Práctica 2 - Limpieza y análisis de datos

Sara García Rodríguez y Javier Mateo Moreno

Tipología y ciclo de vida de los datos

ÍNDICE

I. Descripción del dataset

El dataset que hemos usado para este análisis es un registro de los supervivientes y fallecidos que iban a bordo del Titanic. De cada una de estas personas se incluye no solo si sobrevivieron o no, sino también cierta información personal recogida en otras nueve variables que nos servirán para analizar los perfiles con mayor o menor tendencia a sobrevivir. Los datos se han extraído de la plataforma Kaggle, que tiene en la actualidad una competición relativa al tratamiento de esta misma base de datos (Accesible en: https://www.kaggle.com/c/titanic).

En el contexto de un accidente como puede ser el hundimiento de un barco, asumimos socialmente que sobrevivir o no "es cosa del azar". Pero, ¿lo es realmente? Un primer análisis de los datos disponibles nos indica ciertas tendencias o patrones en los perfiles de quienes sobrevivieron. Por ejemplo, tal y como vemos en los gráficos 1 y 2, a primera vista parece que las mujeres sobrevivieron más que los hombres, y que a mejor estatus socioeconómico de un individuo, mayores sus probabilidades de sobrevivir. El interés del estudio de esta base de datos consiste precisamente en eso: al permitirnos discernir qué características favorecieron la supervivencia de las personas a bordo del Titanic, nos habla de las tendencias de comportamiento humano en situaciones límites.

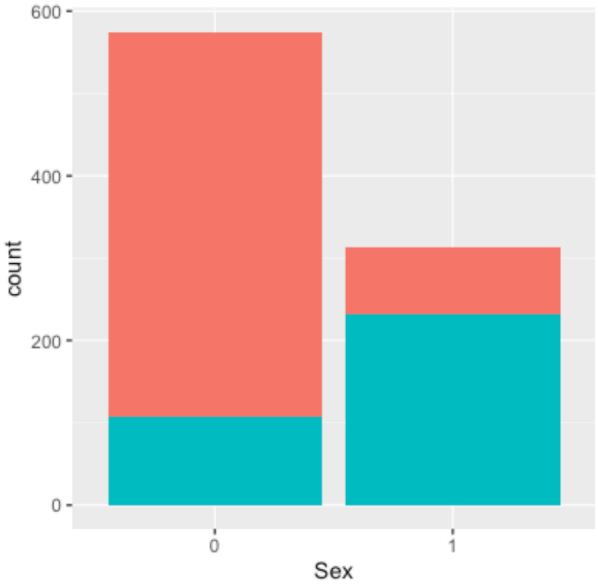


Gráfico 1 - Gráfico de elaboración propia

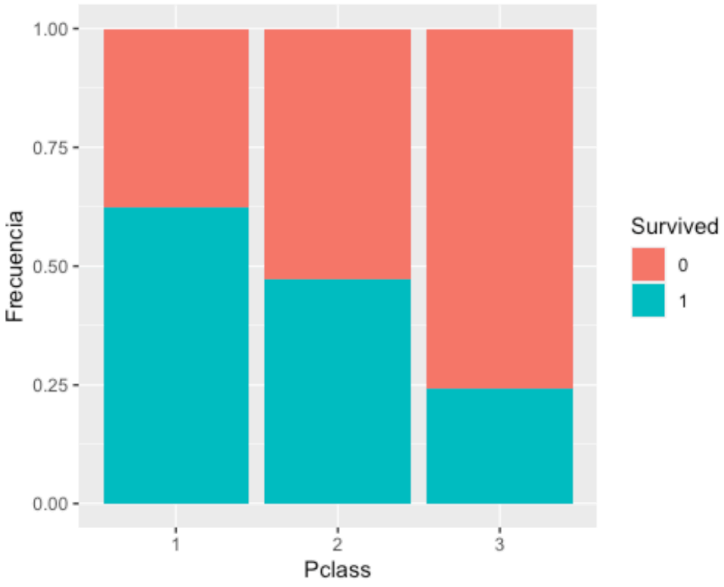


Gráfico 2 - Gráfico de elaboración propia

En suma, el dataset responde la pregunta general: "¿Qué características de una persona favorecieron su supervivencia en el accidente del Titanic?". El análisis de estos datos también nos permite responder a cuestiones más concretas como: ¿hasta qué punto el género y/o la edad de una persona afectó a su supervivencia?, ¿existe una correlación entre pagar más por el billete y aumentar las opciones de sobrevivir? En nuestro caso concreto, hemos fijado tres objetivos distintos para nuestro análisis:

- Análisis de la relación entre el precio del billete y la supervivencia, haciendo uso de correlaciones y de una regresión simple.
- Determinar cómo afectan las variables demográficas de las personas (la edad, el género, la clase social y la familia) a sus posibilidades de sobrevivir, mediante regresiones logísticas.
- Realizar un análisis estadístico para buscar cuál es el perfil o cuáles son los perfiles con más tendencia a sobrevivir, usando métodos de asociación.

II. Integración y selección de los datos de interés

El dataset original recoge información relativa a 891 pasajeros en las siguientes 12 variables:

- PassengerID - Identificador único de cada pasajero.
- Survived - Variable dummy que nos indica si la persona ha sobrevivido (Survived=1) o no (Survived=0).
- Pclass - Variable factorial que nos indica cuál era la clase del billete de cada pasajero. Puede tomar los valores 1 (primera clase), 2 (segunda clase) o 3 (tercera clase) y, tal y como se explica en la web de Kaggle, se puede considerar como una proxy del estatus socioeconómico de la persona.
- Name - Variable de tipo string que indica el nombre del pasajero
- Sex - Codificada inicialmente como 'male' y 'female', indica el género del pasajero.
- Age - Variable numérica que indica la edad del pasajero.
- SibSp - Indica el número de hermanos y/o esposa a bordo del barco.
- Parch - Indica el número de padres o hijos a bordo del barco.
- Ticket - Identificador del número de billete.
- Far.- Variable numérica que indica el coste del pasaje.
- Cabin - Esta variable indica la o las cabinas en las que estaban los pasajeros.
- Embarked - Indica el puerto de embarque, siendo C = Cherbourg, Q = Queenstown y S = Southampton.

Siguiendo los requerimientos de nuestros objetivos de análisis, hemos eliminado de la base de datos tres variables que no nos resultaban de interés analítico: Ticket, Cabin, y Embarked. La información que recogen estas variables podría ser de utilidad para un proyecto más amplio que tenga la capacidad de, por ejemplo, cruzar la supervivencia con la arquitectura del barco, pero estos no son los objetivos que tenemos en este análisis, que requiere más bien de la información que se recoge en las variables sociodemográficas.

Con lo que respecta a la integración de los datos, hemos creado una nueva variable llamada 'Fam' que nos indica si la persona tiene o no familia. Hemos comprobado que lo que nos interesaba saber era si el hecho de tener familiares a bordo alteró de alguna manera la expectativa de supervivencia de los pasajeros, y no necesitábamos saber cuántos familiares tenían o de qué tipo. Por eso, hemos cruzado las variables SibSp y Parch en una nueva variable ('Fam'), de forma que si alguna de las dos variables originales tenían un valor distinto de cero, la variable Fam tendría un valor 1 (indicando que el pasajero sí tenía familia), y si ambas variables originales registraban un cero, entonces Fam valdría también 0 (y, por tanto, esa persona no tendría familia). Con esta nueva variable creada, y habiendo comprobado que funcionaba correctamente, eliminamos las variables SibSp y Parch, que ya no nos resultaban de interés para el análisis.

```
df3 <- transform(df2, Fam = ifelse(SibSp==0 & Parch ==0, 0, 1))
summary(df3$Fam)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3975  1.0000  1.0000

unique(df3$Fam[df3$SibSp==0 & df3$Parch == 0])

## [1] 0
```

III. Limpieza de los datos

Hemos realizado el proceso de limpieza de datos en tres pasos distintos. En primer lugar, identificamos y tratamos los valores nulos y vacíos; en segundo lugar, identificamos y tratamos los outliers o valores extremos; y por último, realizamos diversos ajustes de limpieza en las variables para poder trabajar con claridad.

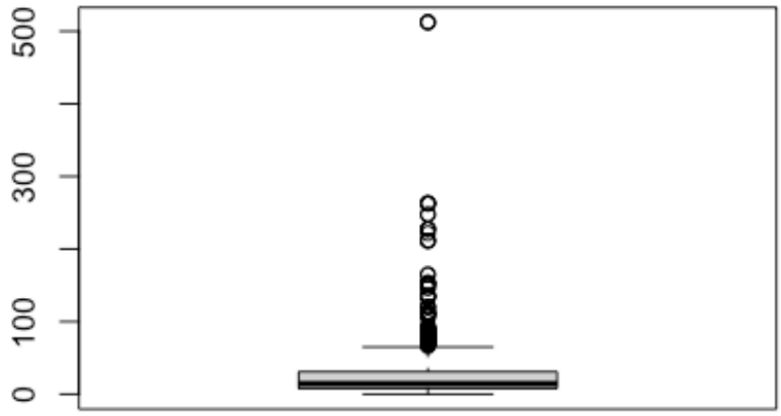
3.1. Valores nulos y vacíos

Tras una primera comprobación vemos que tenemos 177 valores vacíos en la variable 'Age'. Para tratarlos, decidimos sustituir estos valores por la media global de la edad, que equivale en este caso a 29,7 años. Una vez este cambio fue efectivo, comprobamos que ya no quedaba ningún valor perdido.

```
# Vemos que tenemos valores vacíos en la variable Age; los reemplazamos por
la media de la edad
df$Age[is.na(df$Age)] <- mean(df$Age, na.rm=T)

# Comprobamos que no quedan valores vacios en las variables elegidas
colSums(df=="")

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

3.2. Valores extremos

Identificamos valores atípicos en la variable 'Fare'. Al representar los registros haciendo uso de un boxplot, se consideran como atípicos un continuo de valores que, en realidad, no corresponden a valores extremos. En tanto que los billetes podían ser de primera, segunda o tercera clase, existe una alta variación en los precios a los que estos se adquirieron. Como la cantidad de billetes de cada clase no es la misma (habiendo muchos más de tercera clase que de primera), se identifican como outliers los valores más altos que son, en realidad, valores que corresponden a los billetes más caros de primera clase. Sin embargo, sí que existe un valor anómalo que se escapa con mucha diferencia del precio de los demás (Fare=512.33). Tras identificarlo, decidimos que lo más apropiado es eliminar el registro.



3.3. Ajustes de limpieza

Para acabar de ajustar nuestra base de datos hemos realizado otras tareas de limpieza. En primer lugar, redondeamos la variable Fare a dos decimales (ya que es un precio, ajustamos su valor a céntimos) y la variable Age a números enteros. A continuación, cambiamos los registros de la variable Age, 'male' y 'female' a 0 (male) y 1 (female). Por último, analizamos la estructura de las variables y vimos que las variables dummies estaban registradas como números enteros, por lo que las factorizamos y las registramos como variables con dos factores (Survived, Fam, Sex) o con tres (variable Pclass).

```
cols<-c("Survived", "Pclass", "Sex", "Fam")
for (i in cols){
  df_final[,i] <- as.factor(df_final[,i])
}

str(df_final)
```

```
## 'data.frame':    888 obs. of  8 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)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques
Heath (Lily May Peel)" ...
##  $ Sex        : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Age        : int  22 38 26 35 35 29 54 2 27 14 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Fam        : Factor w/ 2 levels "0","1": 2 2 1 2 1 1 1 2 2 2 ...
```

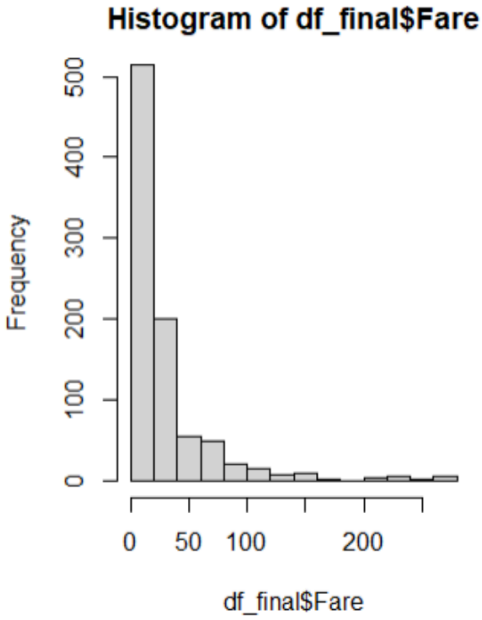
IV. Análisis de los datos

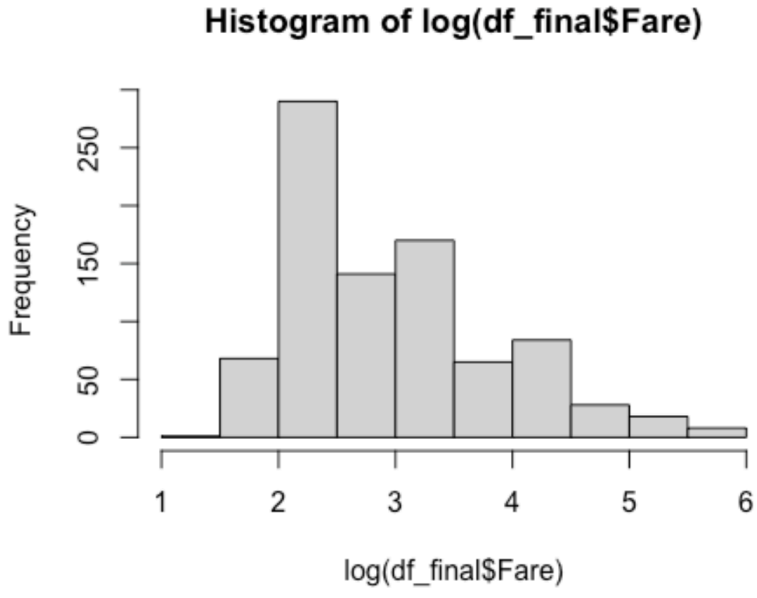
Como especificamos en la descripción de este análisis, tenemos tres objetivos de análisis que se han realizado de forma independiente, cada uno con las pruebas que se estimaron necesarias. Describiremos nuestro análisis siguiendo ese orden, para más detalles sobre el proceso de análisis pueden referirse al repositorio de Github, archivo 'Code' (con el código en bruto) o 'Code-RMarkdown', con el informe.

4.1. Relación entre el precio del billete y la supervivencia

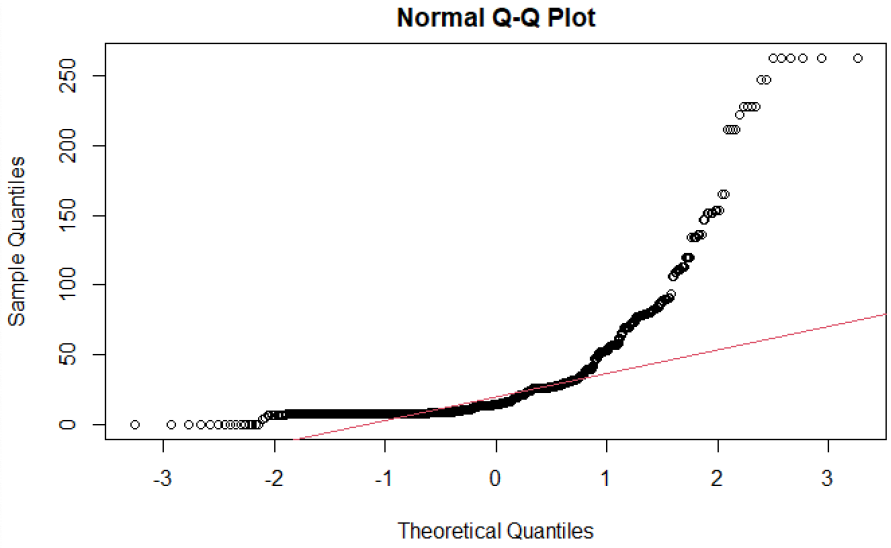
Para esta parte del análisis partimos de la premisa de que cuando uno paga por un billete de primera clase, espera adquirir no solo más lujo y/o exclusividad en el servicio sino también más seguridad. Si esto fuese así, deberíamos poder observar una correlación entre el precio del billete y la supervivencia de los pasajeros. Los análisis posteriores nos permitirán discernir si tras una posible correlación en este aspecto hay o no causalidad.

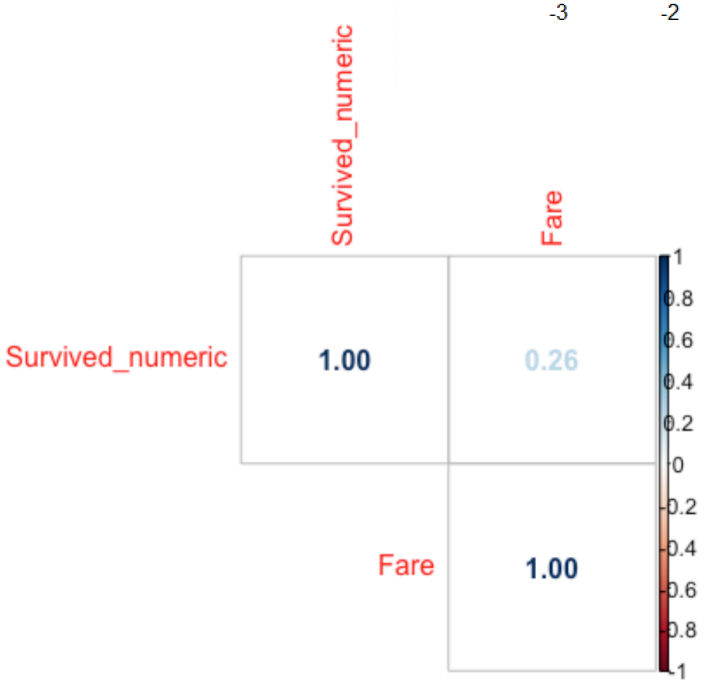
En este caso, los datos de interés son solamente las variables 'Fare' y 'Survived'. Analizando la distribución de Fare, vemos que la disparidad en la dimensión de los registros provoca que se requiera de una escala logarítmica para comprobar la normalidad. Tras implementarla, vemos que se aproxima más a una distribución normal, aunque con una tendencia hacia la izquierda (hacia valores menores en el precio) que se explica con el hecho de que los billetes más baratos son los de tercera clase, y estos son, a su vez, los que más abundan.





Tras hacer también uso de un gráfico QQ-plot para comprobar estas ligeras asimetrías en la normalidad, pasamos a realizar una correlación entre la variable 'Fare' y 'Survived', que hemos leido como numérica para esta ocasión.





Habiendo testeado la correlación, completamos nuestro análisis implementando una regresión logística simple entre las dos variables para testar la capacidad predictiva del precio del billete. La variable dependiente sería 'Survived' que, al ser dummy, se presentaría en la ecuación tal que lnY. La variable independiente sería 'Fare'. Para comprobar la independencia de los resultados del modelo hacemos, uso del método de validación cruzada. Una vez tenemos hechas las dos regresiones (gráfico 3 y 4) comporamos que no hay diferencia en estimadores, p-value y demás medidas. Consecuentemente, podemos considerar como precisos los resultados de la regresión original.

```
regresion_simple <- glm(formula = Survived~Fare, data = df_final, family = binomial)

summary(regresion_simple)

```

```

Call:
glm(formula = Survived ~ Fare, family = binomial, data = df_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4899  -0.8885  -0.8531   1.3458   1.5941

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.941125   0.095192  -9.887  < 2e-16 ***
Fare         0.015189   0.002236   6.794 1.09e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1180.9  on 887  degrees of freedom
Residual deviance: 1117.6  on 886  degrees of freedom
AIC: 1121.6

Number of Fisher Scoring iterations: 4
```

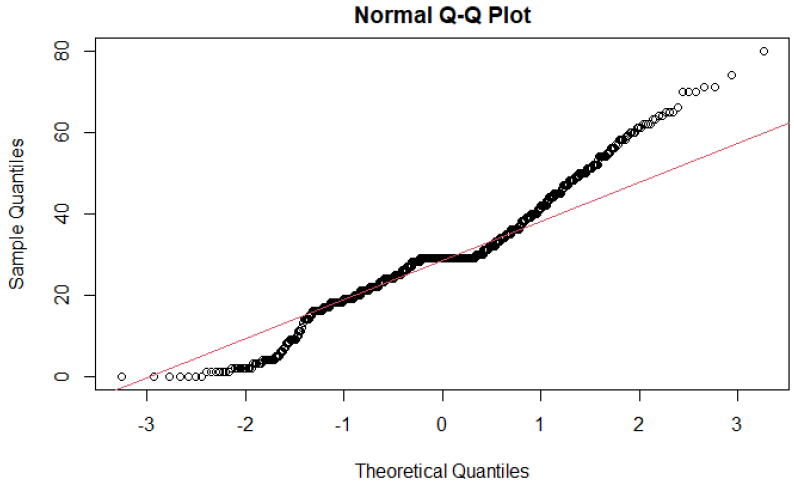
Gráfico 3: Resultados de la regresión original en R

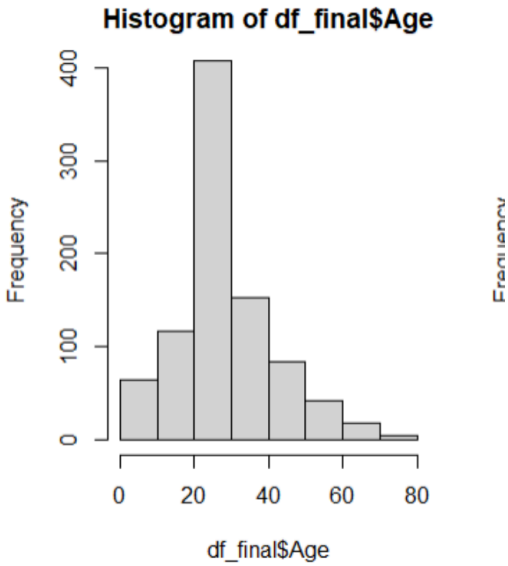
```
library("lattice")
library("caret")
library("ggplot2")

data_ctrl <- trainControl(method = "cv", number = 5)

regresion_simple_cross <- train(Survived ~ Fare,   # model to fit
                    data = df_final,
                    trControl = data_ctrl,             # folds
                    method = "glm",     # specifying regression model
                    family=binomial(), # specifying regression model
                    na.action = na.pass)

summary(regresion_simple_cross)

# Comprobamos que tanto los coeficientes como el valor de AIC es igual en los dos modelos, y por tanto
mantenemos nuestro modelo original.

```
```

Call:
NULL

Deviance Residuals:
    Min   1Q   Median   3Q   Max
-2.4899 -0.8885 -0.8531   1.3458   1.5941

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.941125   0.095192 -9.887 < 2e-16 ***
Fare   0.015189   0.002236   6.794 1.09e-11 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1180.9 on 887 degrees of freedom
Residual deviance: 1117.6 on 886 degrees of freedom
AIC: 1121.6

Number of Fisher Scoring iterations: 4
```

Gráfico 4: Resultados de la regresión para la validación cruzada

4.2. Influencia de las variables demográficas en la supervivencia

Para nuestra segunda parte del análisis nos hemos centrado en cómo los factores internos a la persona (su edad, su género, su clase y la presencia o no de familia a bordo) afectan a las probabilidades de sobrevivir de esta. En este caso, las variables de interés son: Fam, Sex, Age, Pclass y Survived. Comenzamos haciendo las comprobaciones de distribución como en el caso anterior.





A excepción de la variable 'Age', todas las demás son factoriales. Por tanto, implementamos un modelo de regresión logística múltiple para comprobar el potencial explicativo de cada variable. Al igual que en el caso interior, optamos por implementar una validación cruzada para comprobar la efectividad de nuestro modelo. Al crear un nuevo modelo de regresión sobre un subgrupo aleatorio de datos, obtenemos unos resultados en cuanto a estimadores, p-value, AIC y demás mediciones que son iguales a los de la regresión original, quedando así probada la independencia de los resultados con respecto de la muestra.

```
regresion_multiple <- glm(formula = Survived~Age+Sex+Pclass+Fam, data = df_final, family = binomial)
summary(regresion_multiple)

```

```
Call:
glm(formula = Survived ~ Age + Sex + Pclass + Fam, family = binomial,
    data = df_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6326  -0.6498  -0.4265   0.6237   2.4271

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.939319   0.352065   2.668  0.00763 **
Age         -0.033740   0.007533  -4.479 7.50e-06 ***
Sex1         2.639080   0.194296  13.583  < 2e-16 ***
Pclass2     -1.095095   0.259563  -4.219 2.45e-05 ***
Pclass3     -2.312422   0.244644  -9.452  < 2e-16 ***
Fam1        -0.077492   0.188340  -0.411  0.68074
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1180.89  on 887  degrees of freedom
Residual deviance:  801.61  on 882  degrees of freedom
AIC: 813.61

Number of Fisher Scoring iterations: 5
```

Gráfico 5: Resultados de la regresión múltiple en R

En un primer análisis comprobamos que una de las variables ('Fam') no tenía significación estadística, por lo que procedimos a construir una nueva regresión omitiendo tal variable con el fin de mejorar la capacidad predictiva de nuestro modelo (capacidad que medimos con un gráfico pROC). Al realizar las comprobaciones, vemos que eliminar la variable 'Fam' no mejora sino que empeora ligeramente la capacidad predictiva del modelo, por lo que la tendremos en cuenta a la hora de evaluar nuestros resultados. Completados estos dos apartados del análisis, podemos comprender ya: i. Si existe o no una correlación entre el precio del billete y la supervivencia del sujeto, y de qué forma estas variables se afectan entre sí, y ii. Cómo afectaron las variables identitarias de los pasajeros a su supervivencia. Ahora, combinaremos las variables mediante métodos de asociación para analizar qué perfiles han sido más tendenties a sobrevivir.

4.3. Análisis de los perfiles con más probabilidades de sobrevivir

En esta última parte del análisis generamos las reglas de asociación entre las distintas variables categóricas para discernir qué perfiles tuvieron más tendencia a sobrevivir. Las distintas variables que tenemos no operan por separado, sino que se encarnan en sujetos concretos que representan combinaciones específicas de los distintos factores. Por ejemplo, partiendo de que el género del sujeto afecta de forma X y la clase social de forma Y, no tenemos a un sujeto que represente la variable género y a otro que represente la variable clase, sino que ambas variables operan de forma conjunta en personas concretas. Por ese motivo, crear reglas de asociación es un proceso interesante para acabar de comprender cómo funciona la tendencia a sobrevivir.

Así, generamos un set de reglas con distinto soporte, confianza y lift. El soporte nos indica cuántas vecen se han encontrado tales reglas concretas en el dataset; por ejemplo, cuántas veces ha ocurrido que un hombre (sex=0), de clase media (Pclass=2) y sin familia (Fam=0) haya sobrevivido. La confianza nos habla de la probabilidad de que el outcome se produzca (por ejemplo, que se sobreviva) dada la presencia de las condiciones iniciales (siguiendo el ejemplo, dado sex=0, Pclass=2 y Fam=0). En estos dos parámetros, cuánto mayor es el valor mayor es la fuera de la regla de asociación. Por último, el parámetro 'lift' nos indica cuánto de aleatoriedad hay en las reglas, por lo que nos interesa que su valor sea lo más bajo posible. Así, construimos las cinco reglas con un carácter asociativo más fuerte.

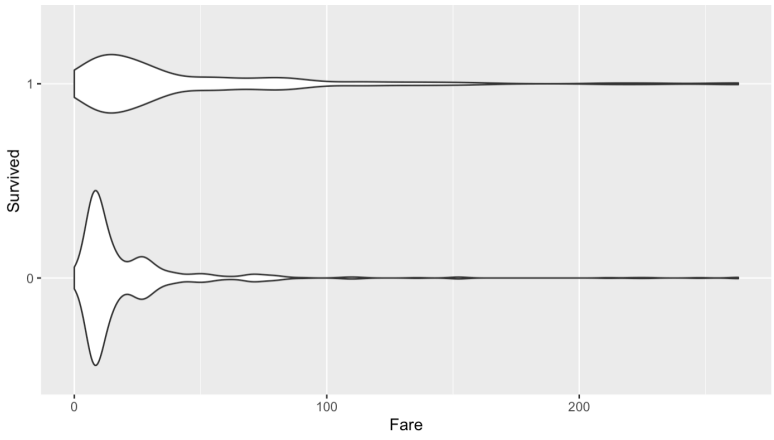
```
library(arules)

df_final1<- select(df_final, "Fam", "Sex", "Pclass", "Survived")
titanic_rules <- apriori(df_final1, parameter = list(support = 0.01, confidence = 0.5))
inspect(head(sort(titanic_rules, by = "confidence"), 5))
```

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{Fam=0, Pclass=1, Survived=0}	=>	{Sex=0}	0.05630631	0.9803922	0.05743243	1.514066	50
[2]	{Fam=0, Sex=1, Pclass=1}	=>	{Survived=1}	0.03603604	0.9696970	0.03716216	2.540091	32
[3]	{Sex=1, Pclass=1}	=>	{Survived=1}	0.10135135	0.9677419	0.10472973	2.534970	90
[4]	{Fam=1, Sex=1, Pclass=1}	=>	{Survived=1}	0.06531532	0.9666667	0.06756757	2.532153	58
[5]	{Pclass=1, Survived=0}	=>	{Sex=0}	0.08671171	0.9625000	0.09009009	1.486435	77

V. Representación de los resultados

Siguiendo la línea de nuestro análisis, analizaremos los resultados en tres apartados para después pasar a sacar unas conclusiones generales que resuelvan las incógnitas que nos planteamos en este estudio.

5.1. Relación entre el precio del billete y la supervivencia



Una primera representación de los datos nos indicaba una tendencia clara: a más caro el billete más sobreviven las personas. De hecho, la tendencia era clara aún disminuyendo notablemente la cantidad de billetes que hay con un precio mayor. Esta relación puede observarse con más claridad haciendo uso de un 'Ridgeline plot' (gráfico 6):

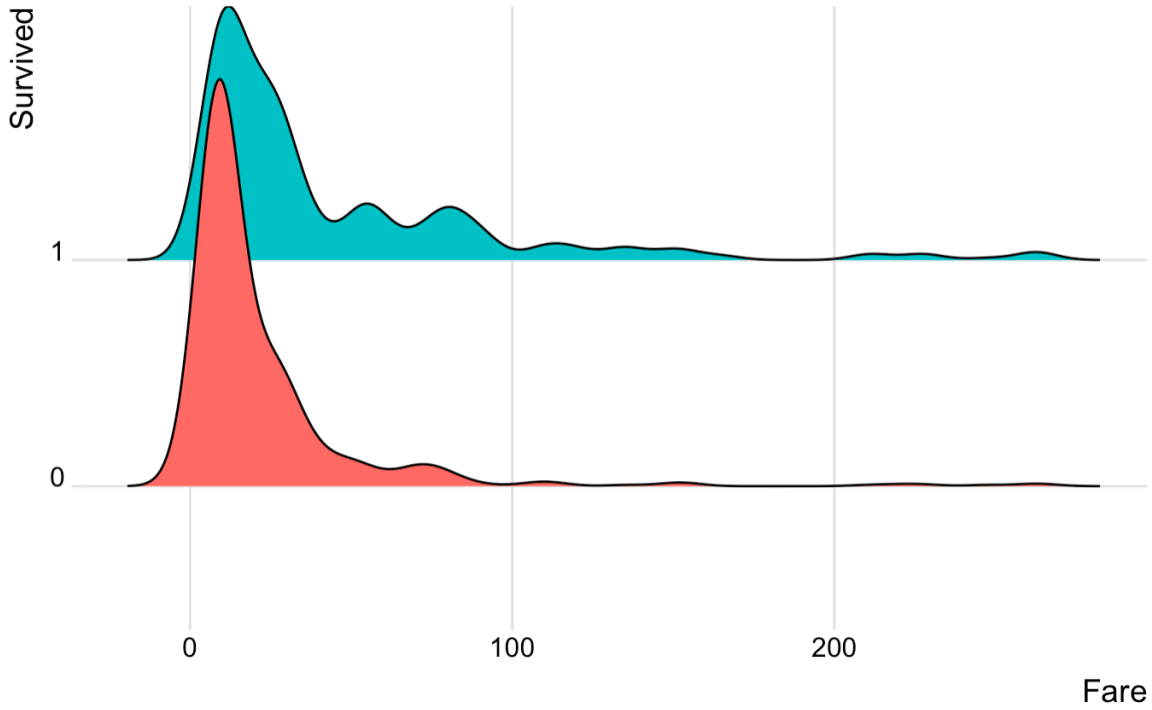
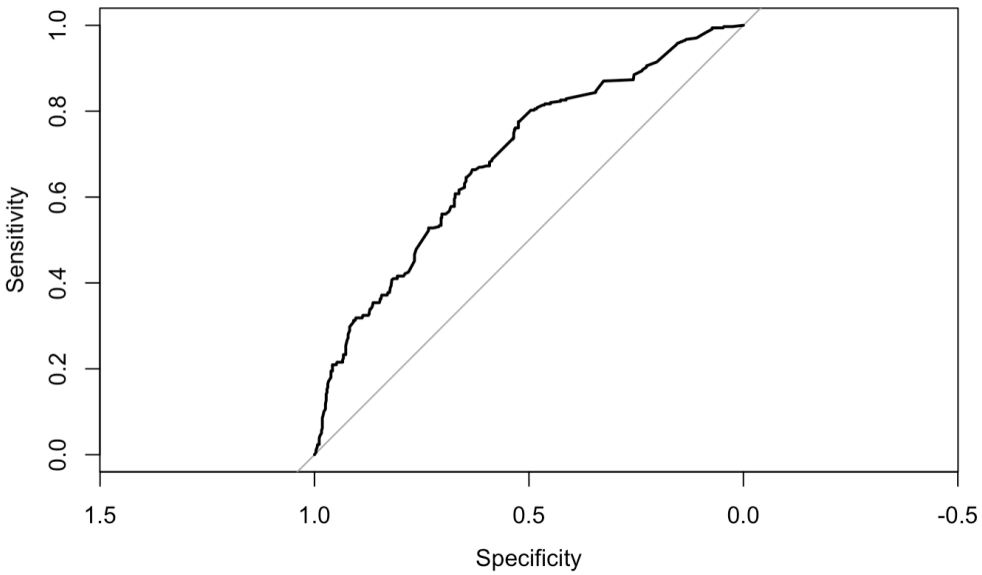


Gráfico 6: Ridgeline plot entre Fare y Survived

Con esta representación vemos que en los billetes con un precio más barato (que son la absoluta mayoría) ha fallecido más gente de la que ha sobrevivido (y de ahí el tamaño del área roja frente a la azul). Y a su vez, a medida que aumentan los precios el tamaño del área azul siempre supera a la roja, siendo así mayor la cantidad de personas que sobrevivieron que la de fallecidos. Al calcular la correlación, sin embargo, obtenemos un valor muy bajo de tan solo **0,26**. Con la enorme variabilidad que existe entre los registros agrupados bajo la variable Fare, pensamos que el valor de esta correlación podría no ser demasiado preciso, por lo que realizamos una regresión entre ambas variables para hacernos una idea más precisa de su relación.

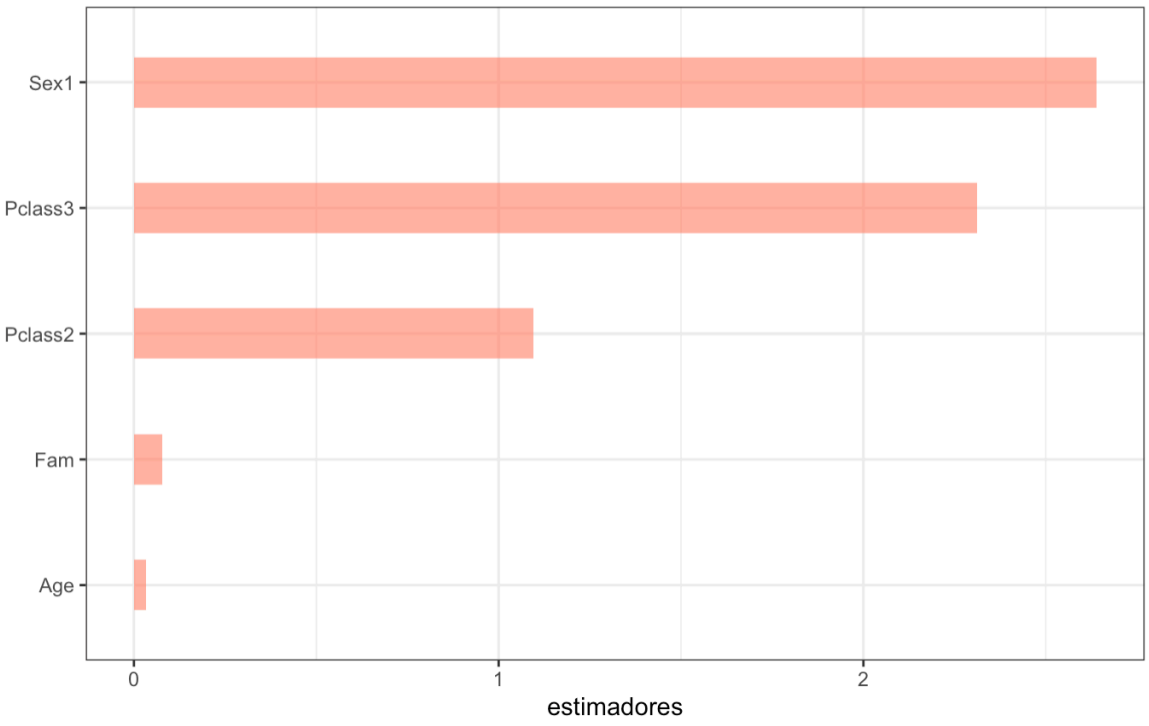
Tras hacer la regresión, vemos que el estimador tiene un valor de 0.015189 y un alto nivel de significación, por lo que sabemos que el precio del billete tiene un efecto positivo y significativo sobre las probabilidades de sobrevivir de una persona. Este estimador no sindica que cuando el precio aumenta en una unidad, las probabilidades de sobrevivir aumentan en un 1,51%. Aunque el efecto de Fare sobre Survived sea pequeño, es estadísticamente significativo, por lo que haremos uso de los análisis posteriores para entender con más precisión las causas de sobrevivir. El p-value es un valor pequeño, aún así comprobamos de forma más precisa la capacidad de discriminación del modelo mediante el gráfico pROC. El valor del área sobre la curva es de 0.6894, por lo que podemos considerar que el modelo discrimina de manera adecuada aunque hay mucho espacio para una mejora predictiva.

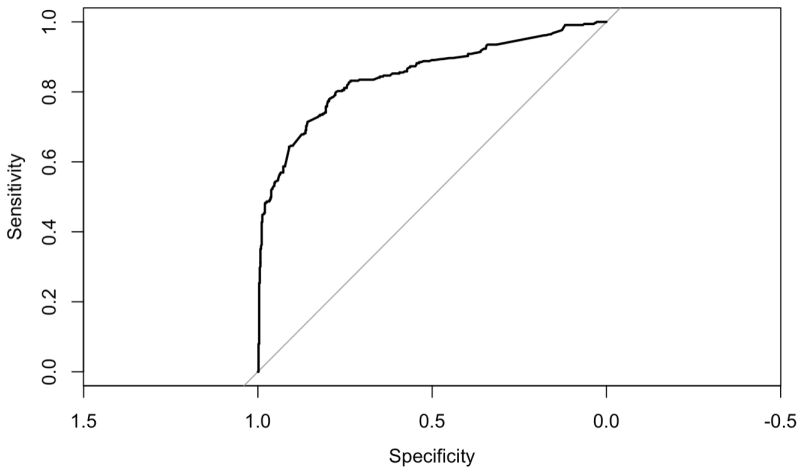


5.2. Relación entre las variables demográficas y la supervivencia

Tras implementar nuestro modelo de regresión con Survived como variable dependiente y Age, Sex, Fam y Pclass como variables independientes, comprobamos que todas menos Fam tienen un alto nivel de significancia estadística. En primer lugar tenemos la variable 'Age', cuyo estimador tiene un valor de -0.033740. Esto nos indica que la edad tiene un efecto negativo sobre las posibilidades de sobrevivir: un aumento de un año en la edad del sujeto, disminuye sus probabilidades de sobrevivir en un 3,37%. Este resultado entra en concordacia con las tendencias que se aprecian en la graficación conjunta de ambas variables.

Por otra parte, vemos que los esitmadores Pclass2 y Pclass3 tienen un valor negativo, y por tanto, será tener un billete de primera clase (Pclass=1) lo que aumente notablemente las probabilidades de sobrevivir de un pasajero. Tenemos por último el indicador de la variable Sex, con un nivel de significación muy alto y un p-value muy bajo, que nos indica que si la persona en cuestión es mujer (si Sex=1), las probabilidades de sobrevivir aumentan en un 26,39%, comprobándose así la tendencia que ya habíamos intuido en nuestra primera visualización de los datos. De todos los estimadores, parece que el género es el que tiene un mayor efecto sobre la variable dependiente. Representamos gráficamente los estimadores para comprobar la magnitud del efecto que cada uno de ellos tiene sobre las probabilidades de sobrevivir de una persona.



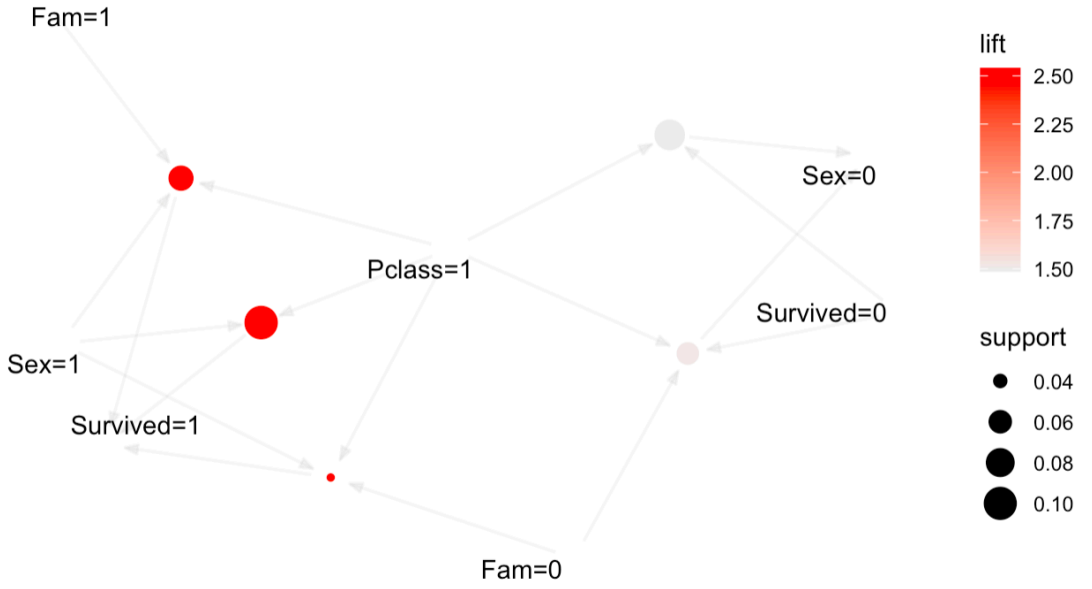


Comprobamos de nuevo la eficacia del modelo haciendo uso de un gráfico pROC. El valor del área bajo la curva es de 0.8479, por lo que se estima que el modelo discrimina de forma excelente. Hemos realizado otra comprobación con un modelo auxiliar que no incluía la variable Fam (ya que esta no resultó ser estadísticamente significante), pero la capacidad predictiva en lugar de mejorar empeoró ligeramente, siendo el área bajo la curva pROC auxiliar de 0.8474. Por tanto, hacemos uso de nuestro modelo original para sacar nuestras conclusiones.

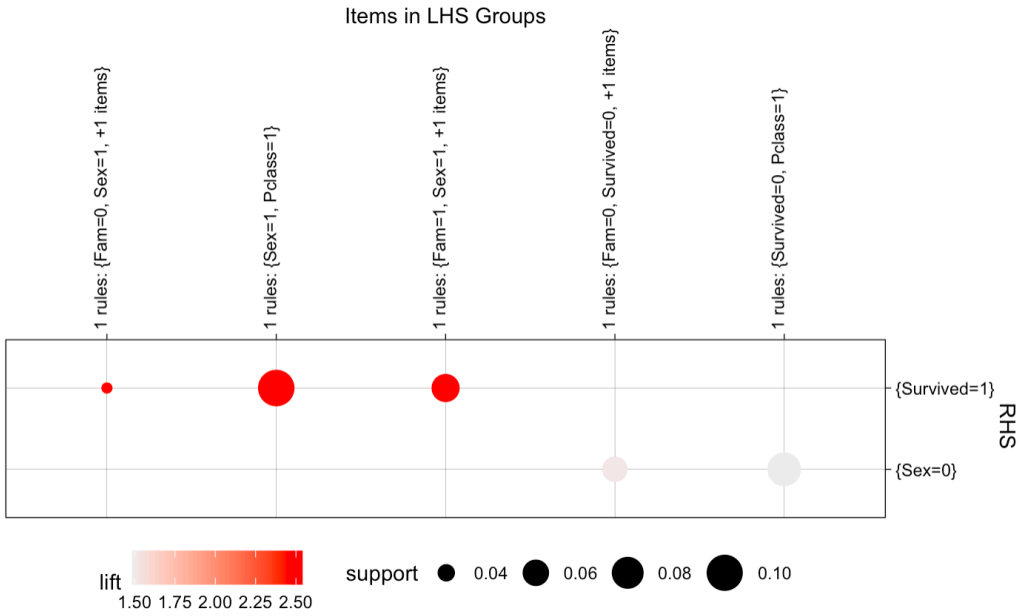
5.3. Perfiles con mayor tendencia a sobrevivir

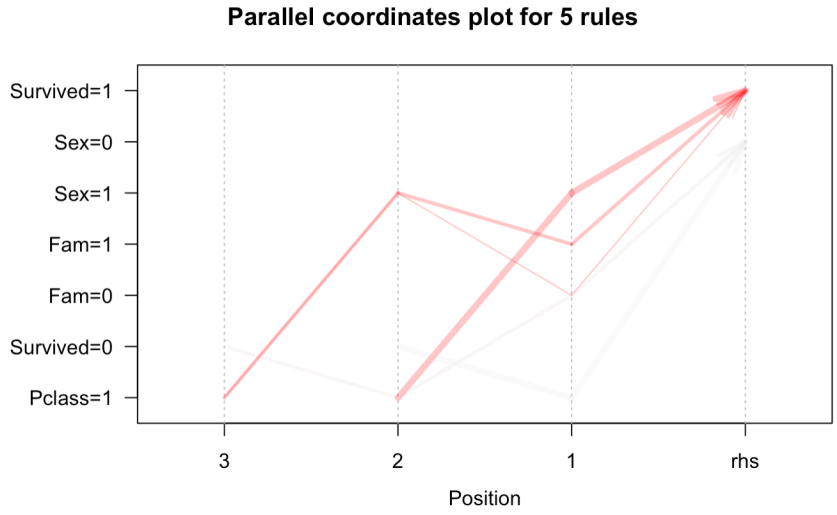
Las cinco reglas de asociación más fuertes que hay entre los registros de esta base de datos son:

1. Si no tiene familia, es de clase 1 y ha muerto; entonces es hombre
2. Si no tiene familia, es mujer y es de clase 1; entonces ha sobrevivido
3. Si es mujer y de clase 1; ha sobrevivido
4. Si tiene familia, es mujer, y es de clase 1; ha sobrevivvido
5. Si es de clase 1 y ha fallecido; entonces era hombre.



Si ponemos como outcome que la persona sobreviva, las tres normas más potentes son la 2, 3 y 4. En las tres el sujeto es mujer de primera clase; las tres normas solo varían en que una considera que tenga familia, la otra que no la tenga, y la última no considera la variable Fam. Por tanto, vemos que la asociación más fuerte; es decir, que el perfil con más probabilidades de sobrevivir ha sido la mujer de primera clase, independientemente de si esta tenía familia o no u otros factores. En los siguientes gráficos podemos ver cómo hay una gran fuerza de asociación en esta variable:





Fijándonos en los valores de las reglas de asociación, vemos que el hecho de que una mujer de clase 1 sobreviviese es un fenómeno que corresponde al 10% de todos los pasajeros. A las tres reglas se les otorga una confianza muy elevada, de casi un 97% ny un lift de 2.5. Podemos concluir que es una asociación robusta.

VI. Resolución del problema

Nuestro análisis nos ha permitido ver de qué manera la edad, el género, la clase, el precio del billete y la tenencia o no de familia han afectado a la supervivencia de las personas a bordo del Titanic. Hemos podido ver que existe una tendencia a que si el precio del billete era alto, la persona sobreviviese. Aunque tanto la correlación como el estimador de la regresión eran bajos, al continuar con nuestro análisis y desarrollar las normas de asociación vemos que esta tendecia precio-supervivencia existe pero se manifiesta solo en mujeres. Si la mujer era de clase alta (y, por tanto, pagaba más por el billete), esta tenía una alta probabilidad de sobervivir. Tras realizar una regresión logarítmica en la que comprobamos cómo el género, la edad, la tenencia de familia y la clase afectaban a la supervivencia, vimos que el factor más determinante era el género, seguido de la clase. El hecho de que la persona tuviese o no familia a bordo no resultó estadísticamente significativo. Estos resultaron encajan con lo comprobado en la última parte del análisis con las reglas de asociación: la relación más fuerte es que las mujeres de clase alta sobreviven, tanto si tienen como si no tienen familia (es decir, el hecho de que tengan o no familia no afecta al hecho de que estadísticamente tuvieron más probabilidades de sobrevivir). Esta era la única tendencia fuerte que encontramos en nuestro análisis. Por supuesto, las causas de que esto fuese así deberían responder a un análisis más amplio de carácter sociológico: la premisa de la caballería, el clasismo que hace que se priorice a alguien de clase alta frente a alguien de clase baja… pero, de cualquier manera, descubrir que existe esta tendencia tan firme es una premisa interesante sobre la que se podría seguir investigando.

VII. Código del análisis

El código que hemos desarrollado para realizar este análisis se encuentra disponible online en un repositorio de Github accesible a través del siguiente link:

https://github.com/sgarciarodriguez6/PRACTICA-2

Se encuentra distribuido en dos archivos distintos:

- Código-Práctica 2 - Con el archivo del código en brutp
- Con el informe de Rmarkdown, que incluye código y resoluciones