Tesina di introduzione
alla statistica computazionale

Titanic Dataset

A cura di:
Edoardo Marabini
Marco Belardinelli

Obiettivo della presentazione

- Trovare un’eventuale correlazione tra i sopravvissuti e le altre variabili presenti all’interno del nostro dataset
- Le altre variabili sono correlate tra loro?
- Possiamo predire i superstiti con i mezzi a nostra disposizione?

Pulizia dataset

```
  PassengerId Survived Pclass                                                Name    Sex Age SibSp Parch           Ticket    Fare Cabin Embarked
1           1        0      3                             Braund, Mr. Owen Harris   male  22     1     0        A/5 21171  7.2500              S
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0         PC 17599 71.2833   C85        C
3           3        1      3                              Heikkinen, Miss. Laina female  26     0     0 STON/O2. 3101282  7.9250              S
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0           113803 53.1000  C123        S
5           5        0      3                            Allen, Mr. William Henry   male  35     0     0           373450  8.0500              S
6           6        0      3                                    Moran, Mr. James   male  NA     0     0           330877  8.4583              Q
```

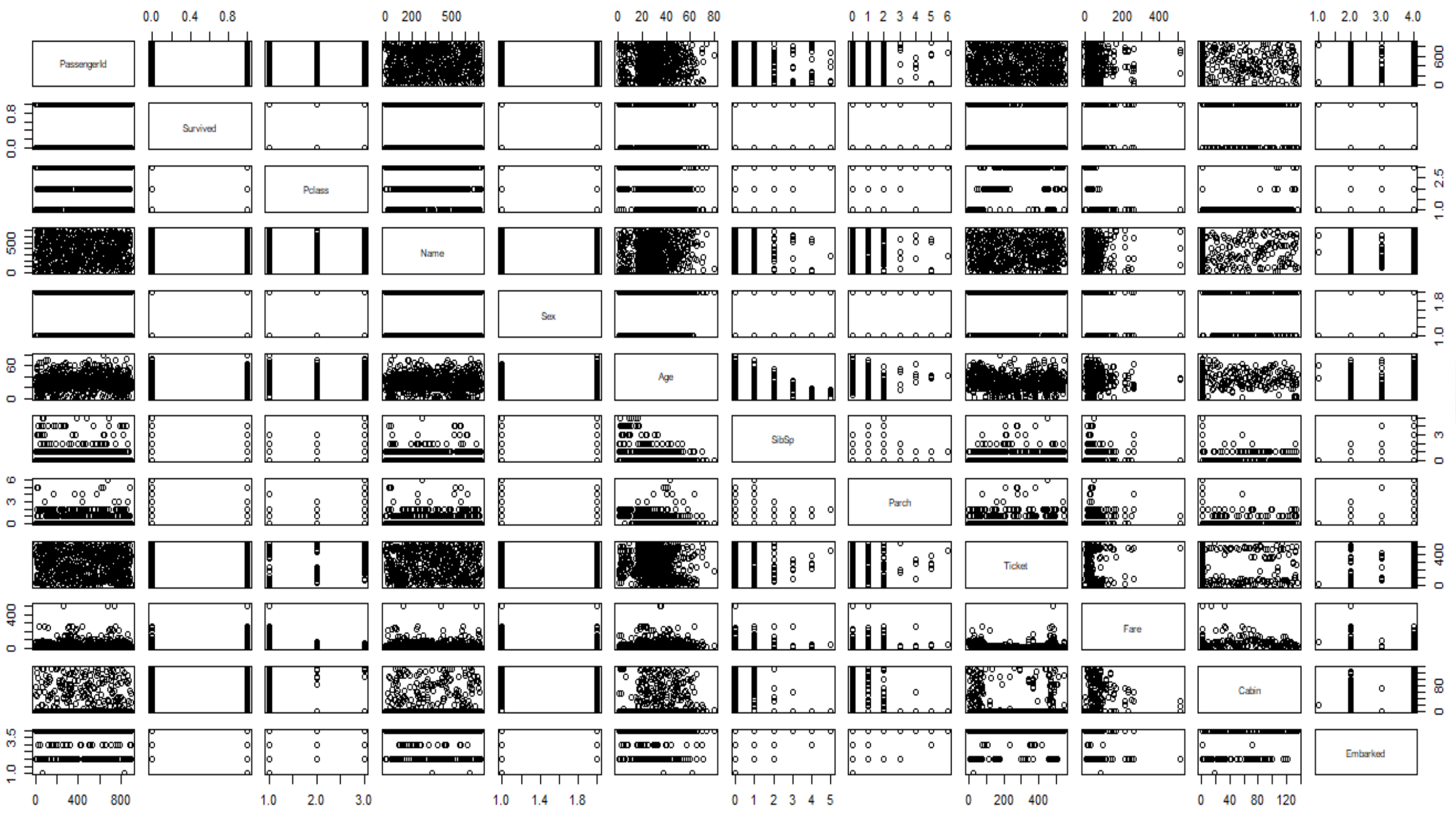
Nel dataset troviamo 12 variabili e 861 osservazioni.

```
> duplicated(data1)
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [26] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [51] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [76] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[101] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[126] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[151] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

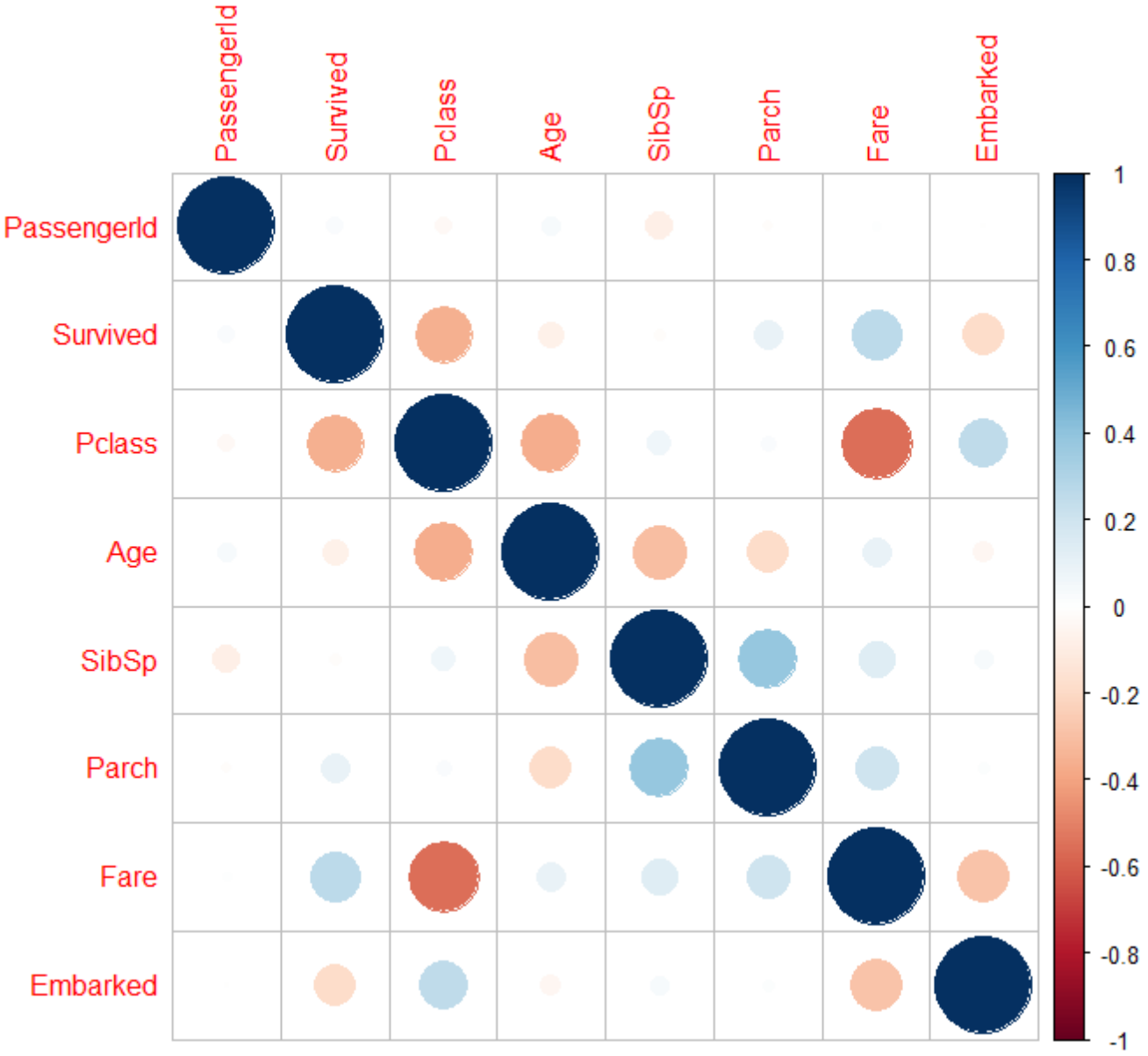
Non sono presenti osservazioni duplicate.

```
> table(data1$Survived)

  0   1
549 342
> table(data1$Pclass)

  1   2   3
216 184 491
> data=drop_na(data1)
> table(data$Survived)

  0   1
424 290
> table(data$Pclass)

  1   2   3
186 173 355
```

Alcune delle osservazioni purtroppo presentano molti valori mancanti costringendoci ad eliminarle (notiamo già che la maggior parte dei dati persi appartenevano a passeggeri della terza classe).


PassengerId
Survived
Pclass
Name
Sex
Age
SibSp
Parch
Ticket
Fare
Cabin
Embarked

Correlazioni

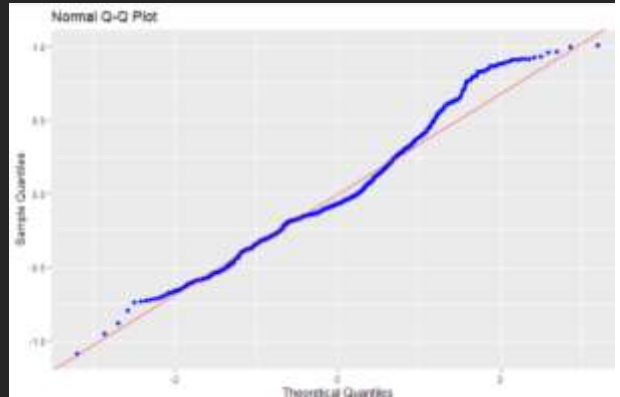


Con questa tabella e il plot precedente cerchiamo eventuali correlazioni tra le variabili, notiamo tuttavia che non sono presenti correlazioni così elevate da necessitare della rimozione di una o più variabili.

Alcune variabili (come l'identificativo del biglietto) sono tuttavia state omesse da questa tabella poiché presentavano valori alfanumerici di difficile interpretazione.

Regressione lineare

Procediamo con la costruiamo del modello di regressione lineare. Mettiamo in relazione le nostre variabili indipendenti con «Survived» ovvero la variabile dicotomica che identifica morti e sopravvissuti. Questa, d'ora in poi, sarà la nostra variabile dipendente (Y).
Il VIF ci conferma che la correlazioni tra le variabili è trascurabile.




```
> summary(model11)

Call:
lm(formula = Survived ~ . - Name - Ticket - Cabin, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08883 -0.23205 -0.06948  0.22923  1.00816

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.440e+00  1.026e-01  14.038  < 2e-16 ***
PassengerId  5.357e-05  5.542e-05   0.967   0.3340
Pclass      -1.897e-01  2.263e-02  -8.385 2.75e-16 ***
Sexmale     -4.855e-01  3.141e-02 -15.453  < 2e-16 ***
Age         -6.449e-03  1.127e-03  -5.720 1.58e-08 ***
SibSp       -5.048e-02  1.743e-02  -2.896   0.0039 **
Parch       -1.075e-02  1.903e-02  -0.565   0.5723
Fare         2.017e-04  3.464e-04   0.582   0.5604
Embarked    -3.069e-02  1.916e-02  -1.602   0.1096
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3817 on 705 degrees of freedom
Multiple R-squared:  0.4035,    Adjusted R-squared:  0.3968
F-statistic: 59.62 on 8 and 705 DF,  p-value: < 2.2e-16
```

```
> vif(model11)
PassengerId      Pclass     Sexmale         Age       SibSp       Parch        Fare    Embarked
   1.009016    1.760276    1.121660    1.312681    1.285432    1.289833    1.644126    1.120922
```

La regressione appena usata può esser letta come:
Y= β0+ β1*X1+ β2*X2+…+ βk*Xk.

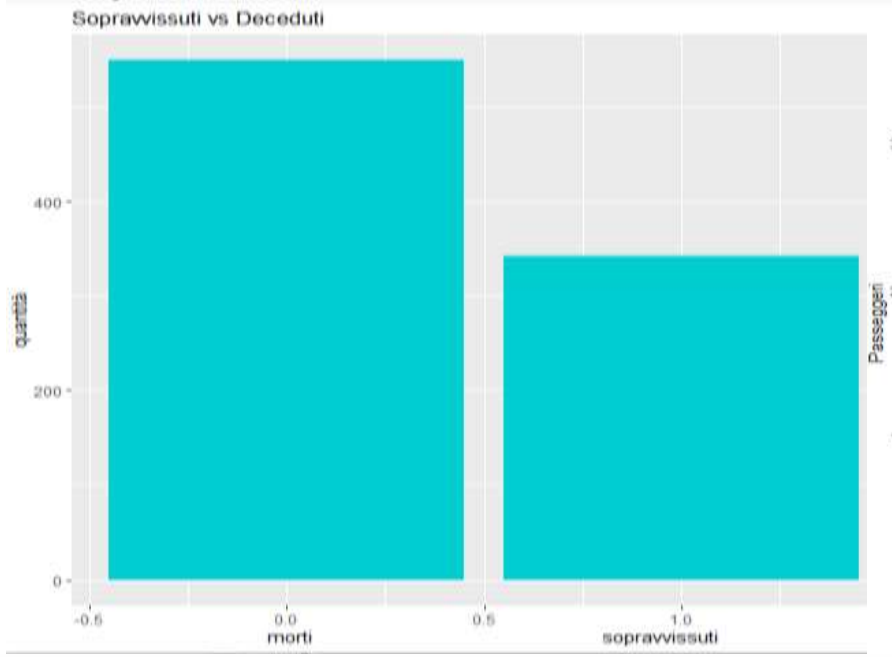
- Y= Survived (sopravvissuti / deceduti)
- Intercetta: 1,440
- β1= -1,897e-01
- β2= -4,855e-01
- Etc

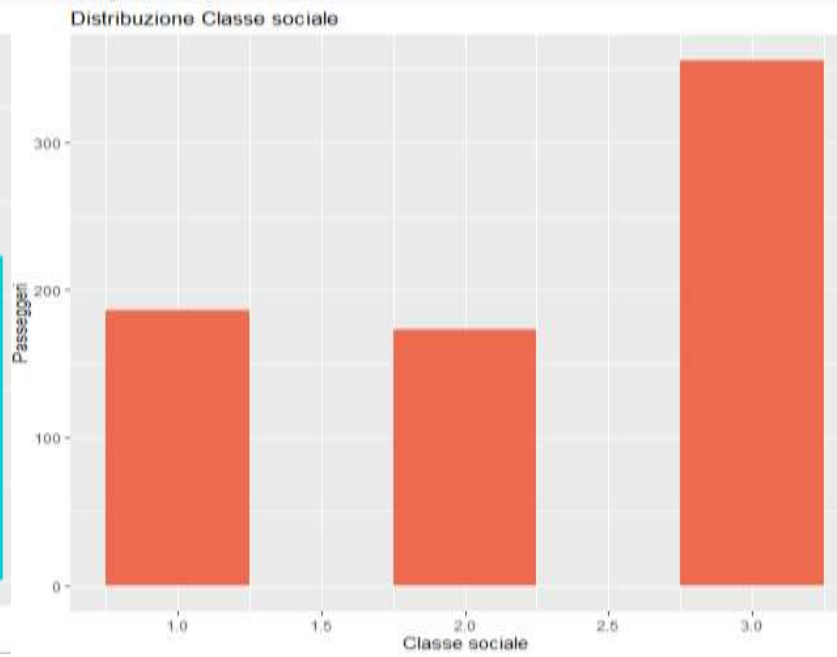
Y= 1,440 – (1,897e-01)*X1 – (4,855e-01)*X2+ β3…

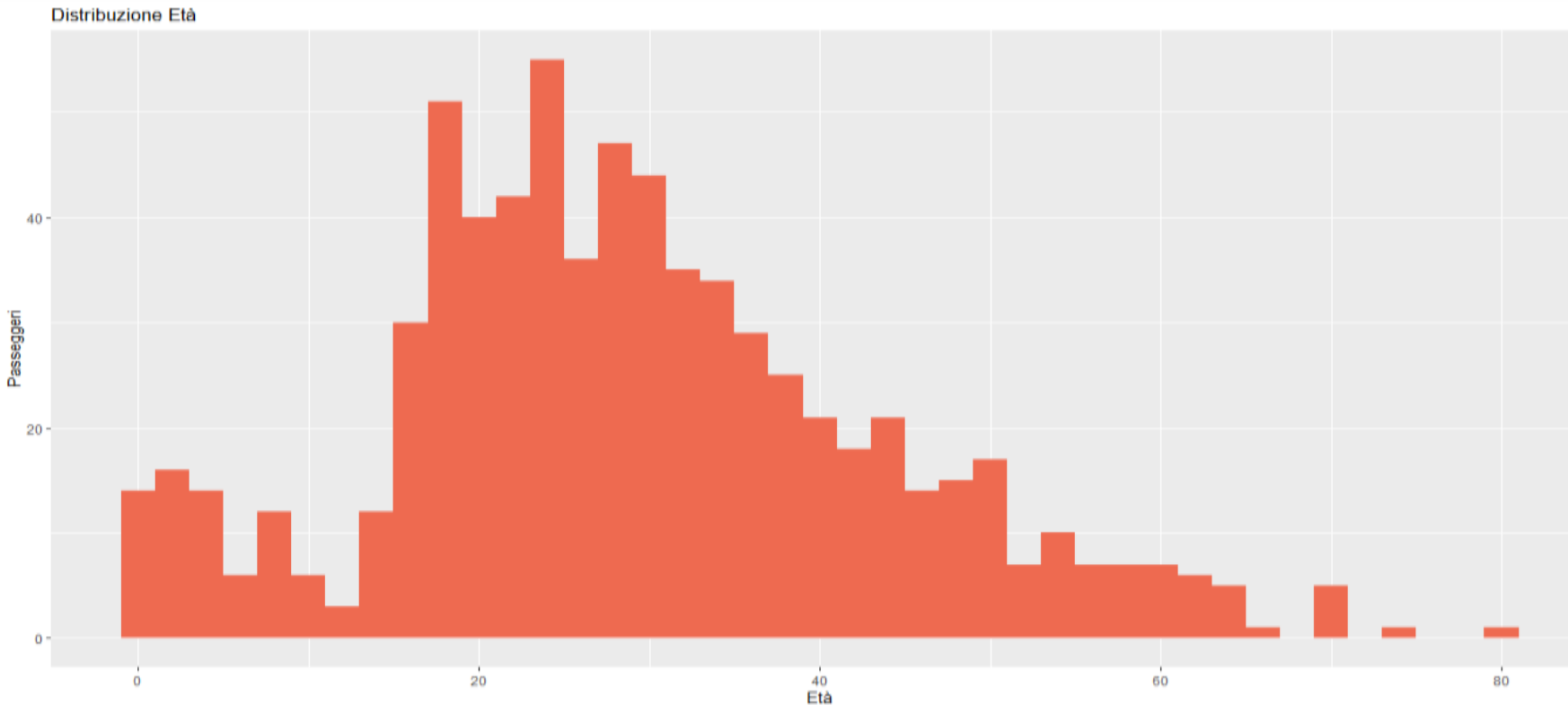
L'*R-quadro* risultante dalla regressione è 0,4035, questo valore ci fa sapere che l'affidabilità del nostro modello è del 40% circa; l'*R-quadro ADJ* invece è di poco inferiore (0,3968). (Aggiungendo variabili alla regressione l'R-quadro continuerà ad aumentare mentre l'R-quadro ADJ diminuirà nel caso in cui la variabile inserita non dovesse essere significativa).

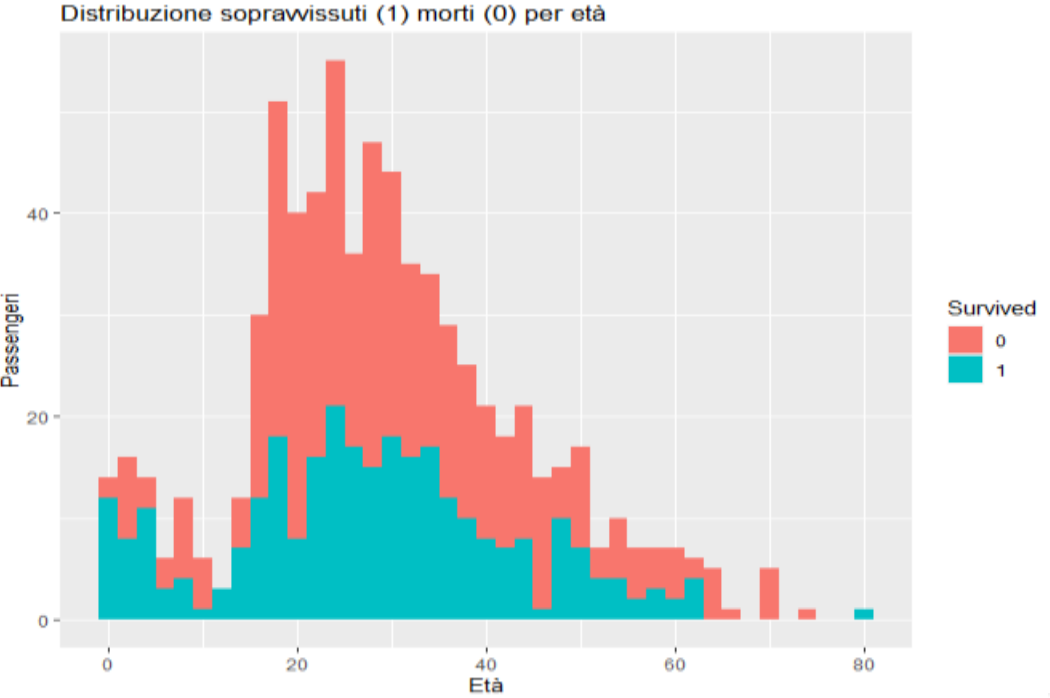
Osservazione grafica dei dati

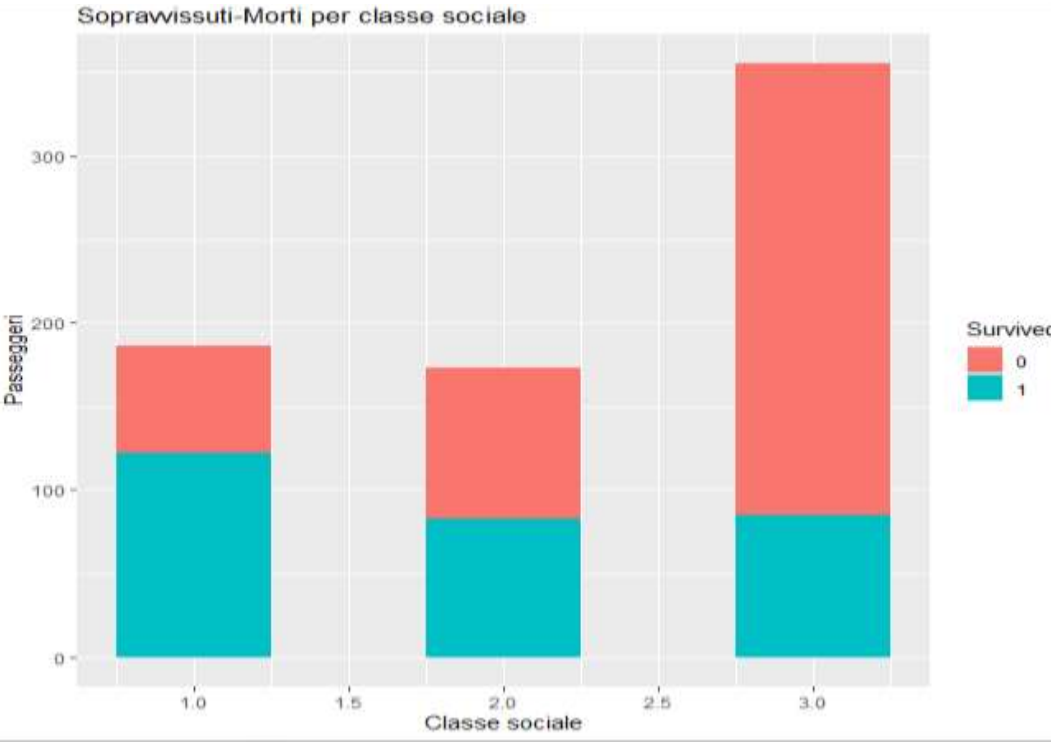
- Osserviamo ora le variabili che la regressione ha rilevato come significative.
- La rappresentazione grafica aiuta a capire meglio il dataset e la situazione in cui ci troviamo.

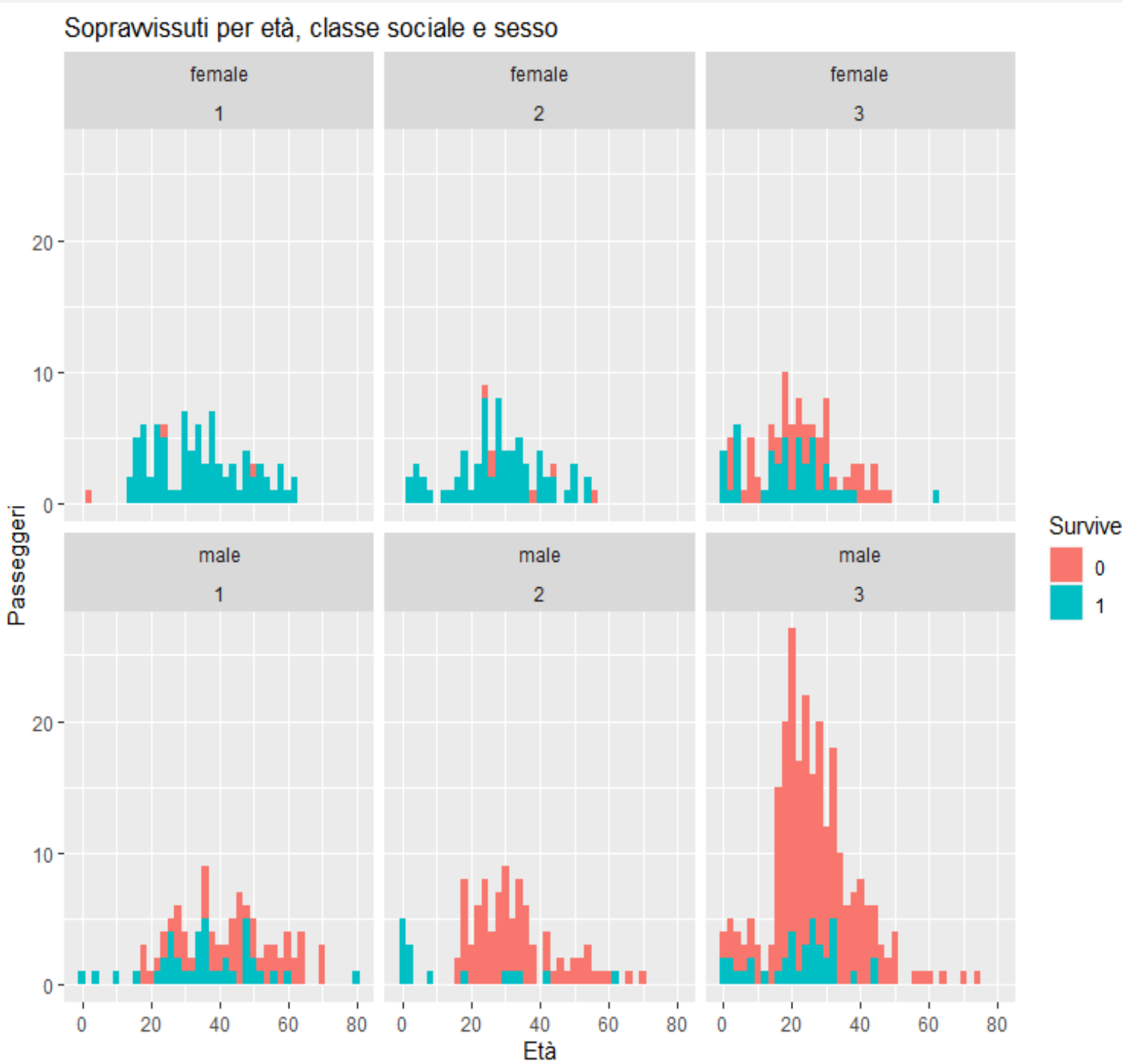







Distribuzione sopravvissuti (1) morti (0) per età
Passengeri
Età
Survived
0
1


Sopravvissuti-Morti per classe sociale
Passeggeri
Classe sociale
Survived
0
1


Sopravvissuti per età, classe sociale e sesso
female
1
female
2
female
3
male
1
male
2
male
3
Passeggeri
Età
Survive
0
1

Previsioni

- Ci avviamo ora verso l'ultimo punto della nostra analisi. Dopo aver creato un nuovo dataset composto da tutte le variabili a nostra disposizione, trasformate in numeriche proseguiamo usando una logit poichè la nostra Y è una variabile dicotomica.

```
> View(a)
> data_elabora=data.frame(a)
> out_LR <-glm(d3~.,  family=binomial(link="logit") , data=data_elabora)#
> yprob <-predict(out_LR,data_elabora,"response")
> y_prev<- ifelse(yprob>0.5,'1','0')
> y_oss <- data_elabora$d3 ;
> n <- nrow(data_elabora)
> T0 = sum((y_prev==0&y_oss==0))
> T1 = sum((y_prev==1&y_oss==1))
> F0 = sum((y_prev==0&y_oss==1))
> F1 = sum((y_prev==1&y_oss==0))
> accuracy=((T0+T1)/n)*100; specificita=(T1/(T1+F0))*100;sensitivita =(T0/(T0+F1))*100
> accuracy
[1] 80.2521
> specificita # raramente sbaglia classificazione morti / falsi positivi
[1] 71.72414
> sensitivita # raramente sbaglia classificazione vivi/ falsi negativi
[1] 86.08491
>
```

I risultati sembrano buoni, l'accuratezza è alta, così come la specificità e la sensitività!

Previsioni

- Ci avvaliamo della stepwise per scegliere la combinazione di variabili migliore per provare ad aumentare i nostri valori.
- Continuiamo creando un nuovo dataset con le variabili suggeriteci dalla stepwise.
- Procediamo come prima

42	87	3		d1	d2	d9	0.0444942026	0.0404568541	423.397204
51	88	3		d1	d5	d9	0.0369953842	0.0329263506	432.260723
63	89	3		d2	d5	d6	0.0181591188	0.0140104953	454.524979
39	90	3		d1	d2	d6	0.0134506000	0.0092820815	460.090396
48	91	3		d1	d5	d6	0.0126824871	0.0085107230	460.998297
38	92	3		d1	d2	d5	0.0086794653	0.0044907870	465.729824
130	93	4	d2	d4	d5	d8	0.3997184503	0.3963318125	5.526102
137	94	4	d2	d4	d8	d9	0.3932181203	0.3897948093	13.209419
133	95	4	d2	d4	d6	d8	0.3923240049	0.3888956495	14.266254
96	96	4	d1	d2	d4	d8	0.3913651867	0.3879314219	15.399566
135	97	4	d2	d4	d7	d8	0.3902013000	0.3867609688	16.775267
153	98	4	d4	d5	d8	d9	0.3744222800	0.3708929276	35.425893
117	99	4	d1	d4	d8	d9	0.3730998048	0.3695629913	36.989044
156	100	4	d4	d6	d8	d9	0.3725014889	0.3689612998	37.696247
157	101	4	d4	d7	d8	d9	0.3722741242	0.3687326524	37.964989
151	102	4	d4	d5	d7	d8	0.3720754856	0.3685328931	38.199778

d2= Età d4= Classe Passeggero d5= Fratelli/Sorelle/Partnet a bordo d8= Sesso

```
> #nomi=names(a)
> data_elaborab=data.frame(b)
> out_LRb <-glm(d3~.,  family=binomial(link="logit") , data=data_elaborab)#
> yprob <-predict(out_LRb,data_elaborab,"response")
> y_prev<- ifelse(yprob>0.5,'1','0')
> y_oss <- data_elaborab$d3 ;
> n <- nrow(data_elaborab)
> T0 = sum((y_prev==0&y_oss==0))
> T1 = sum((y_prev==1&y_oss==1))
> F0 = sum((y_prev==0&y_oss==1))
> F1 = sum((y_prev==1&y_oss==0))
> accuracy=((T0+T1)/n)*100; specificita=(T1/(T1+F0))*100; sensitivita =(T0/(T0+F1))*100
> accuracy
[1] 80.81232
> specificita
[1] 73.10345
> sensitivita
[1] 86.08491
> out_LRc <-lm(d3~., data=data_elaborab)
```

L'accuratezza è la sensibilità sono rimaste invariate ma la specificità è aumentata!

Previsioni

- Dividiamo il nostro dataset in 70-30 per procedere con il nostro lavoro ma…

```
> ind <- sample(2, nrow(data_elaborab), replace=TRUE, prob=c(0.70, 0.30))
> ftable(ind)
ind   1   2

    508 206
```

- Data quindi la natura «sbilanciata» del dataset procediamo dividendo il dataset in due subset, uno contenente tutti i morti e uno tutti i superstiti.

```
> morti=subset(data_elaborab, d3== 0)
> vivi=subset(data_elaborab, d3== 1)
```

- Creiamo ora due dataset, uno contenente il 70% del primo e del secondo dataset e uno che contiene i restanti 30%.

```
> morti7<-sample_frac(morti, 0.7)
> kek<-as.numeric(rownames(morti7))
> morti3<- morti[-kek,]
> vivi7<-sample_frac(vivi, 0.7)
> wew<-as.numeric(rownames(vivi7))
> vivi3<- vivi[-wew,]
> training=rbind(vivi7,morti7)
> test=rbind(morti3,vivi3)
```

Previsioni

Ci muoviamo ora nello stesso modo di prima per farci restituire i valori di accuratezza, specificità e sensitività.
Anche dopo aver creato i due subset, poiché i dati erano sbilanciati, ci ritroviamo con le strade intraprese prima che ci portano purtroppo allo stesso risultato, un'accuratezza molto bassa!

```
> out_LA <-glm(d3 ~.,  family=binomial(link="logit"), data=data.training)
> xpred=predict(out_LA,data.test,"response")
> x_prev<- ifelse(xpred>0.5,'1','0')
> x_oss <- data.test$d3
> tab <- xtabs(~x_prev+x_oss)
> tab
      x_oss
x_prev   0   1
     0 110  24
     1  17  63
> n <- nrow(data_elaborab)
> T0 = sum((x_prev==0&x_oss==0))
> T1 = sum((x_prev==1&x_oss==1))
> F0 = sum((x_prev==0&x_oss==1))
> F1 = sum((y_prev==1&y_oss==0))
> accuracy=((T0+T1)/n)*100
> specificita=(T1/(T1+F0))*100
> sensitivita =(T0/(T0+F1))*100
> accuracy;specificita; sensitivita
[1] 24.22969
[1] 72.41379
[1] 65.08876
```

Specificità e sensibilità sono un pò diminuite ma l'accuratezza è COLATA A PICCO!

Conclusioni

Possiamo quindi affermare che:

- Le variabili iniziali non presentavano correlazione tra loro.
- Il modello sembra avere buoni valori, tuttavia dobbiamo tenere a mente che il nostro *R-quadro ADJ* era abbastanza basso.
- Probabilmente con mezzi più adeguati e tecniche più avanzate saremmo riusciti ad ottenere un risultato migliore.

Grazie per l’attenzione