Pràctica 2: Neteja i validació de dades

Autor: Gerard Masllorens Fuentes

Desembre 2019

Contents

1.Descripció del dataset

En aquest exercici treballaré amb el dataset dels passatgers del Titanic. Aquest dataset està format per dos subset que en total sumen 1309 observacions i 12 variables. Les variables són:

- **Survival**: Variable dummy que indica si el passatger va sobreviure.
- **pclass**: Classe en la que viatjava el passatger: primera, segona o tercera.
- **sex**: Sexe del passatger.
- **Age**: Edat del passatger.
- **sibsp**: nombre de germans i/o parelles a bord del Titanic.
- **parch**: nombre de pares i/o fills a bord del Titanic.
- **ticket**: Número de tiquet.
- **fare**: Tarifa del tiquet.
- **cabin**: Número de cabina.
- **embarked**: Port d'embarcament C = Cherbourg, Q = Queenstown, S = Southampton.

En general aquest dataset és interessant per aprendre sobre un esdeveniment històric i per poder practicar tècniques d'anàlisis de dades amb dades reals.

2.Integració i selecció de les dades d'interès a analitzar.

Comencem carregant les dades. En aquest cas el recurs estava dividit en dos dataset: "test" i "train". Això és degut a que és un joc de dades per fer exercicis d'aprenentatge automàtic. En aquesta pràctica no necessitem un dataset per entrenar les dades i un altre per provar el resultats, per tant, ajuntem els dos datasets.

A continuació mirem l'estructura del dataset.

```r
# Carreguem els paquets R que utilitzarem
library(ggplot2)
library(dplyr)

# Guardem el joc de dades test i train en un únic dataset
test <- read.csv('test.csv',stringsAsFactors = FALSE)
train <- read.csv('train.csv', stringsAsFactors = FALSE)

# Unim els dos jocs de dades en un només
Titanic <- bind_rows(train,test)
filas=dim(train)[1]

# Verifiquem l'estructura del joc de dades
str(Titanic)
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

3. Neteja de les dades.

3.1. Dades que contenen zeros o elements buits

Treballem els atributs amb valors buits. Comencem mirant quines dades tenen valors buits. Per fer-ho mirarem quines dades tenen elements NA (not available) o bé estan buides.

```r
# Estadístiques de valors buits
colSums(is.na(Titanic))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
```

```r
colSums(Titanic=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0          NA           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0          NA        1014           2
```

Veiem que en aquest dataset els valors buits es poden representar de les dues maneres que hem descrit abans. Concretament hi ha valors buits amb el valor NA a Survived, Age, Fare. També hi ha valors buits representades deixen la cela buida a Cabin i Embarked. A continuació els treballem.

Comencem per l'edat. En aquest cas és complicat assignar valors mitjançant mètodes probabilístics com knn perquè tenim poques dades individualitzades. De fet, moltes són dades que depenen d'estrats (sexe, tarifa, classe, etc.). En aquest cas he optat per assignar la mitjana de l'edat als valors perduts d'edat.

```
# Prenem la mitjana per a valors buits de la variable "Age"
Titanic$Age[is.na(Titanic$Age)] <- mean(Titanic$Age,na.rm=T)
```

Per fare i embarketd veiem que només hi ha 3 observacions entre totes que tenen valors buits. En aquest cas opto per treure aquestes observacions

```
# eliminem les observacions buides d'embarked i fare
Titanic<-Titanic[!(Titanic$Embarked==""),]
Titanic<-Titanic[!(is.na(Titanic$Fare)),]
```

Pel que fa a la cabina veiem que la majoria d'observacions no tenen informació sobre la cabina on estaven els passatger. En aquest cas aquesta variable ens aporta molt poca informació i decideixo eliminar-la.

```
# eliminem la variable cabina
Titanic<-Titanic[, -(11)]
```

Finalment amb la variable Survived tenim un problema més important ja que, en principi, és la nostra variable d'interès. En aquest cas opto per deixar-la com estar i fer l'anàlisi deixant els valors buits.

```
# comprovem
colSums(is.na(Titanic))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         417           0           0           0           0
##       SibSp       Parch      Ticket        Fare    Embarked
##           0           0           0           0           0
```

```
colSums(Titanic=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0          NA           0           0           0           0
##       SibSp       Parch      Ticket        Fare    Embarked
##           0           0           0           0           0
```

3.2 Identificació i tractament de valors extrems.

Primer de tot discretitzem quan té sentit.

```
# Per a quines variables tindria sentit un procés de discretització?
apply(Titanic,2, function(x) length(unique(x)))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##        1306           3           3        1304           2          98
##       SibSp       Parch      Ticket        Fare    Embarked
##           7           8         927         280           3
```

```
# Discretitzem les variables amb poques classes
cols<-c("Survived","Pclass","Sex","Embarked")
for (i in cols){
  Titanic[,i] <- as.factor(Titanic[,i])
}

# Després dels canvis, analitzem la nova estructura del joc de dades
str(Titanic)
```

```
## 'data.frame':    1306 obs. of  11 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
```

A continuació podem buscar outlires a les variables numèriques: Age, SibSp, Parch, Fare

```
boxplot.stats(Titanic$Age)$out
```

```
##  [1]  2.00 58.00 55.00  2.00 66.00 65.00  0.83 59.00 71.00 70.50  2.00
## [12] 55.50  1.00 61.00  1.00 56.00  1.00 58.00  2.00 59.00 62.00 58.00
## [23] 63.00 65.00  2.00  0.92 61.00  2.00 60.00  1.00  1.00 64.00 65.00
## [34] 56.00  0.75  2.00 63.00 58.00 55.00 71.00  2.00 64.00 62.00 62.00
## [45] 60.00 61.00 57.00 80.00  2.00  0.75 56.00 58.00 70.00 60.00 60.00
## [56] 70.00  0.67 57.00  1.00  0.42  2.00  1.00  0.83 74.00 56.00 62.00
## [67] 63.00 55.00 60.00 60.00 55.00 67.00  2.00 76.00 63.00  1.00 61.00
## [78] 64.00 61.00  0.33 60.00 57.00 64.00 55.00  0.92  1.00  0.75  2.00
## [89]  1.00 64.00  0.83 55.00 55.00 57.00 58.00  0.17 59.00 55.00 57.00
```

```
boxplot.stats(Titanic$SibSp)$out
```

```
##  [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3
## [36] 5 4 3 4 8 4 3 4 8 4 8 3 4 5 3 4 8 4 8 4 3 3
```

```
boxplot.stats(Titanic$Parch)$out
```

```
##   [1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2 1
##  [36] 2 1 1 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1
##  [71] 1 2 1 2 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2
## [106] 2 3 4 1 2 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1
## [141] 2 1 1 2 5 2 1 1 1 2 1 5 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 1 3 2 1 1 1
## [176] 1 2 1 2 3 1 2 1 2 2 1 1 2 1 2 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1
## [211] 1 5 2 1 1 1 1 3 1 2 2 1 2 1 2 1 2 4 1 1 2 1 1 1 4 6 2 3 1 1 2 2 2 1 1
## [246] 2 5 2 3 2 1 1 1 2 1 2 2 2 1 2 1 1 2 1 2 1 2 1 2 2 1 1 1 1 1 2 1 1 2 1
## [281] 1 1 2 1 2 9 1 1 1 2 2 2 1 9 1 1 2 2 1 1 2 1 1 1 1 1 1
```

```
boxplot.stats(Titanic$Fare)$out
```

```
##   [1]  71.2833 263.0000 146.5208  82.1708  76.7292  83.4750  73.5000
##   [8] 263.0000  77.2875 247.5208  73.5000  77.2875  79.2000  66.6000
##  [15]  69.5500  69.5500 146.5208  69.5500 113.2750  76.2917  90.0000
##  [22]  83.4750  90.0000  79.2000  86.5000 512.3292  79.6500 153.4625
##  [29] 135.6333  77.9583  78.8500  91.0792 151.5500 247.5208 151.5500
##  [36] 110.8833 108.9000  83.1583 262.3750 164.8667 134.5000  69.5500
##  [43] 135.6333 153.4625 133.6500  66.6000 134.5000 263.0000  75.2500
##  [50]  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000 120.0000
##  [57] 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042  91.0792
##  [64]  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792
##  [71] 106.4250  71.0000 106.4250 110.8833 227.5250  79.6500 110.8833
```

```
##  [78]  79.6500  79.2000  78.2667 153.4625  77.9583  69.3000  76.7292
##  [85]  73.5000 113.2750 133.6500  73.5000 512.3292  76.7292 211.3375
##  [92] 110.8833 227.5250 151.5500 227.5250 211.3375 512.3292  78.8500
##  [99] 262.3750  71.0000  86.5000 120.0000  77.9583 211.3375  79.2000
## [106]  69.5500 120.0000  93.5000  83.1583  69.5500  89.1042 164.8667
## [113]  69.5500  83.1583  82.2667 262.3750  76.2917 263.0000 262.3750
## [120] 262.3750 263.0000 211.5000 211.5000 221.7792  78.8500 221.7792
## [127]  75.2417 151.5500 262.3750  83.1583 221.7792  83.1583  83.1583
## [134] 247.5208  69.5500 134.5000 227.5250  73.5000 164.8667 211.5000
## [141]  71.2833  75.2500 106.4250 134.5000 136.7792  75.2417 136.7792
## [148]  82.2667  81.8583 151.5500  93.5000 135.6333 146.5208 211.3375
## [155]  79.2000  69.5500 512.3292  73.5000  69.5500  69.5500 134.5000
## [162]  81.8583 262.3750  93.5000  79.2000 164.8667 211.5000  90.0000
## [169] 108.9000
```

En general, sembla que hi ha forces outliers a les variables numèriques. Tanmateix, si mirem bé els valors que prenen aquests outliers veiem que sembla que en tots els casos, malgrat ser outliers, poden ser dades perfectament creïbles i que siguin reals. Així doncs, opto per deixar els outliers tal com estan.

Finalment guardem el dataset i passem a un anàlisi més formal.

```
# Finalment guardem el nou dataset
write.csv(Titanic, "Titanic_clean.csv")
```

4.Anàlisi de les dades.

En aquest apartat intentaré descobrir quins eren els factors més importants per sobreviure a l'enfonsament del titanic.

4.1 Selecció dels grups de dades que es volen analitzar/comparar

Comencem seleccionant un grup interessant homes vs. dones. En teoria hauríem d'esperar que les dones haguessin sobreviscut més ja que van ser evacuades abans.

```
#Separem per sexe
titanic.homes <- Titanic[Titanic$Sex == "male",]
titanic.dones <- Titanic[Titanic$Sex == "female",]
```

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Si tinguéssim variables continues fer una comprovació de la normalitat i de la variància seria un bona pràctica ja que ens ajudar a determinar si podem utilitzar test paramètrics, o bé, hem d'utilitzar tests no-paramètrics.

Per fer una comprovació de la normalitat podem utilitzar el test Shapiro-Wilk o fent una visualització gràfica amb les corbes Q-Q. Per comprovar que la variància és similar entre els dos grups podem utilitzar Fligner-Killeen.

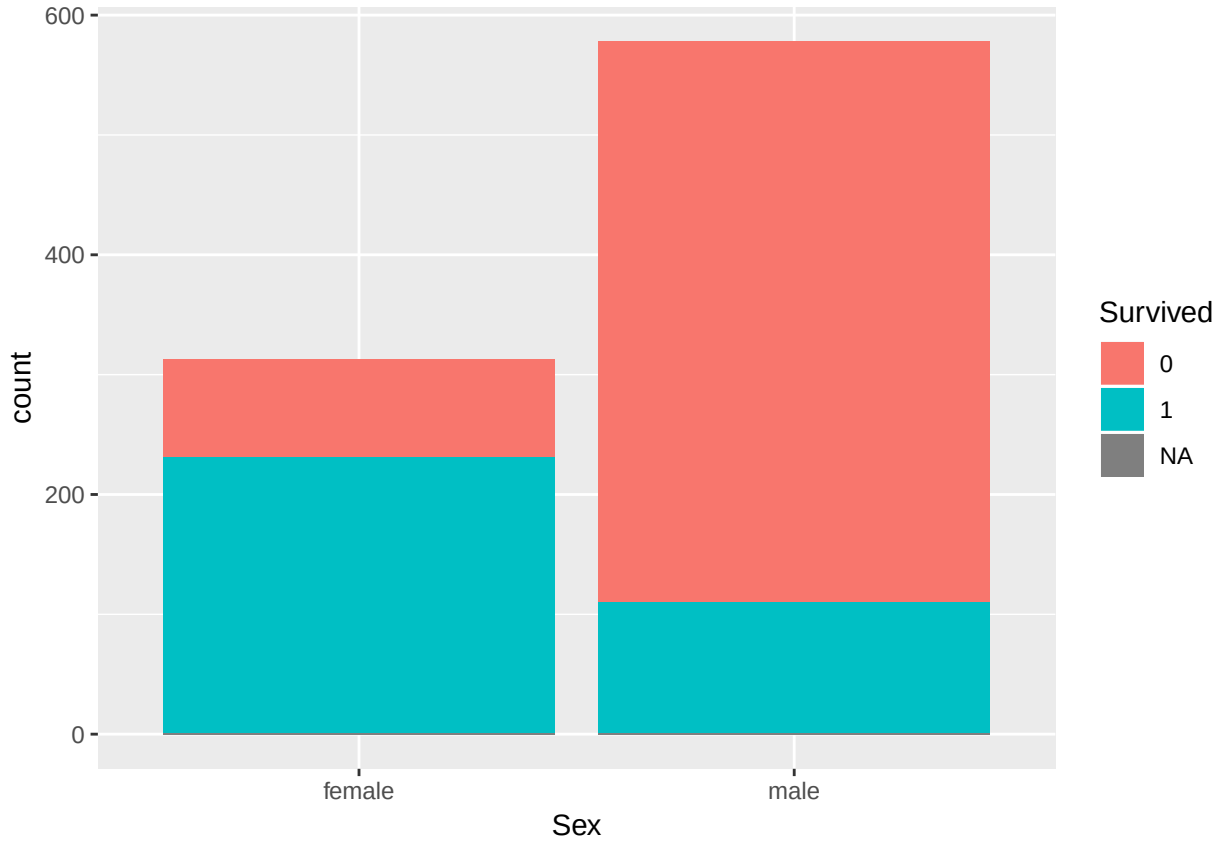
En aquest cas, la majoria de variables (incloent la dependent) són variables binàries o categòriques en general. Així doncs, no té sentit fer cap test de normalitat ja que la distribució normal és una distribució contínua i és evident que una variable categoria no pot seguir una distribució normal.

4.3 Aplicació de proves estadístiques per comparar els grups de dades

Visualització del dataset i taules de freqüències.

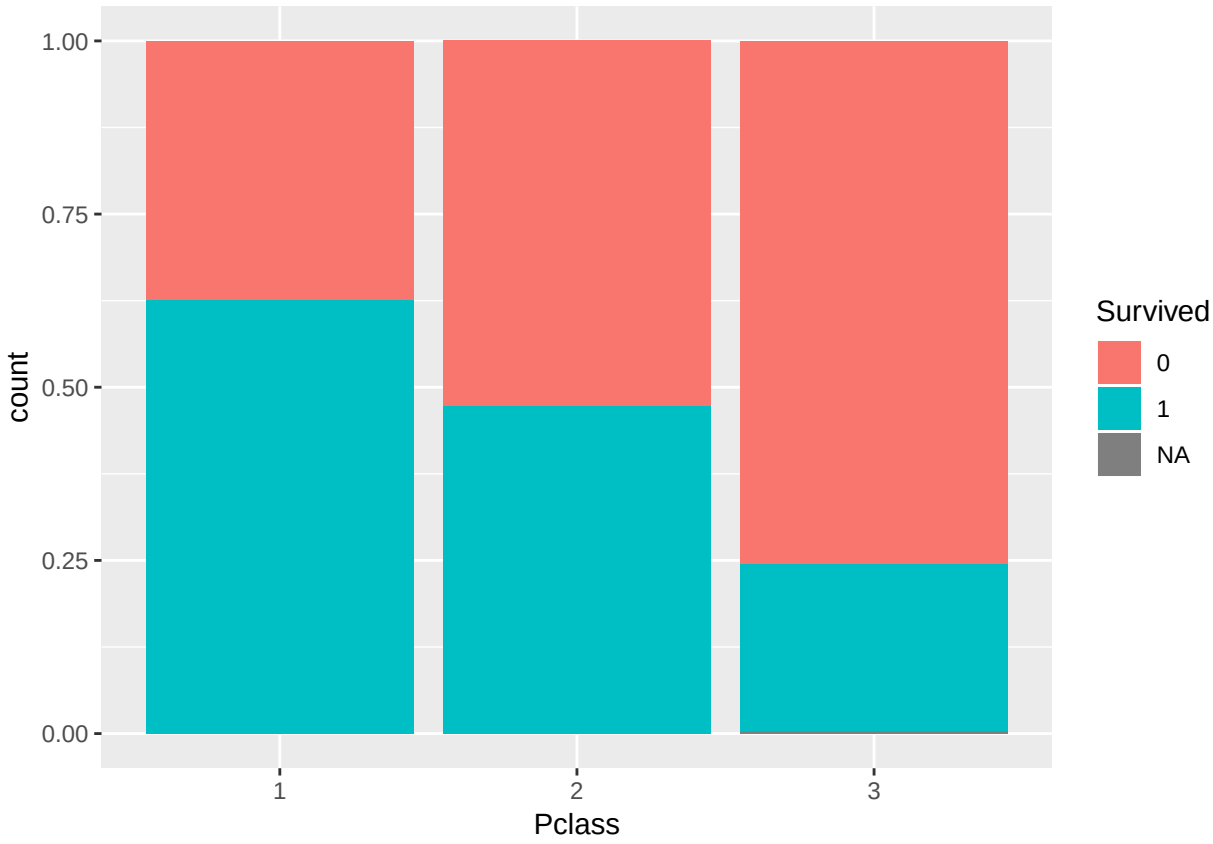
Abans de fer un anàlisi més profund, és interessant visualitzar les relacions i variables que tenim simplement amb gràfics.

```
# Visualitzem la relació entre les variables "sex" i "survival":
ggplot(data=Titanic[1:filas,],aes(x=Sex,fill=Survived))+geom_bar()
```



Un altre factor important podia ser la classe on s'anava

```
# Visualitzem la relació entre les variables "classe" i "survival":
ggplot(data = Titanic[1:filas,],aes(x=Pclass,fill=Survived))+geom_bar(position="fill")
```



Chi-square test

Tal com s'explica a l'apartat anterior per a les variables categòriques no es poden utilitzar els mateixos mètodes que per variables continues. En aquest cas si volem fer una comparació de mitjanes, els mètodes per variables contínues no són els més adients. En aquest cas, doncs, opto fer fer un chi-square test per saber si hi ha un diferència significativa en els freqüències relatives entre sexe i sobreviure.

Comencem fent una taules de freqüència que en cas de variables categòriques expressen millor la informació que les correlacions.

```
#Taula de freqüències sexe i sobreviure
t<-table(Titanic[1:filas,]$Sex,Titanic[1:filas,]$Survived)
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

```
##
##                 0        1
##   female 25.96154 74.03846
##   male   81.10919 18.89081
```

A continuació apliquem el test chi

```
#Fem el test Chi amb la taula de freqüencies que hem fet a l'apartat anterior
chisq.test(t)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  t
## X-squared = 58.934, df = 1, p-value = 1.631e-14
```

En aquest cas veiem un p-valor molt petit cosa que suggereix que podem rebutjar la hipòtesis nula i per tant que sobreviure no és independent del sexe. En aquest cas, a més, si observem la taula de freqüències veiem que les dones tenien una probabilitat més gran de sobreviure.

Regressió logísitca

Com a últim mètode d'anàlisi faré una regressió logística (recordem que la variable sobreviure és binaria).

```
#Model logit

logit <- glm(Survived ~ Pclass + Sex + Age, data = Titanic, family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age, family = "binomial",
##     data = Titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6473  -0.6631  -0.4195   0.6291   2.4279
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.540622   0.365483   9.688  < 2e-16 ***
## Pclass2     -1.115637   0.257864  -4.326 1.52e-05 ***
## Pclass3     -2.321195   0.240917  -9.635  < 2e-16 ***
## Sexmale     -2.603899   0.186879 -13.934  < 2e-16 ***
## Age         -0.033531   0.007382  -4.542 5.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.82  on 888  degrees of freedom
## Residual deviance:  804.66  on 884  degrees of freedom
##   (417 observations deleted due to missingness)
## AIC: 814.66
##
## Number of Fisher Scoring iterations: 5
```

En la taula anterior veiem que hi ha varis factors significatius per explicar el fet de sobreviure a l'accident del titanic. Primerament, i tal com ja he apuntat en apartats anteriors, sembla que ser home reduïa significativament les probabilitat de sobreviure. Això es dedueix si observem el coeficient de la variable Sex(male) que és negatiu i que té tres estrelles.

També podem veure que ser de segona o tercera classe reduïa la probabilitat de sobreviure respecte ser de primera classe.

Finalment veiem que com més jove també hi havia més probabilitat de morir.

5 Conclusions

En aquesta pràctica he treballat el dataset sobre els passatgers del Titanic. Primerament he netejat les dades tractant els valors buits i outliers. A continuació he fet un estudi preliminar per deduir quins factors dels passatgers estan associats a sobreviure la tragèdia.

En general s'ha vist que les dones tenien una probabilitat més gran de sobreviure. Això possiblement és degut a una societat masclista de l'època que va fer que primer evacuessin a les dones del vaixell.

Igualment veiem que la societat també era classista ja que la supervivència també depenia de forma crucial de la classe on es viatjava. Així la gent de segona classe tenia una probabilitat menor de sobreviure que la gent de primera classe i la gent de tercera tenir una probabilitat menor que la gent de segona i de primera.

Finalment veiem també que els joves ho tenien pitjor. Una vegada més imagino que primer es va evacuar a la gent gran del vaixell.