate rispetto ai maschi. Questo potrebbe suggerire un'interazione tra l'essere femmina e l'avere età inferiore ai 50 anni.

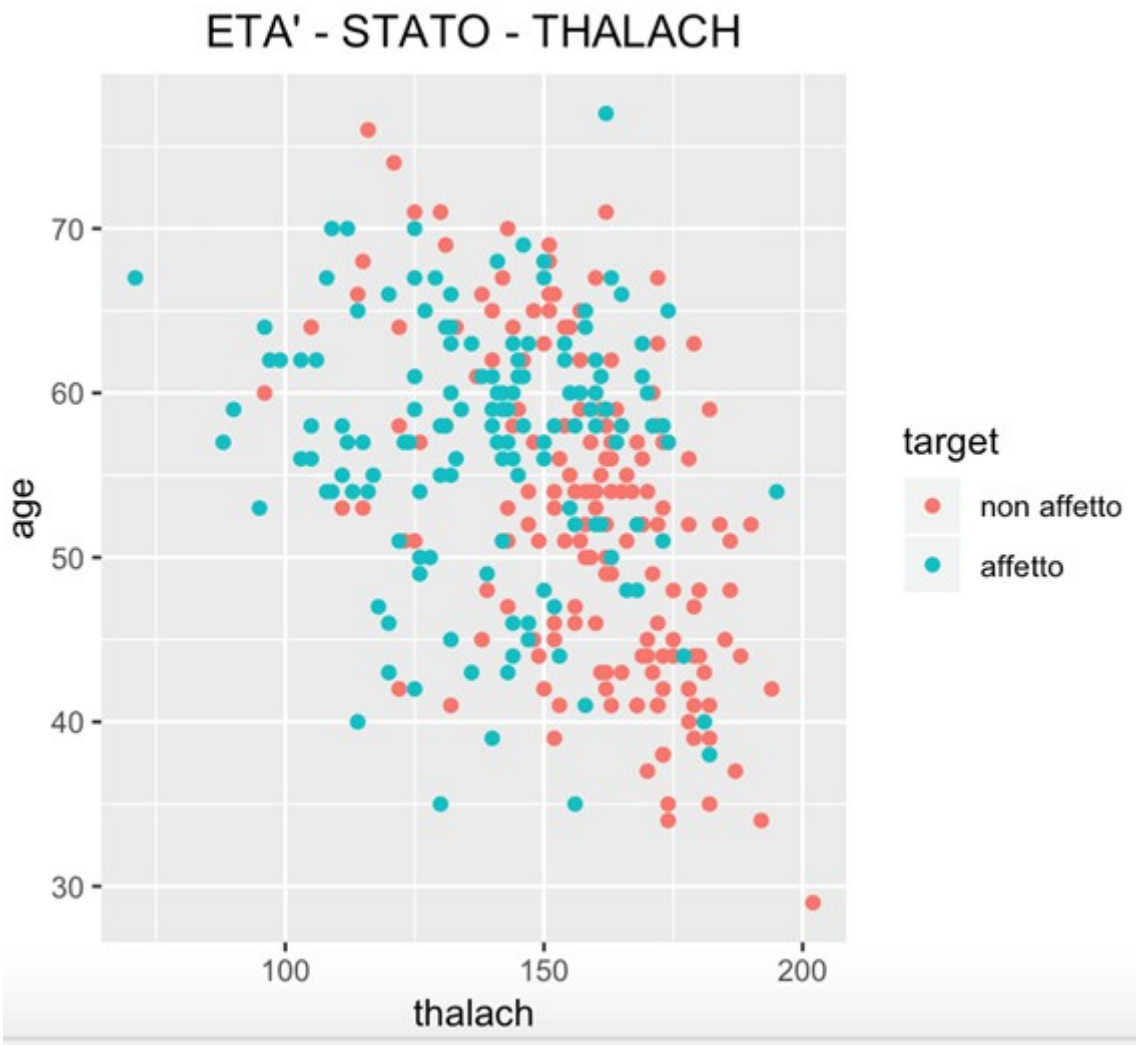


Fig.4 Si nota che individui con valori più elevati di massima frequenza cardiaca raggiunta (**thalach**) hanno **età** mediamente inferiori. Inoltre, come già notato con il relativo boxplot, i non affetti raggiungono livelli di frequenza cardiaca più elevata. Anche in questo caso l'effetto dell'età potrebbe essere parzialmente assorbito, vista la correlazione pari a -0.3985219.

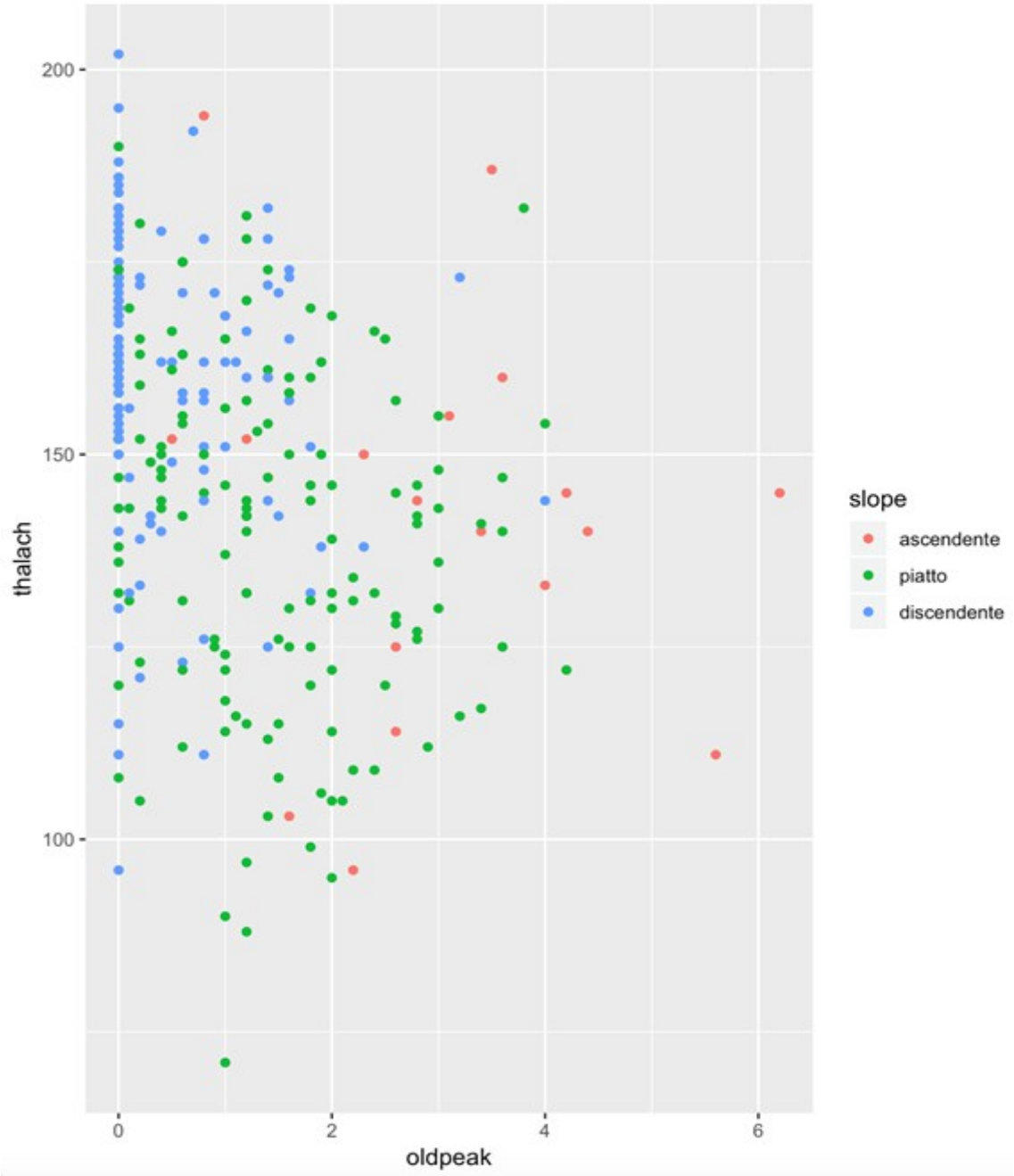


Fig.5 In figura si osserva come **Oldpeak**, valore connesso ad un risultato dell'elettrocardiogramma in fase di recupero, assume mediamente valori più bassi in corrispondenza di valori di **Thalach** più elevati. Inoltre sembra esserci prevalenza della modalità "discendente" di **Slope** per valori più bassi di **Oldpeak** ed elevati di **Thalach**.

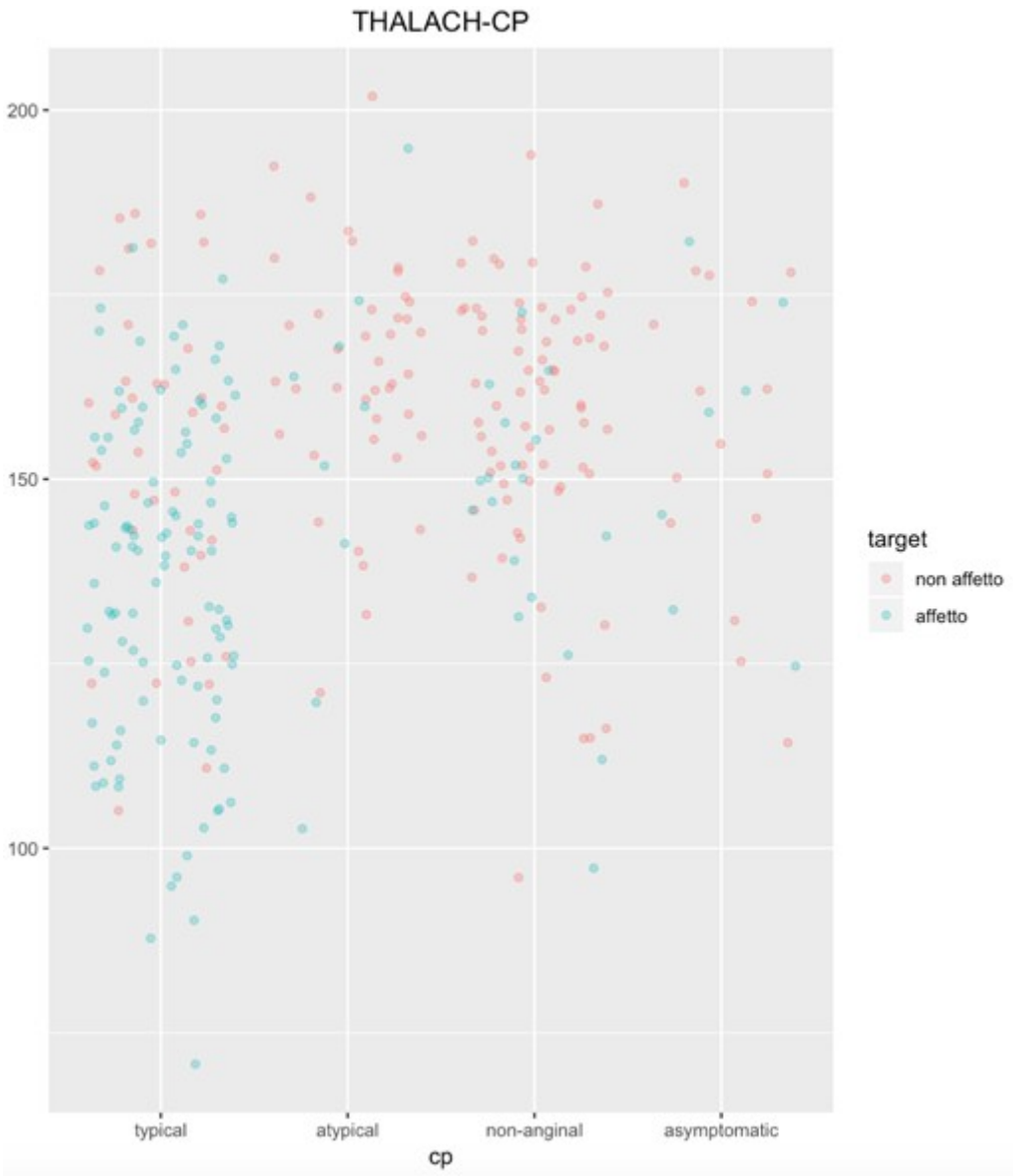


Fig.6. La modalità "typical" di tipo di dolore al petto (**cp**) si registra per individui che hanno un valore medio molto più basso di massima frequenza cardiaca raggiunta (**thalach**), mentre non sembra esserci un legame tra le altre tre modalità e i livelli di frequenza cardiaca.

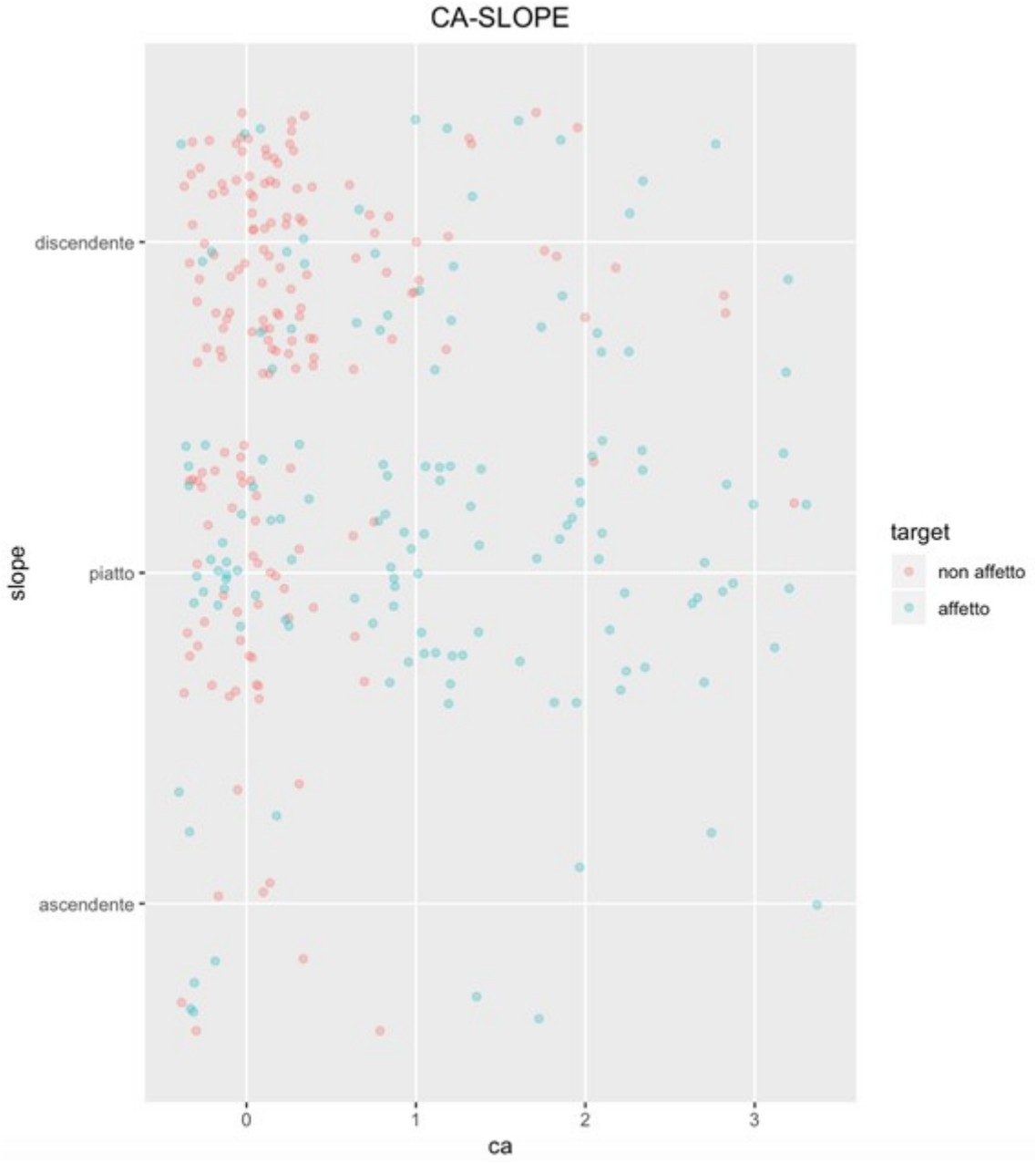


Fig.7 Sembra esserci un effetto congiunto del risultato dell'elettrocardiogramma **slope** e del numero di arterie colorate dalla fluoroscopia **(ca)**: la modalità "discendente" di slope appare un indice di basso rischio di malattie cardiache se presentata congiuntamente alla modalità 0 di ca, mentre la modalità "piatto" se abbinata a valori di ca superiori ad 1 rappresenta un fattore di rischio.

3 Stima del modello

La variabile di interesse per l'analisi è la variabile dicotomica Target. Si specifica un modello distributivo per le y_i :

$$Y_i \sim Be(p_i) \quad i \in \{1, \dots, n\} \quad n=303$$

Si definisce un predittore lineare $\eta_i = x_i^T \beta$ che si lega al parametro di interesse p_i nel seguente modo:

$$\eta_i = \log\left(\frac{p_i}{1-p_i}\right) \quad .$$

3.1 Stima del modello con tutte le variabili

Iniziamo con un modello di regressione logistica con tutte le variabili esplicative:

```
Call:
glm(formula = target ~ ., family = binomial, data = dati)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1026  -0.4512  -0.1111   0.2795   2.9335

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -2.574860   2.924093  -0.881 0.378552
age                                -0.026486   0.025415  -1.042 0.297338
sexmale                             1.840116   0.564766   3.258 0.001121 **
cpatypical                         -0.949492   0.573059  -1.657 0.097543 .
cpnon-anginal                      -2.015389   0.526577  -3.827 0.000130 ***
cpasymptomatic                     -2.434082   0.714461  -3.407 0.000657 ***
trestbps                            0.025906   0.011804   2.195 0.028182 *
chol                                0.004067   0.004230   0.962 0.336276
fbsfasting blood sugar > 120 mg/dl -0.372699   0.571061  -0.653 0.513986
restecgwave abnorm                 -0.447875   0.397271  -1.127 0.259583
thalach                            -0.019575   0.011832  -1.654 0.098037 .
exangyes                            0.809397   0.448574   1.804 0.071172 .
oldpeak                             0.418650   0.239346   1.749 0.080267 .
slopepiatto                         0.775628   0.876875   0.885 0.376407
slopediscendente                   -0.656826   0.946934  -0.694 0.487912
ca1                                 2.319460   0.522409   4.440 9.00e-06 ***
ca2                                 3.424436   0.806591   4.246 2.18e-05 ***
ca3                                 2.224318   0.922841   2.410 0.015940 *
thaldifNormale                     -0.316441   0.810112  -0.391 0.696083
thaldifReversibile                  1.364369   0.435188   3.135 0.001718 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 181.25  on 283  degrees of freedom
AIC: 221.25

Number of Fisher Scoring iterations: 6
```

```
> anova(fit)
Analysis of Deviance Table

Model: binomial, link: logit

Response: target

Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			302	417.64
age	1	15.777	301	401.86
sex	1	31.287	300	370.57
cp	3	77.654	297	292.92
trestbps	1	4.925	296	287.99
chol	1	2.376	295	285.62
fbs	1	0.024	294	285.59
restecg	1	2.241	293	283.35
thalach	1	20.122	292	263.23
exang	1	5.675	291	257.56
oldpeak	1	17.365	290	240.19
slope	2	5.447	288	234.74
ca	3	41.251	285	193.49
thal	2	12.237	283	181.25

3.2 Stima di un modello senza chol, fbs, e restecg

Tra le variabili non significative provo dapprima a togliere quelle che, seppur aggiunte sequenzialmente come quinta, sesta e settima variabile rispettivamente, non portano ad una grande riduzione della devianza, ovvero **chol, fbs,** e **restecg.**

```
Call:
glm(formula = target ~ age + sex + cp + trestbps + thalach +
    exang + oldpeak + slope + ca + thal, family = binomial, data = dati)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1141  -0.4540  -0.1205   0.2929   2.9226

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.34051    2.76728  -0.846 0.397676
age                -0.02051    0.02474  -0.829 0.407067
sexmale             1.69866    0.54024   3.144 0.001665 **
cpatypical         -0.97811    0.57003  -1.716 0.086184 .
cpnon-anginal      -2.11307    0.51858  -4.075 4.61e-05 ***
cpasymptomatic     -2.44238    0.70139  -3.482 0.000497 ***
trestbps            0.02588    0.01141   2.268 0.023349 *
thalach            -0.01766    0.01160  -1.523 0.127746
exangyes            0.77612    0.44215   1.755 0.079205 .
oldpeak             0.44780    0.23295   1.922 0.054571 .
slopepiatto         0.80434    0.85497   0.941 0.346819
slopediscendente   -0.64829    0.92177  -0.703 0.481863
ca1                 2.33682    0.51251   4.560 5.13e-06 ***
ca2                 3.29299    0.77964   4.224 2.40e-05 ***
ca3                 2.20969    0.91215   2.423 0.015414 *
thaldifNormale     -0.46996    0.79280  -0.593 0.553320
thaldifReversibile  1.33040    0.42712   3.115 0.001840 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 184.34  on 286  degrees of freedom
AIC: 218.34

Number of Fisher Scoring iterations: 6
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			302	417.64
sex	1	24.841	301	392.80
cp	3	84.189	298	308.61
trestbps	1	9.812	297	298.80
exang	1	13.151	296	285.64
slope	2	26.883	294	258.76
age	1	5.304	293	253.46
thalach	1	3.821	292	249.64
oldpeak	1	11.270	291	238.37
ca	3	41.268	288	197.10
thal	2	12.755	286	184.34

	Df	Deviance	Resid. Df	Resid. Dev
NULL			302	417.64
sex	1	24.841	301	392.80
thalach	1	59.107	300	333.69
cp	3	51.660	297	282.03
trestbps	1	10.342	296	271.69
exang	1	5.877	295	265.81
slope	2	14.086	293	251.72
age	1	2.088	292	249.64
oldpeak	1	11.270	291	238.37
ca	3	41.268	288	197.10
thal	2	12.755	286	184.34

Il secondo modello stimato contiene tre variabili (**Age, Thalach** e **Slope**) che non sono significativamente diverse da zero. Inoltre, il comando *Anova* permette di vedere la riduzione di devianza che si ottiene, aggiungendo al modello nullo le varie variabili

nell'ordine in cui appaiono, una alla volta. Modificando l'ordine di aggiunta delle variabili cambia, ovviamente, la riduzione in varianza. Variare l'ordine permette di captare eventuali relazioni tra le esplicative.

Per esempio, l'effetto di **Thalach** viene probabilmente assorbito da **Exang** e **Cp**, visti anche i rispettivi grafici del confronto tra coppie di esplicative, in quanto se **Thalach** viene inserita come seconda è molto significativa, ma lo restano anche **Exang** e **Cp**, mentre non vale viceversa.

Ca potrebbe essere quello che in parte assorbe l'effetto di **Slope**, viste le relazioni grafiche e il fatto che **Ca** è molto significativo nel modello ad ogni livello.

L'effetto dell'**età** appare non significativo perché vi sono molte altre misurazioni più specifiche che sono connesse con l'età.

3.3 Modello senza Thalach e age

Ristimo un terzo modello cui tolgo le variabili **Thalach** e **Age**

```
Call:
glm(formula = target ~ sex + cp + trestbps + exang + oldpeak +
    slope + ca + thal, family = binomial, data = dati)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9700  -0.4613  -0.1101   0.3371   3.0163

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.64754    1.77933  -3.174 0.001504 **
sexmale             1.61776    0.52007   3.111 0.001867 **
cpatypical         -1.09548    0.56304  -1.946 0.051698 .
cpnon-anginal      -2.21593    0.51333  -4.317 1.58e-05 ***
cpasymptomatic     -2.54793    0.69785  -3.651 0.000261 ***
trestbps            0.02209    0.01091   2.025 0.042868 *
exangyes            0.86995    0.43378   2.006 0.044906 *
oldpeak             0.49233    0.23068   2.134 0.032825 *
slopepiatto         0.89075    0.82795   1.076 0.281989
slopediscendente   -0.67810    0.89852  -0.755 0.450437
ca1                 2.35836    0.49627   4.752 2.01e-06 ***
ca2                 3.09153    0.75123   4.115 3.87e-05 ***
ca3                 2.27140    0.89492   2.538 0.011146 *
thaldifNormale     -0.27823    0.77190  -0.360 0.718509
thaldifReversibile  1.38324    0.42300   3.270 0.001075 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 186.85  on 288  degrees of freedom
AIC: 216.85

Number of Fisher Scoring iterations: 6
```

```
> anova(fit1, fit2, test="Chisq")
Analysis of Deviance Table

Model 1: target ~ age + sex + cp + trestbps + thalach + exang + oldpeak +
    slope + ca + thal
Model 2: target ~ sex + cp + trestbps + exang + oldpeak + slope + ca +
    thal
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       286     184.34
2       288     186.85 -2  -2.5102    0.285
```

Il test *Anova,* che equivale al test del log rapporto di verosimiglianza, fa preferire quest'ultimo modello (**fit2**).

3.4 Modello senza Slope

```
> anova(fit2, fit3, test="Chisq")
Analysis of Deviance Table

Model 1: target ~ sex + cp + trestbps + exang + oldpeak + slope + ca +
    thal
Model 2: target ~ sex + cp + trestbps + exang + oldpeak + ca + thal
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       288     186.85
2       290     199.64 -2  -12.789 0.001671 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il test *Anova* mi porta a concludere che si preferisce il modello in cui questa esplicativa non viene omessa.

Un ulteriore parametro di confronto per la scelta del modello è il criterio di informazione di *Akaike* (**AIC**), in base al quale si preferisce il modello con l'AIC minore. Il valore AIC di **fit2** è 216.8535, quello del modello senza **Slope** è 225.6424.

Il modello scelto, **fit2**, è lo stesso che viene selezionato eseguendo il comando STEP a partire da tutte le variabili e con *direction="both".*

Per quanto riguarda il modello scelto, l'intepretazione dei coefficienti è la seguente. Si consideri il coefficiente associato al livello "Yes" della variabile esplicativa **Exang** (angina da sforzo), 0.86995: questo significa che, tenendo fisse tutte le altre variabili

esplicative, il passaggio dal livello base (**Exang = "no"**) alla modalità "yes" comporta un aumento dell'odds ratio pari a $e^{0.86995} \cong 3.38$

4 Previsione

Si prevede ora, sulla base del modello scelto, la probabilità di insorgenza di malattie cardiache per un nuovo individuo, cui si assegnano dei valori ipotetici alle variabili esplicative. La scelta della valori è stata fatta privilegiando modalità/valori delle esplicative che hanno effetto positivo sulla probabilità di riscontrare malattie cardiache.

```
>new=data.frame(sex="female",cp="atypical", trestbps= 130, thalach=163,
exang="no",     oldpeak=4, slope="piatto", ca="2", thal="difFisso")
>predict(fit2,newdata=new,type="response")
```

0.8889378

```
>new=data.frame(sex="male",cp="atypical", trestbps= 130, thalach=163, exang="no",
oldpeak=4, slope="piatto", ca="2", thal="difFisso")
>predict(fit2,newdata=new,type="response")
```

0.9758186

Le stessa probabilità, usando il legame "probit", sono pari a 0.8261215 e 0.9668829 rispettivamente.