

A3 - Modelado Predictivo

Pablo A. Delgado

14 de May, 2021

Contents

Introduccion	2
1. Modelo de regresión lineal	2
1.1. Modelo de regresión lineal (regresores cuantitativos)	6
1.1.a Regresion Lineal Simple	6
1.1.b Regresion Lineal Simple Multiple	8
1.1.c Regresion Lineal Multiple (2 submuestas)	9
1.2. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos)	10
1.3. Diagnosis del modelo	11
1.4. Predicción del modelo	14
2. Modelo de regresión logística.	14
2.1. Estudio de relaciones entre variables.	14
2.1.a Analisis con dos variables independientes	15
2.1.b Asociacion entre variables	19
2.2. Creacion de Modelos de regresión logística.	20
2.2.a. Con 1 una variable independiente: DAY_OF_WEEK	20
2.2.b. Con 1 una variable independiente: AIRLINE.	22
2.2.c. Con 2 variables independientes: DAY_OF_WEEK y DISTANCE	23
2.2.d. Seleccion de las variables mas significativas	24
2.3. Predicción	26
2.4. Bondad del ajuste	26
2.5. Curva ROC	27
3. Conclusiones del análisis	28
Referencias	29

Introducción

En esta actividad usaremos un conjunto de datos sobre el aeropuerto internacional de San Francisco (dat_SFO). Ha sido galardonado dos veces, como el mejor aeropuerto en América del Norte. En este estudio se analizarán los datos de vuelos recogidos durante el año 2015. El archivo contiene aproximadamente 145000 registros y 28 variables

Las principales variables son:

- Month: Día del mes de salida del vuelo.
- Day of week: Día de la semana de salida del vuelo.
- Airline: Nombre en siglas de la compañía aérea.
- Destination Airport: Aeropuerto de destino.
- Scheduled Departure: Hora de salida del vuelo estimada por la compañía.
- Departure Time: Hora de salida real del vuelo.
- Departure Delay: Diferencia entre la hora de salida estimada y la real.
- Air Time: Tiempo real de vuelo en aire,
- Distance: Distancia entre los aeropuertos origen y llegada.
- Scheduled Arrival: Hora de llegada del vuelo estimada por la compañía.
- Arrival Time: Hora de llegada del vuelo
- Arrival Delay: Diferencia entre la hora de llegada estimada y la real.
- Late Aircraft Delay: Retraso por llegada tarde del avión.
- Diverted: Indicador de vuelo desviado, siendo cero si el vuelo se ha efectuado con normalidad y uno si ha sido desviado.
- Cancelled: Indicador de vuelo cancelado, siendo cero si el vuelo se ha efectuado y uno si no.

Cada año una cantidad considerable de vuelos de diferentes aerolíneas se retrasa o cancela, costando al sistema de transporte aéreo miles de millones de euros en pérdidas de tiempo y dinero. En esta actividad se pretende realizar un estudio de los retrasos de los vuelos, tanto en salidas como llegadas. Para ello, se estudiarán las relaciones entre los mismos y varias variables. Primero se estudiarán las relaciones lineales y posteriormente se evaluarán los posibles factores de riesgo de estos retrasos.

A continuación, se especifican los pasos a seguir. En la entrega, se debe respetar la misma numeración de los apartados del índice.

1. Modelo de regresión lineal

Como primer paso definimos que librerías estaremos usando. Para ellos generamos un vector con todas las posibles librerías que necesitaremos, instalamos las que no tengamos, para finalmente mediante el uso de lapply cargarlas.

```
packages <- c("ggplot2", "gridExtra", "vcd", "ResourceSelection", "pROC")
new <- packages[!(packages %in% installed.packages()[, "Package"])]
if(length(new)) install.packages(new)
foo=lapply(packages, require, character.only=TRUE)
```

```
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.5

## Loading required package: gridExtra
```

```

## Warning: package 'gridExtra' was built under R version 4.0.5

## Loading required package: vcd

## Warning: package 'vcd' was built under R version 4.0.5

## Loading required package: grid

## Loading required package: ResourceSelection

## Warning: package 'ResourceSelection' was built under R version 4.0.5

## ResourceSelection 0.3-5 2019-07-22

## Loading required package: pROC

## Warning: package 'pROC' was built under R version 4.0.5

## Type 'citation("pROC")' for a citation.

## 
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
## 
##     cov, smooth, var

```

Dado que el input de datos es un archivo csv, y en una inspeccion visual manual hemos visto que posee como separador de columna la coma, usaremos la funcion read.csv() para cargar los datos en un dataframe. Hemos podido constatar que el csv cuenta con 28 columnas y 145954 lineas, la primera linea corresponde al header y la ultima una linea en blanco, siendo asi 145952 lineas con datos.

Dicho esto carguemos el archivo y hagamos una verificacion rapida de las primeras y ultimas 5 filas del dataframe para algunas de las variables.

```

vuelos = read.csv("SFO.csv", stringsAsFactors = FALSE, header=TRUE)
cols = c('YEAR','AIRLINE','FLIGHT_NUMBER','ORIGIN_AIRPORT','DESTINATION_AIRPORT',
        'SCHEDULED_TIME','ELAPSED_TIME','DISTANCE','ARRIVAL_DELAY',
        'DEPARTURE_DELAY','CANCELLED')
head(vuelos[,cols])

```

	YEAR	AIRLINE	FLIGHT_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_TIME
## 1	2015	F9	668	SFO	DEN	150
## 2	2015	00	6287	SFO	SBA	71
## 3	2015	AA	2207	SFO	DFW	208
## 4	2015	00	5647	SFO	SBA	75
## 5	2015	00	6508	SFO	PSP	92
## 6	2015	00	6289	SFO	MMH	63
	ELAPSED_TIME	DISTANCE	ARRIVAL_DELAY	DEPARTURE_DELAY	CANCELLED	
## 1	139	967	-45	-34	0	
## 2	99	262	-4	-32	0	
## 3	247	1464	8	-31	0	
## 4	87	262	-14	-26	0	
## 5	110	421	-7	-25	0	
## 6	71	193	-14	-22	0	

```
tail(vuelos[,cols])
```

```
##          YEAR AIRLINE FLIGHT_NUMBER ORIGIN_AIRPORT DESTINATION_AIRPORT
## 145947 2015      AA           12          SFO          JFK
## 145948 2015      AA          193          SFO          DFW
## 145949 2015      AA           16          SFO          JFK
## 145950 2015      AA          1145         SFO          ORD
## 145951 2015      AA          2293         SFO          DFW
## 145952 2015      AA          1454         SFO          DFW
##          SCHEDULED_TIME ELAPSED_TIME DISTANCE ARRIVAL_DELAY DEPARTURE_DELAY
## 145947            347           319     2586        1098        1126
## 145948            204           202     1464        1143        1145
## 145949            353           344     2586        1167        1176
## 145950            275           253     1846        1187        1209
## 145951            201           207     1464        1371        1365
## 145952            212           214     1464        1498        1496
##          CANCELLED
## 145947          0
## 145948          0
## 145949          0
## 145950          0
## 145951          0
## 145952          0
```

Veamos la cantidad de columnas, filas y como quedaron los tipos de datos en el dataframe para entender si hubo algun dato importado incorrectamente.

```
str(vuelos)
```

```
## 'data.frame': 145952 obs. of 28 variables:
## $ YEAR : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
## $ MONTH : int 12 1 11 5 5 3 4 4 4 6 ...
## $ DAY : int 31 8 27 30 23 29 11 18 25 2 ...
## $ DAY_OF_WEEK : int 4 4 5 6 6 7 6 6 6 2 ...
## $ AIRLINE : chr "F9" "OO" "AA" "OO" ...
## $ FLIGHT_NUMBER : int 668 6287 2207 5647 6508 6289 304 304 317 5459 ...
## $ ORIGIN_AIRPORT : chr "SFO" "SFO" "SFO" "SFO" ...
## $ DESTINATION_AIRPORT: chr "DEN" "SBA" "DFW" "SBA" ...
## $ SCHEDULED_DEPARTURE: int 2000 1350 1200 1840 2230 831 2135 2135 2145 900 ...
## $ DEPARTURE_TIME : int 1926 1318 1129 1814 2205 809 2113 2113 2123 838 ...
## $ DEPARTURE_DELAY : int -34 -32 -31 -26 -25 -22 -22 -22 -22 -22 ...
## $ TAXI_OUT : int 14 52 43 36 39 25 23 15 20 35 ...
## $ WHEELS_OFF : int 1940 1410 1212 1850 2244 834 2136 2128 2143 913 ...
## $ SCHEDULED_TIME : int 150 71 208 75 92 63 84 84 124 66 ...
## $ ELAPSED_TIME : int 139 99 247 87 110 71 95 84 136 75 ...
## $ AIR_TIME : int 109 44 177 47 64 41 67 65 111 36 ...
## $ DISTANCE : int 967 262 1464 262 421 193 421 421 679 190 ...
## $ WHEELS_ON : int 2229 1454 1709 1937 2348 915 2243 2233 2334 949 ...
## $ TAXI_IN : int 16 3 27 4 7 5 5 4 5 4 ...
## $ SCHEDULED_ARRIVAL : int 2330 1501 1728 1955 2 934 2259 2259 2349 1006 ...
## $ ARRIVAL_TIME : int 2245 1457 1736 1941 2355 920 2248 2237 2339 953 ...
## $ ARRIVAL_DELAY : int -45 -4 8 -14 -7 -14 -11 -22 -10 -13 ...
```

```

## $ DIVERTED      : int 0 0 0 0 0 0 0 0 0 ...
## $ CANCELLED     : int 0 0 0 0 0 0 0 0 0 ...
## $ CANCELLATION_REASON: chr "" "" "" ...
## $ AIR_SYSTEM_DELAY : int 0 0 0 0 0 0 0 0 0 ...
## $ LATE_AIRCRAFT_DELAY: int NA 0 NA NA NA NA NA NA ...
## $ WEATHER_DELAY    : int NA NA NA NA NA NA NA NA ...

```

```
nrow(vuelos)
```

```
## [1] 145952
```

```
ncol(vuelos)
```

```
## [1] 28
```

En principio como vemos, 145952 obs. of 28 variables, la cantidad de filas y columnas en el dataframe coinciden con la inspeccion visual que hemos hecho sobre el archivo csv. Y los tipos de datos de cada variable corresponden a los esperados.

Lo que si es visible que tenemos cancellation reasons sin asignar, que serian los casos en que no hubo cancelacion. Valores perdidos a los cuales se le podra imputar por ej un “No Aplica” si se trata de los vuelos no cancelados. Ademas vemos varios NA en las dos ultimas variables. En ambos casos trataremos segun corresponda al momento de realizar el analisis y prediccion de los datos en los siguientes pasos de esta practica.

```
colSums(is.na(vuelos))
```

	YEAR	MONTH	DAY	DAY_OF_WEEK
##	0	0	0	0
##	AIRLINE	FLIGHT_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT
##	0	0	0	0
##	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_OUT
##	0	0	0	40
##	WHEELS_OFF	SCHEDULED_TIME	ELAPSED_TIME	AIR_TIME
##	40	0	461	461
##	DISTANCE	WHEELS_ON	TAXI_IN	SCHEDULED_ARRIVAL
##	0	173	173	0
##	ARRIVAL_TIME	ARRIVAL_DELAY	DIVERTED	CANCELLED
##	173	461	0	0
##	CANCELLATION_REASON	AIR_SYSTEM_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY
##	0	0	93699	116418

```
colSums(vuelos=="")
```

	YEAR	MONTH	DAY	DAY_OF_WEEK
##	0	0	0	0
##	AIRLINE	FLIGHT_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT
##	0	0	0	0
##	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_OUT
##	0	0	0	NA
##	WHEELS_OFF	SCHEDULED_TIME	ELAPSED_TIME	AIR_TIME
##	NA	0	NA	NA

```

##          DISTANCE      WHEELS_ON      TAXI_IN  SCHEDULED_ARRIVAL
##            0             NA           NA           0
## ARRIVAL_TIME    ARRIVAL_DELAY    DIVERTED    CANCELLED
##            NA             NA           0           0
## CANCELLATION_REASON   AIR_SYSTEM_DELAY LATE_AIRCRAFT_DELAY WEATHER_DELAY
##            145860                  0           NA           NA

```

1.1. Modelo de regresión lineal (regresores cuantitativos)

1.1.a Regresion Lineal Simple

Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable DEPARTURE_DELAY en función de la variable ARRIVAL_DELAY. Se evaluará la bondad del ajuste, a partir del coeficiente de determinación. Calcular el coeficiente de correlación y explicar su relación con el coeficiente de determinación.

```
m11a = lm(DEPARTURE_DELAY~ARRIVAL_DELAY,data=vuelos)
summary(m11a)
```

```

##
## Call:
## lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY, data = vuelos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -165.056   -6.299    0.144    6.576  109.598
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.8560098  0.0309626 189.1 <2e-16 ***
## ARRIVAL_DELAY 0.9215929  0.0007684 1199.4 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.69 on 145489 degrees of freedom
## (461 observations deleted due to missingness)
## Multiple R-squared:  0.9082, Adjusted R-squared:  0.9082
## F-statistic: 1.439e+06 on 1 and 145489 DF,  p-value: < 2.2e-16

```

Siendo el coeficiente de determinación (R^2 o R-squared) una medida de calidad del modelo que toma valores entre 0 y 1, se comprueba cómo el DEPARTURE_DELAY y el ARRIVAL_DELAY se correlacionan fuertemente, dando lugar a un R-squared de 0.9082.

Pero que pasa si imputamos un valor al arrival_delay? Para eso creamos una nueva variable sin los NA, aplicando un calculo simple de imputacion: asignar la media. Obviamente se podria usar regresion para imputar. Pero a modo de ejemplificar y validar rapidamente usamos este metodo.

```
vuelos$ARRIVAL_DELAY_NO_NA = vuelos$ARRIVAL_DELAY

vuelos$ARRIVAL_DELAY_NO_NA[is.na(vuelos$ARRIVAL_DELAY_NO_NA)] = mean(vuelos$ARRIVAL_DELAY_NO_NA,na.rm=T)
```

Volvamos a validar:

```
m11a_bis = lm(DEPARTURE_DELAY~ARRIVAL_DELAY_NO_NA,data=vuelos)
summary(m11a_bis)
```

```
##
## Call:
## lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY_NO_NA, data = vuelos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -165.09    -6.36     0.11     6.54   379.81
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.8913116  0.0319084 184.6   <2e-16 ***
## ARRIVAL_DELAY_NO_NA  0.9215929  0.0007931 1162.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.07 on 145950 degrees of freedom
## Multiple R-squared:  0.9025, Adjusted R-squared:  0.9025
## F-statistic: 1.35e+06 on 1 and 145950 DF, p-value: < 2.2e-16
```

Vemos que practicamente no hubo cambios en el coeficiente, solo un delta negativo de -0.0057 entre ambos coeficientes, osea empeora minimamente la predicción porque en realidad en el caso anterior directamente LM descarto los valores missing y por ende no fueron tenidas en cuenta. Mientras que al imputarle solo la media, y estabamos hablando de vuelos comerciales con todo lo que eso implica, aplicar la media de los delay en la demora no parece ser una buena medida de imputacion de valores missings, al menos para la muestra con la que contamos.

Ahora volviendo a los resultados sin descartar missings, dijimos que el coeficiente de determinacion R^2 es 0.9082, pero que implica este valor?

La función de `lm()` de R nos calcula la recta de regresión por mínimos cuadrados, la cual minimiza la suma de los cuadrados de los residuos. Pero como sabemos si ese ajuste es lo suficiente bueno? Sabemos que visualmente podemos visualizarlo con gráficos de dispersión, pero si hablamos de un valor numérico que nos ayude precisarlo es aquí cuando entra en juego el coeficiente de determinación, el cual es la medida más importante de la bondad del ajuste. La fórmula es:

$$R^2 = \frac{\text{Varianza explicada por la recta de regresión}}{\text{Varianza total de los datos}}$$

Por lo tanto como R^2 nos explica la proporción de variabilidad de los datos que queda explicada por el modelo de regresión, cuanto más cercano a la unidad esté, mejor es el ajuste.

Como hemos dicho antes, con el diagrama de dispersión podemos ver si hay algún tipo de relación entre dos variables X e Y . Pero además con el coeficiente de correlación podemos representar en números esta relación. Dicho eso calculemos el r de nuestras variables :

```
cor(x=vuelos$DEPARTURE_DELAY, y=vuelos$ARRIVAL_DELAY, use = "pairwise.complete.obs")
```

```
## [1] 0.9529701
```

Y siendo que el coeficiente de correlación se caracteriza por estar entre $-1 \leq r \leq 1$, de forma que:

- $r = -1$ o $r = 1$ cuando haya una asociación lineal exacta entre las variables (en el primer caso positiva y en el segundo, negativa).

- $-1 < r < 1$ cuando la relación entre las variables no sea lineal de forma exacta.

- Para los otros valores siempre se formula la misma pregunta: ¿a partir de qué valor de r podemos decir que la relación entre las variables es fuerte? Una regla razonable es decir que la relación es débil si $0 < |r| < 0,5$, fuerte si $0,8 < |r| < 1$, y moderada si tiene otro valor.

Nuestro valor de r obtenido está representado por las dos últimas situaciones ya que primero la relación entre nuestras dos variables en análisis no tienen una relación lineal exacta y segundo que tienen una correlación fuerte dado que:

$$0,8 < |r = 0.9529701| < 1$$

Recordar, que si el r obtenido hubiese estado cerca de 0, implicaría que no existe correlación entre las variables.

Y esto cumple con la regla $R^2 = r^2$ en regresiones simples ya que:

$R^2 = 0.9082$ y si calculamos el cuadrado de r tenemos:

```
r=0.9529701
r^2
```

```
## [1] 0.908152
```

o sea $r^2 = 0.908152 = R^2$

Recordar entonces que:

- R-Squared: mide la proporción de variación de la variable dependiente explicada por la variable independiente.
- r : mide el grado de asociación entre las dos variables.
- Y en la regresión lineal simple siempre tendremos que $R^2 = r^2$, como hemos podido comprobar.

También es importante tener presente que r nos da más información que R^2 . El signo de r nos informa de si la relación es positiva o negativa. Así pues, con el valor de r siempre podremos calcular el valor de R-Squared, pero al revés siempre nos quedará indeterminado el valor del signo a menos que conozcamos la pendiente de la recta. Por ejemplo, dado nuestro R-squared de 0.9082, si sabemos que la pendiente de la recta de regresión es positiva, entonces podremos afirmar que el coeficiente de correlación (realizando $\sqrt{R^2}$) será $r = 0.9529$ (hubiese sido -0.9529 si la pendiente era negativa).

1.1.b Regresión Lineal Simple Múltiple

Se añadirá al modelo anterior la variable independiente DISTANCIA. ¿Existe una mejora del ajuste?

```
m11b = lm(DEPARTURE_DELAY~ARRIVAL_DELAY+DISTANCE,data=vuelos)
summary(m11b)
```

```
##
## Call:
## lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY + DISTANCE, data = vuelos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0000 -1.0000  0.0000  1.0000 10.0000
```

```

## -166.337   -5.925     0.698     6.704   104.300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.350e+00 5.032e-02 46.70 <2e-16 ***
## ARRIVAL_DELAY 9.246e-01 7.499e-04 1232.96 <2e-16 ***
## DISTANCE     2.904e-03 3.335e-05  87.08 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.4 on 145488 degrees of freedom
##   (461 observations deleted due to missingness)
## Multiple R-squared:  0.9127, Adjusted R-squared:  0.9127
## F-statistic: 7.605e+05 on 2 and 145488 DF,  p-value: < 2.2e-16

```

Al introducir DISTANCE, el R-squared mejora hasta 0.9127 ya que también se correlaciona con esta nueva variable, aunque en menor medida que con ARRIVAL_DELAY como puede verse en el resultado del summary.

1.1.c Regresion Lineal Multiple (2 submuestras)

Ahora procederemos a dividir la muestra en dos, según los vuelos sean o no más largos. Se tomará por larga distancia aquéllos con un recorrido superior a 600 millas.

```

vuelos_short_distance=vuelos[vuelos$DISTANCE<=600,]
vuelos_long_distance=vuelos[vuelos$DISTANCE>600,]

m11c_short_distance = lm(DEPARTURE_DELAY~ARRIVAL_DELAY+DISTANCE,data=vuelos_short_distance)
summary(m11c_short_distance)

```

```

##
## Call:
## lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY + DISTANCE, data = vuelos_short_distance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -97.876  -4.732    0.852    5.734   56.659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.3786458  0.1252898 11.00 <2e-16 ***
## ARRIVAL_DELAY 0.9299034  0.0010159  915.32 <2e-16 ***
## DISTANCE     0.0057171  0.0003179   17.99 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.206 on 63195 degrees of freedom
##   (173 observations deleted due to missingness)
## Multiple R-squared:  0.9299, Adjusted R-squared:  0.9299
## F-statistic: 4.189e+05 on 2 and 63195 DF,  p-value: < 2.2e-16

```

```
m11c_long_distance = lm(DEPARTURE_DELAY~ARRIVAL_DELAY+DISTANCE, data=vuelos_long_distance)
summary(m11c_long_distance)
```

```
##
## Call:
## lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY + DISTANCE, data = vuelos_long_distance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -165.666   -7.060    0.515    7.764  108.767
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.858e+00 1.260e-01 14.74 <2e-16 ***
## ARRIVAL_DELAY 9.218e-01 1.051e-03 877.08 <2e-16 ***
## DISTANCE     3.127e-03 6.409e-05 48.80 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.82 on 82290 degrees of freedom
##   (288 observations deleted due to missingness)
## Multiple R-squared:  0.9034, Adjusted R-squared:  0.9034
## F-statistic: 3.848e+05 on 2 and 82290 DF, p-value: < 2.2e-16
```

Vemos que el coeficiente de determinacion mejora cuando solo trabajamos con el grupo de vuelos cortos, mientras que se decrementa minimamente para los vuelos largos. Aparentemente seria mas costoso predecir cual sera el delay de los vuelos largos que de los cortos. Lo cual se podria esperar, ya que podria afectarnos los factores climaticos por mas tiempo? esta hipotesis se podria validar si contaramos con informacion climatica mas precisa como tambien con cualquier otro dato que consultando con los expertos en el area nos puedan sugerir utilizar.

1.2. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos)

En este apartado se estudiará la relación de DEPARTURE_DELAY, con las variables explicativas ARRIVAL_DELAY y LATE_AIRCRAFT_DELAY. Para ello se procederá a la recodificación de la variable LATE_AIRCRAFT_DELAY, en mayor y menor o igual a 15 minutos.

```
vuelos$LATE_AIRCRAFT_FLAG[vuelos$LATE_AIRCRAFT_DELAY<15
                           | is.na(vuelos$LATE_AIRCRAFT_DELAY)] = "< 15"
vuelos$LATE_AIRCRAFT_FLAG[vuelos$LATE_AIRCRAFT_DELAY>=15] = ">= 15"

table(vuelos$LATE_AIRCRAFT_FLAG)
```

```
##
## < 15  >= 15
## 132068 13884
```

```
nrow(vuelos)
```

```
## [1] 145952
```

```
m12 = lm(DEPARTURE_DELAY~ARRIVAL_DELAY+LATE_AIRCRAFT_FLAG,data=vuelos)
summary(m12)
```

```
##
## Call:
## lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY + LATE_AIRCRAFT_FLAG,
##      data = vuelos)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -159.570  -6.281    0.056   6.376 152.447 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             5.3437945  0.0318818 167.61 <2e-16 ***
## ARRIVAL_DELAY          0.8933306  0.0009045  987.64 <2e-16 ***
## LATE_AIRCRAFT_FLAG>= 15 7.0705130  0.1227849   57.59 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.56 on 145488 degrees of freedom
## (461 observations deleted due to missingness)
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9102 
## F-statistic: 7.373e+05 on 2 and 145488 DF,  p-value: < 2.2e-16
```

En este ultimo caso (ARRIVAL_DELAY+LATE_AIRCRAFT_FLAG) hemos obtenido un R-Squared de 0.9102, menor al que obtuvimos con la combinacion ARRIVAL_DELAY+DISTANCE ya que el mismo era de 0.9127 (sin realizar el split de DISTANCE).

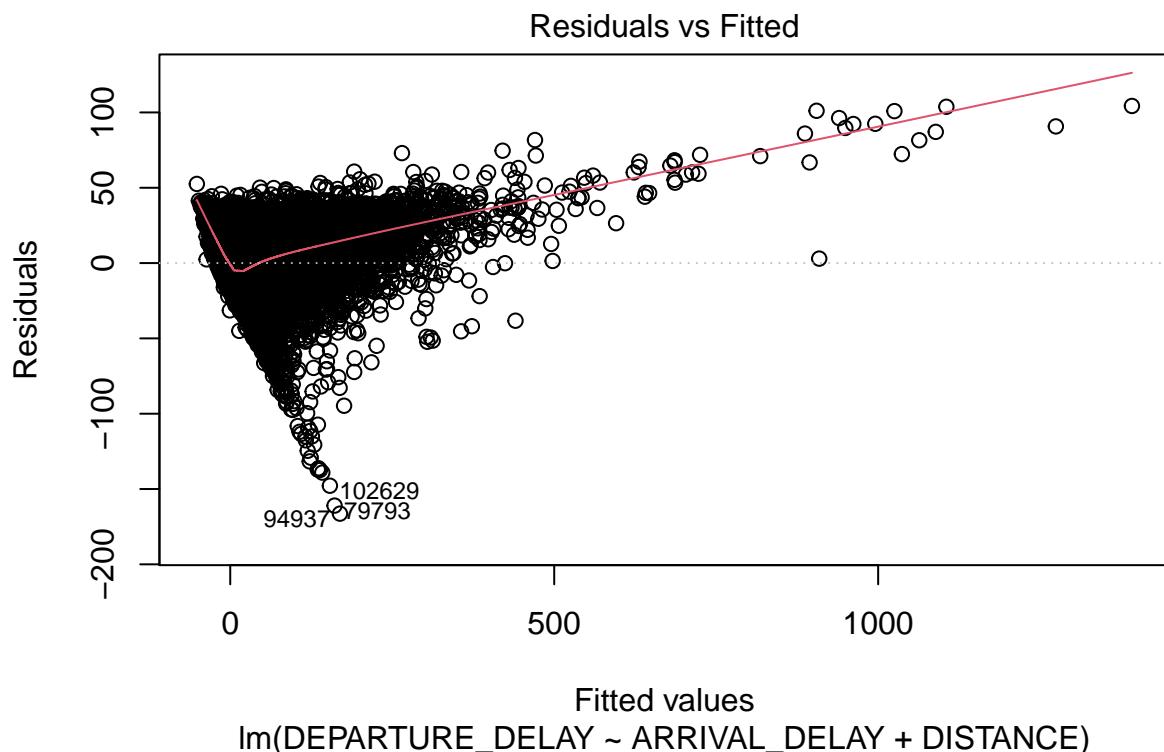
1.3. Diagnosis del modelo

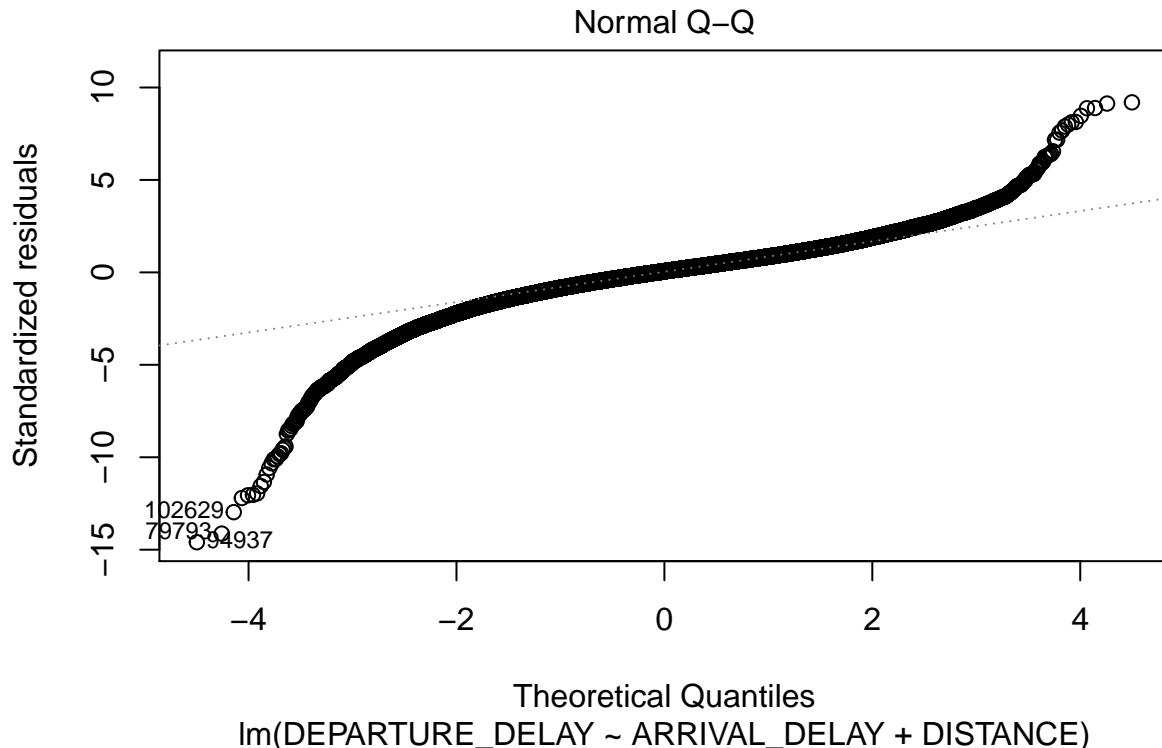
Para la diagnosis se escoge el modelo construído en el apartado 1.1.b y se pintarán dos gráficos:

- uno con los valores ajustados frente a los residuos (que nos permitirá ver si la varianza es constante) y
- el gráfico cuantil-cuantil que compara los residuos del modelo con los valores de una variable que se distribuye normalmente (QQ plot).

Para ello usamos el metodo plot pasandole por parametro los dos graficos que necesitamos analizar:

```
plot(m11b, which = c(1,2), caption = list("Residuals vs Fitted", "Normal Q-Q"))
```





Si bien R² es grande y el ajuste es bueno, tenemos igualmente una distorsión en los residuos.

Vemos el gráfico “Residuals vs Fitted” una posible tendencia en los datos, si bien para valores menores a 500 los puntos paracen concentrarse...y hubiesemos dicho que estan todos los puntos o residuos dispersos sin una forma específica, pero sin embargo a medida que los valores aumentan, tambien comienzan a aumentar los valores de los residuos con una clara tendencia lineal. Aquí puede estar sucediendo que el modelo es bueno para valores menores a 500 y deja de ser útil para valores estimados mayor a 500.

Mientras que en el QQ Plot, vemos que en los extremos inferior y superior nos alejamos de una distribución normal. Como se alejan podemos decir que los residuos no siguen una distribución normal. Una de las asunciones para la validez del modelo lineal es que los residuos presenten homoscedasticidad (valores de varianza similares) y normalidad.

Que los residuos de un modelo de regresión lineal se distribuyan de forma normal es una condición necesaria para que la significancia (p-value) y los intervalos de confianza asociados a los predictores (calculados a partir de modelos teóricos) sean precisos.

Pero el QQplot que vemos nos lleva a pensar que no se cumple la normalidad de los residuos. Por tanto, deberíamos concluir que los residuos no siguen una distribución normal. Esto invalidaría la asunción del modelo y nos podría hacer pensar en buscar modelos alternativos, quizás otro tipo de regresiones (no lineales). O sea, dado que nos encontramos con un patrón que se aleja de la normalidad se tendría que estudiar si existe relaciones no lineales que se podrían modelar.

En resumen si queremos determinar si encontramos modelos mejor ajustados o validados, tendríamos que profundizar más en las variables para entender si hay relaciones no lineales, osea exponenciales o logarítmicas, para ajustar más el modelo.

1.4. Predicción del modelo

Según el modelo del apartado b), calcular el retraso en la salida de un avión, que después de recorrer 2500 millas ha llegado a su destino con 30 minutos más tarde.

```
# Usamos un valor negativo de arrival delay por definicion de la variable.  
# Arrival Delay: Diferencia entre la hora de llegada estimada y la real,  
# y siendo que el avion llega 30 minutos mas tarde que la hora estimada =>  
# Arrival Delay = Hora Estimada - (Hora Estimada + 30)  
predict.df<-data.frame(DISTANCE=2500,ARRIVAL_DELAY=-30)  
predict(m11b,newdata=predict.df)
```

```
##      1  
## -18.128
```

La predicción a partir del modelo definido nos dice que la hora de salida del avión será aproximadamente 18 minutos después de la hora establecida.

2. Modelo de regresión logística.

2.1. Estudio de relaciones entre variables.

Se quiere estudiar la probabilidad que tiene un avión de sufrir un retraso.

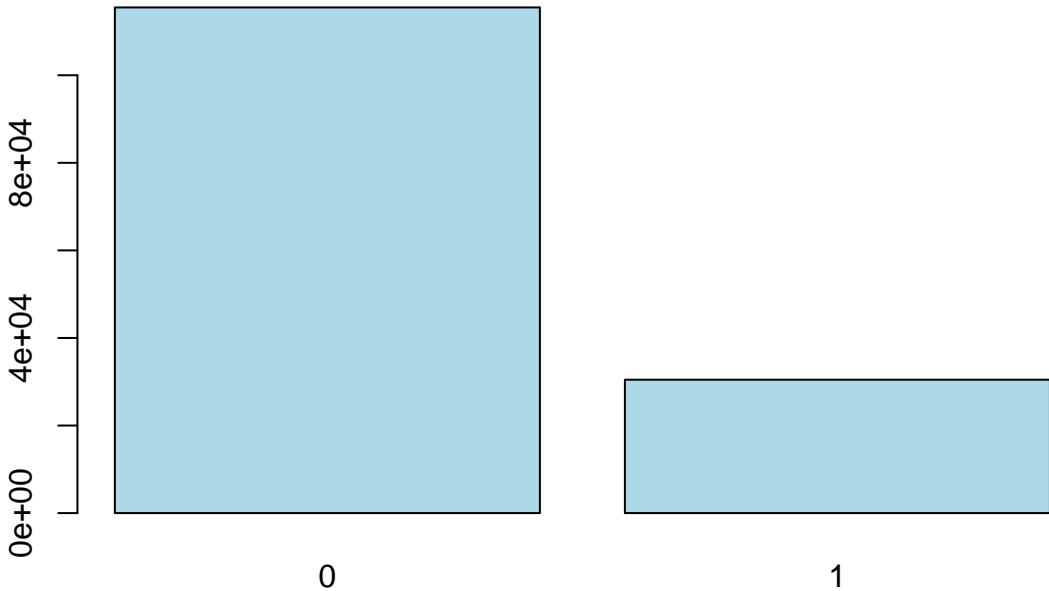
Para ello, primero se creará una nueva variable dicotómica llamada `delay_SFO`. Esta nueva variable está relacionada con los valores de la variable `Departure_Delay`. Se codificará de la siguiente: Si el valor de dicha variable es menor a 15 minutos, se puede asumir que el vuelo no va con retraso y se codificará con el valor 0, en caso contrario, se codificará con el valor 1.

Primero crearemos la variable:

```
vuelos[vuelos$DEPARTURE_DELAY < 15, 'DELAY_SFO'] = 0  
vuelos[vuelos$DEPARTURE_DELAY >= 15, 'DELAY_SFO'] = 1  
vuelos$DELAY_SFO = factor(vuelos$DELAY_SFO)
```

Y demosles un rápido vistazo de como se distribuyen estos dos nuevos valores:

```
barplot(table(vuelos$DELAY_SFO), col = c("lightblue"))
```



2.1.a Análisis con dos variables independientes

Visualizar la relación entre delay_SFO y las variables independientes:DAY_OF_WEEK y AIRLINE. Calcular las frecuencias relativas. Interpretar el significado. Visualizar con barplot.

Para resolver esto representemos las tablas de contingencias.

Veamos primero las Tablas de frecuencias absolutas para ambas variables:

```
AIRLINE_TABLE = table(vuelos$DELAY_SFO,vuelos$AIRLINE )
DAY_OF_WEEK_TABLE = table(vuelos$DELAY_SFO,vuelos$DAY_OF_WEEK )
AIRLINE_TABLE
```

```
##
##          AA      AS      B6      DL      F9      HA      OO      UA      US      VX      WN
## 0  10112  4258  3769  8063  1490   602 26824 34304  2324 13098 10644
## 1   1950    887  1121  1596    452     63  7399 10831    269  2750  3146
```

```
DAY_OF_WEEK_TABLE
```

```
##
##          1      2      3      4      5      6      7
## 0 16431 16911 17669 16747 16734 15068 15928
## 1  5311  4343  3975  4956  4681  2493  4705
```

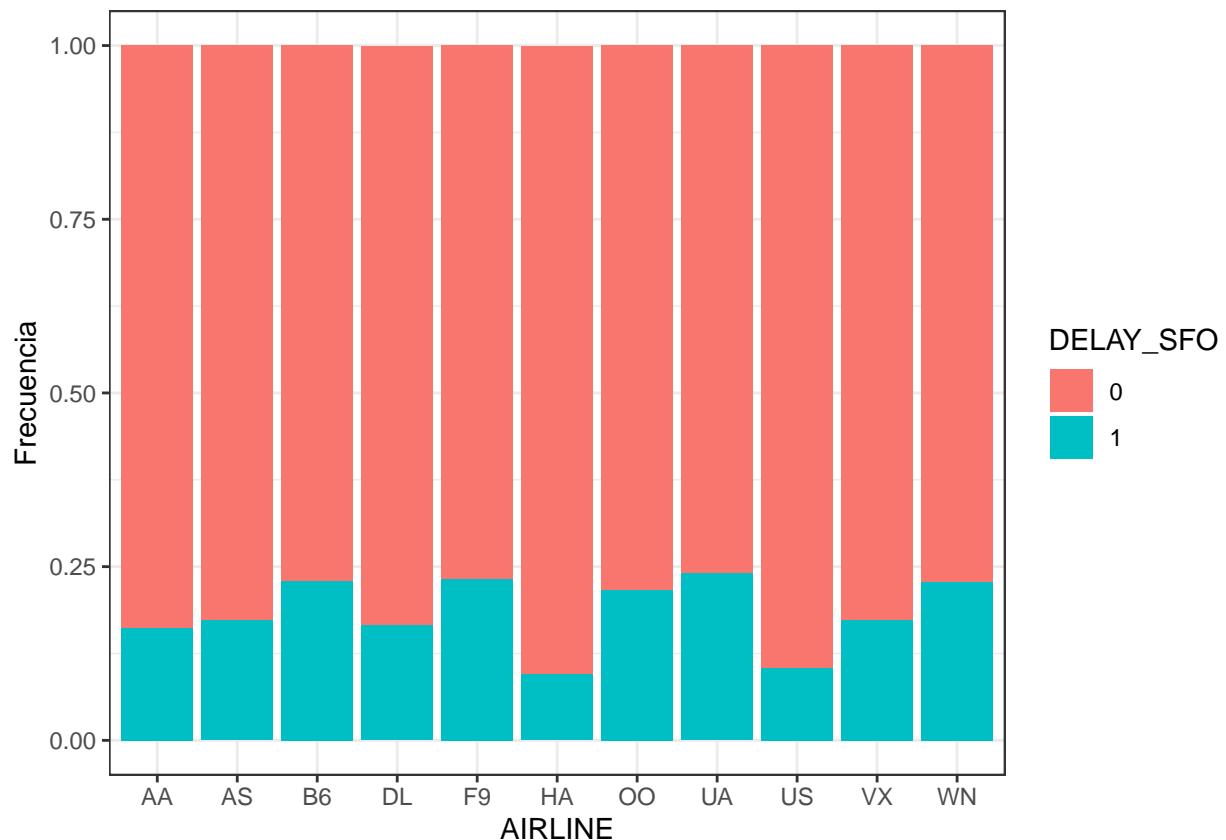
Mientras que ahora representemos las Tablas de frecuencias relativas y visualicemos graficamente esos valores:

```
AIRLINE_TABLE_RELATIVE = prop.table(AIRLINE_TABLE, margin=2)
DAY_OF_WEEK_TABLE_RELATIVE = prop.table(DAY_OF_WEEK_TABLE, margin=2)

AIRLINE_TABLE_RELATIVE
```

```
##
##          AA         AS         B6         DL         F9         HA
##  0  0.83833527 0.82759961 0.77075665 0.83476550 0.76725026 0.90526316
##  1  0.16166473 0.17240039 0.22924335 0.16523450 0.23274974 0.09473684
##
##          OO         UA         US         VX         WN
##  0  0.78380037 0.76003102 0.89625916 0.82647653 0.77186367
##  1  0.21619963 0.23996898 0.10374084 0.17352347 0.22813633
```

```
ggplot(vuelos,aes(x=AIRLINE,fill=DELAY_SFO)) +
  geom_bar(position="fill") +
  theme_bw() +
  ylab("Frecuencia")
```



```
DAY_OF_WEEK_TABLE_RELATIVE
```

```
##
```

```

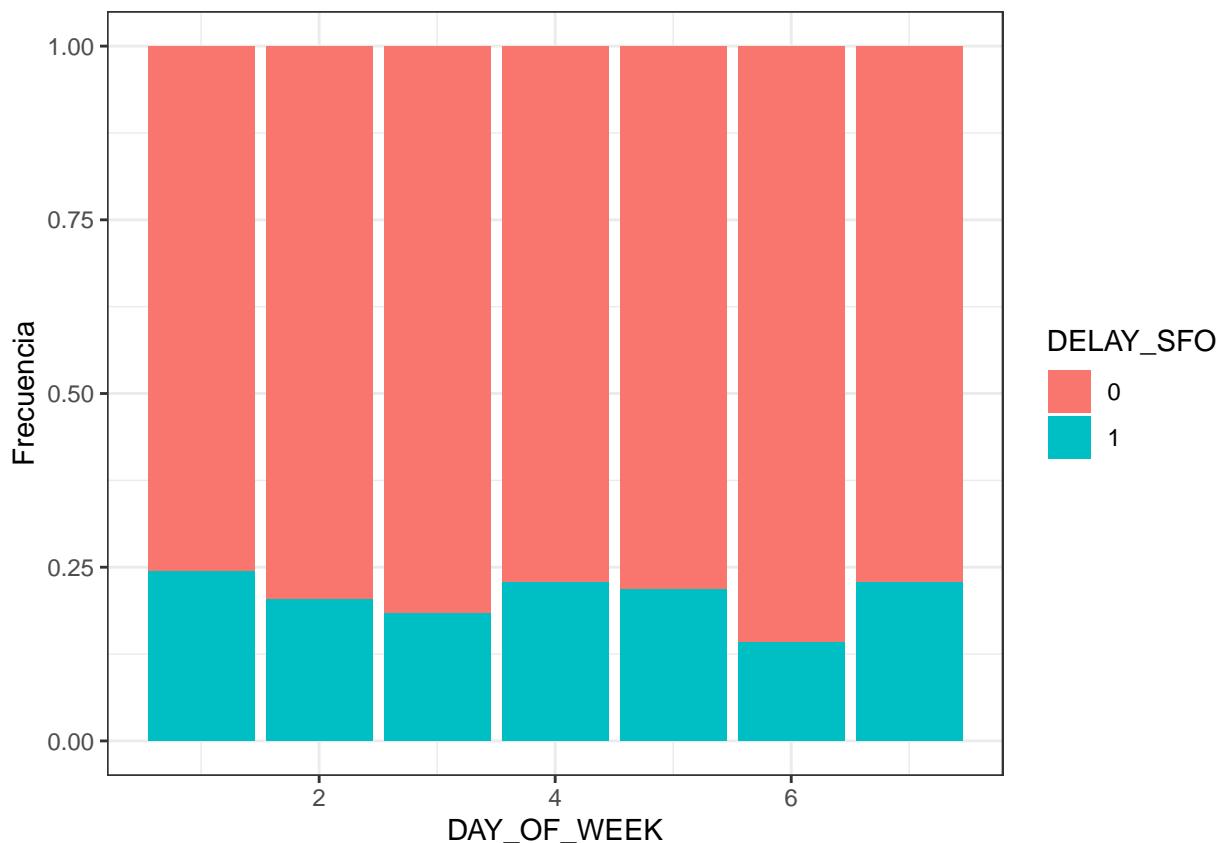
##          1         2         3         4         5         6         7
## 0 0.7557262 0.7956620 0.8163463 0.7716445 0.7814149 0.8580377 0.7719672
## 1 0.2442738 0.2043380 0.1836537 0.2283555 0.2185851 0.1419623 0.2280328

```

```

ggplot(vuelos,aes(x=DAY_OF_WEEK,fill=DELAY_SFO)) +
  geom_bar(position="fill") +
  theme_bw() +
  ylab("Frecuencia")

```



Por lo que podemos ver, las 4 peores aerolineas en cuanto a retrasos son B6, F9, UA y WN. Mientras que HA y US son las que menos retrasan tienen.

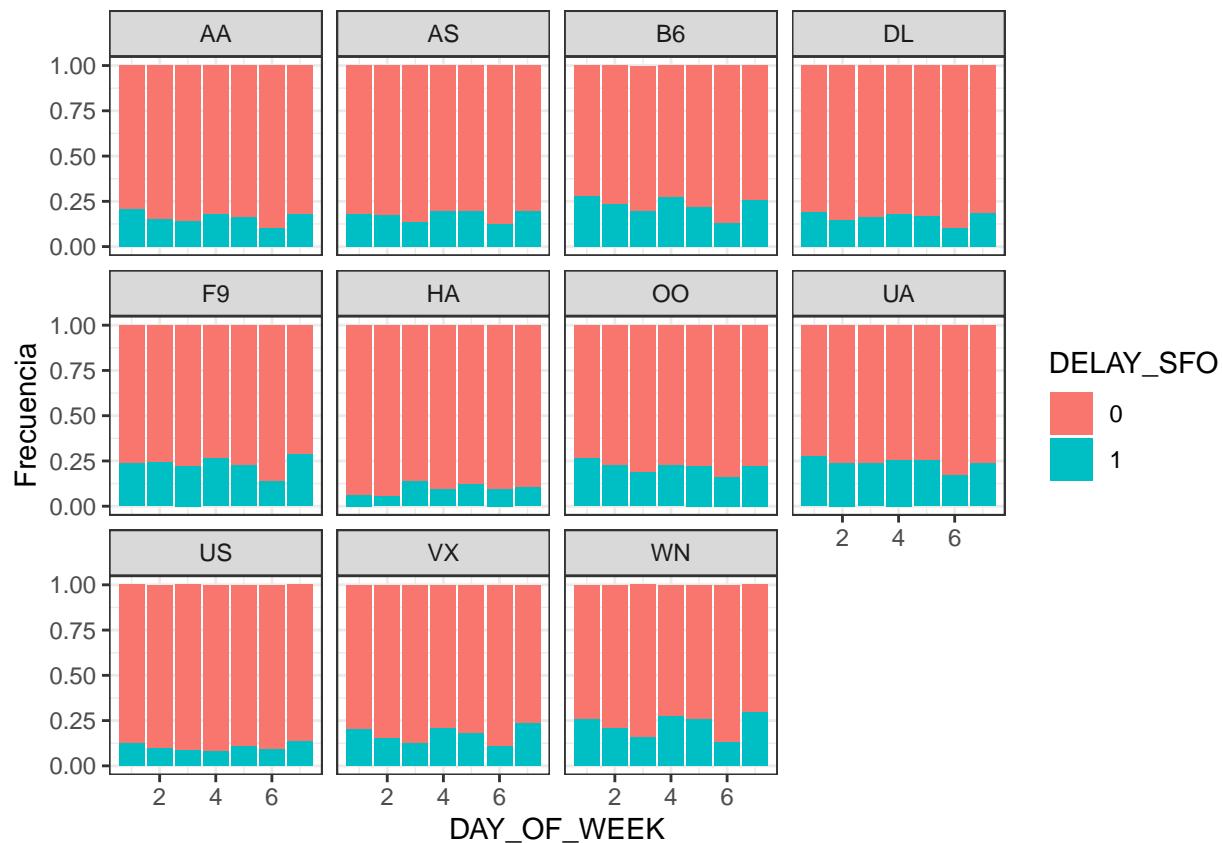
En cuanto a los dias no hay una diferencia significativa. Aunque se ve una ligera mejora los dias 6 de la semana con una menor cantidad de retrasos. Tal vez una opcion seria agrupar entre dias de la semana y fines de la semana para entender si podemos sacar mejores conclusiones.

Pero que pasa ahora si combinamos los valores de Dia y Semana a la vez? Veamoslo:

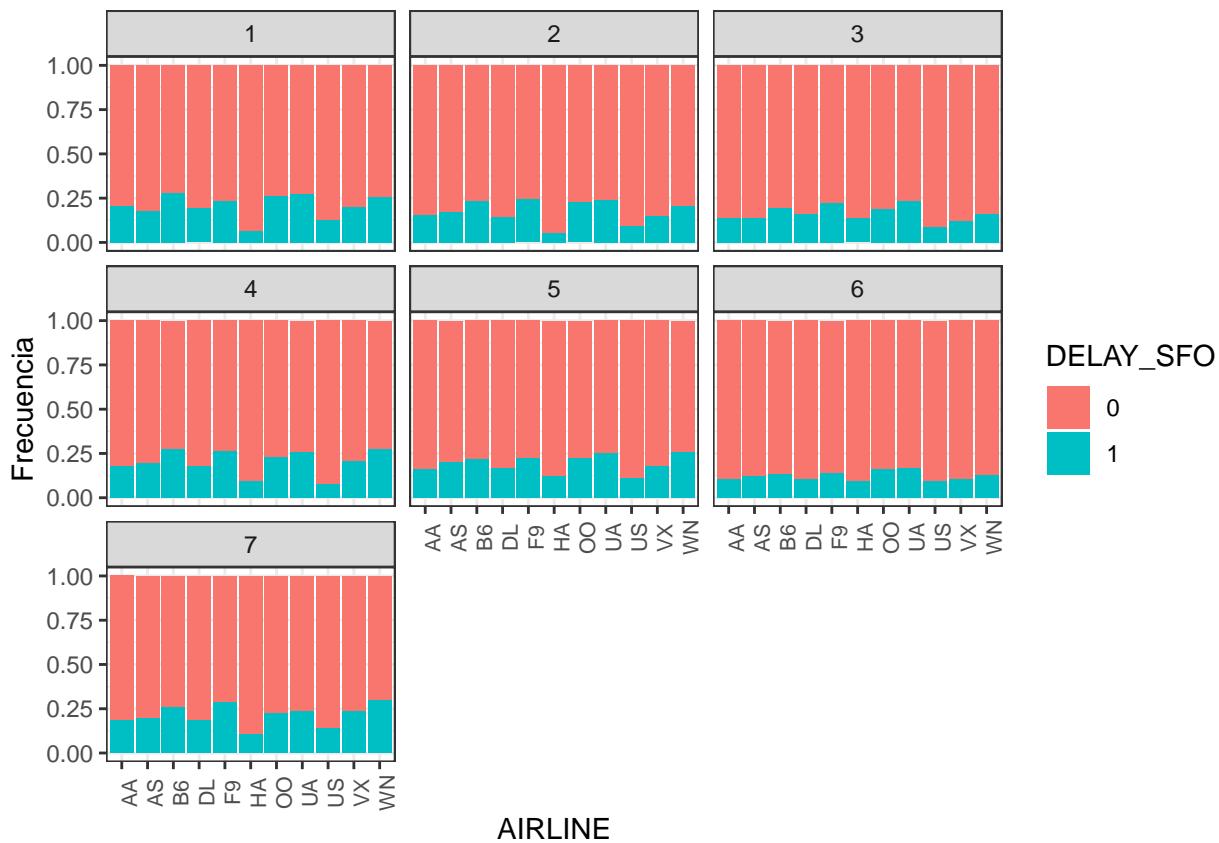
```

ggplot(vuelos,aes(x=DAY_OF_WEEK,fill=DELAY_SFO)) +
  geom_bar(position="fill") +
  theme_bw() +
  ylab("Frecuencia") +
  facet_wrap(~AIRLINE)

```



```
ggplot(vuelos,aes(x=AIRLINE,fill=DELAY_SFO)) +
  geom_bar(position="fill") +
  theme_bw() +
  ylab("Frecuencia") +
  facet_wrap(~DAY_OF_WEEK) +
  theme(axis.text.x=element_text(size=8,angle=90))
```



Con estas graficas no podemos sacar muchas mas conclusiones. El unico insigh rapido que obtenemos de aqui es que si bien vimos antes que la aerolinea HA tiene baja cantidad de retrasos a nivel total respecto a las demas aerolineas, esa gran diferencia se da solo en los dias 1 y 2. En los demas dias los retrasos son similares a las demas.

2.1.b Asociacion entre variables

Para comprobar si existe asociación entre las variable dependiente y cada una de las variables explicativas, se aplicará el test Chi-cuadrado de Pearson. Un resultado significativo nos dirá que existe asociación.

Con R podemos aplicar este test de la siguiente forma:

```
chi_1 <- chisq.test(AIRLINE_TABLE)
chi_1
```

```
##
##  Pearson's Chi-squared test
##
## data: AIRLINE_TABLE
## X-squared = 986.78, df = 10, p-value < 2.2e-16
```

```
chi_2 <- chisq.test(DAY_OF_WEEK_TABLE)
chi_2
```

```
##
```

```

## Pearson's Chi-squared test
##
## data: DAY_OF_WEEK_TABLE
## X-squared = 834.95, df = 6, p-value < 2.2e-16

```

Como vemos el p-value es significativo por lo que podemos asumir asociacion, ya que se rechaza la hipotesis nula de independencia entre variables.

O podemos usar esta otra funcion del paquete vcd que nos da varios coeficientes a la vez entre ellos el de Pearson.

```
assocstats(AIRLINE_TABLE)
```

```

##          X^2 df P(> X^2)
## Likelihood Ratio 1040.55 10      0
## Pearson         986.78 10      0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.082
## Cramer's V        : 0.082

```

```
assocstats(DAY_OF_WEEK_TABLE)
```

```

##          X^2 df P(> X^2)
## Likelihood Ratio 874.22  6      0
## Pearson         834.95  6      0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.075
## Cramer's V        : 0.076

```

2.2. Creacion de Modelos de regresión logística.

2.2.a. Con 1 una variable independiente: DAY_OF_WEEK

Estimar el modelo de regresión logística tomando como variable dependiente delay_SFO y variable explicativa DAY_OF_WEEK. Se tomará como día de referencia el lunes. Se puede considerar que el día de la semana es un factor de riesgo? Justifica tu respuesta.

Primero definimos como referencia el dia 1 y luego generamos el modelo

```

DAY_OF_WEEK_REF_=relevel(factor(vuelos$DAY_OF_WEEK), ref = '1')

model_22a=glm(formula=DELAY_SFO~DAY_OF_WEEK_REF_,family=binomial(link=logit), data=vuelos)

summary(model_22a)

```

```

##
## Call:
## glm(formula = DELAY_SFO ~ DAY_OF_WEEK_REF_, family = binomial(link = logit),
##      data = vuelos)
## 
```

```

## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7484 -0.7195 -0.6761 -0.5534  1.9760
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.12939   0.01578 -71.551 < 2e-16 ***
## DAY_OF_WEEK_REF_2  -0.23001   0.02321 -9.911 < 2e-16 ***
## DAY_OF_WEEK_REF_3  -0.36240   0.02361 -15.351 < 2e-16 ***
## DAY_OF_WEEK_REF_4  -0.08823   0.02260 -3.904 9.44e-05 ***
## DAY_OF_WEEK_REF_5  -0.14454   0.02286 -6.323 2.56e-10 ***
## DAY_OF_WEEK_REF_6  -0.66970   0.02677 -25.017 < 2e-16 ***
## DAY_OF_WEEK_REF_7  -0.09006   0.02290 -3.933 8.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 149532 on 145951 degrees of freedom
## Residual deviance: 148658 on 145945 degrees of freedom
## AIC: 148672
##
## Number of Fisher Scoring iterations: 4

```

Tal como vemos aqui, segun el p-value, todos los dias son significativos.

Pero como saber si el dia de la semana es un factor de riesgo o no? *Calculando el Odds Ratio.*

Como sabemos los odds es la razón de la probabilidad de ocurrencia de un suceso entre la probabilidad de su no ocurrencia. En nuestro caso la probabilidad de ocurrencia del retraso.

$$Odds = \frac{p}{1-p}$$

donde p es la probabilidad de que el individuo tome el valor “1” en la variable dicotómica.

Mientras que el Odds Ratio se obtiene como el cociente entre ambos odds. Donde la variable respuesta Y está presente entre los individuos, es decir, toma el valor $Y = 1$, y la variable independiente X puede estar presente o no, es decir, tomar los valores $X = 1$ y $X = 0$. Teniendo:

$$OR = \frac{\frac{p(Y=1/X=1)}{1-p(Y=1/X=1)}}{\frac{p(Y=1/X=0)}{1-p(Y=1/X=0)}} = e^{b_1}$$

Pudiendo darse 3 situaciones:

- Un OR = 1 implica que no existe asociación entre la variable respuesta y la covariable.
- Un OR inferior a la unidad se interpreta como un factor de protección, es decir, el suceso es menos probable en presencia de dicha covariable.
- Un OR mayor a la unidad se interpreta como un factor de riesgo, es decir, el suceso es más probable en presencia de dicha covariable

Veamos entonces que obtenemos con los datos de nuestro modelo, para ello podemos directamente ejecutar la funcion R exp, con la cual se calculará la exponencial de los coeficientes del modelo obtenido

```
exp(coefficients(model_22a))
```

```
##           (Intercept) DAY_OF_WEEK_REF_2 DAY_OF_WEEK_REF_3 DAY_OF_WEEK_REF_4
##             0.3232305      0.7945262      0.6960058      0.9155498
## DAY_OF_WEEK_REF_5 DAY_OF_WEEK_REF_6 DAY_OF_WEEK_REF_7
##             0.8654193      0.5118637      0.9138735
```

Tal como se ve todos los OR son menores a 1 por lo que podemos decir que el dia de la semana no es un factor de riesgo del retraso. Esto concuerda con el analisis visual que hicimos en el punto 2.1.

2.2.b. Con 1 una variable independiente: AIRLINE.

Idem al anterior tomando como variable explicativa AIRLINE. Se tomará como aerolínea de referencia AA. Se puede considerar que la aerolínea es un factor de riesgo?

Realicemos lo mismo que en el punto a. Generaremos el modelo y calculemos el OR.

```
AIRLINE_REF_=relevel(factor(vuelos$AIRLINE), ref = 'AA')

model_22b=glm(formula=vuelos$DELAY_SFO~AIRLINE_REF_,family=binomial(link=logit))

summary(model_22b)
```

```
##
## Call:
## glm(formula = vuelos$DELAY_SFO ~ AIRLINE_REF_, family = binomial(link = logit))
##
## Deviance Residuals:
##    Min      1Q   Median      3Q      Max
## -0.7408  -0.7216  -0.6980  -0.5939   2.1710
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.64589   0.02473 -66.547 < 2e-16 ***
## AIRLINE_REF_AS 0.07718   0.04443   1.737  0.08235 .
## AIRLINE_REF_B6 0.43330   0.04206  10.302 < 2e-16 ***
## AIRLINE_REF_DL 0.02611   0.03691   0.707  0.47934
## AIRLINE_REF_F9 0.45304   0.05912   7.663 1.82e-14 ***
## AIRLINE_REF_HA -0.61123   0.13469  -4.538 5.68e-06 ***
## AIRLINE_REF_OO 0.35794   0.02800  12.782 < 2e-16 ***
## AIRLINE_REF_UA 0.49304   0.02708  18.209 < 2e-16 ***
## AIRLINE_REF_US -0.51044   0.06899  -7.399 1.37e-13 ***
## AIRLINE_REF_VX 0.08503   0.03243   2.622  0.00874 **
## AIRLINE_REF_WN 0.42703   0.03199  13.348 < 2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 149532  on 145951  degrees of freedom
## Residual deviance: 148491  on 145941  degrees of freedom
## AIC: 148513
```

```
##  
## Number of Fisher Scoring iterations: 4
```

Como puede verse no todas los p-value de cada variable son significativas. Por ej la aerolina VX, que es medianamente significativa o la DL que directamente no lo es. O incluso As.

Pero que pasa con el OR?

```
exp(coefficients(model_22b))
```

```
##      (Intercept) AIRLINE_REF_AS AIRLINE_REF_B6 AIRLINE_REF_DL AIRLINE_REF_F9
## 0.1928402     1.0802404    1.5423464    1.0264521    1.5730938
## AIRLINE_REF_HA AIRLINE_REF_OO AIRLINE_REF_UA AIRLINE_REF_US AIRLINE_REF_VX
## 0.5426834     1.4303817    1.6372924    0.6002313    1.0887550
## AIRLINE_REF_WN
## 1.5326970
```

A partir de los OR obtenidos volvemos a validar el análisis visual del punto 2.1 donde habíamos determinado que las aerolíneas US y HA tenían menos retrasos. Y aquí rápidamente con el OR podemos ver que solo estas dos aerolíneas no son un factor de riesgo del retraso. Siendo la mayoría de ellas un factor de riesgo.

2.2.c. Con 2 variables independientes: DAY_OF_WEEK y DISTANCE

Se creará un modelo con la variable dependiente y las variables explicativas DAY_OF_WEEK (la obtenida en el apartado a) y DISTANCE. ¿Se observa una mejora con referencia a los anteriores?

Incluimos entonces en el modelo la variable DISTANCE.

```
model_22c=glm(formula=DELAY_SFO~DAY_OF_WEEK_REF_+DISTANCE,family=binomial(link=logit), data=vuelos)
summary(model_22c)
```

```
##  
## Call:  
## glm(formula = DELAY_SFO ~ DAY_OF_WEEK_REF_ + DISTANCE, family = binomial(link = logit),  
##       data = vuelos)  
##  
## Deviance Residuals:  
##      Min        1Q      Median        3Q       Max  
## -0.7525   -0.7185   -0.6782   -0.5526    1.9834  
##  
## Coefficients:  
##                               Estimate Std. Error z value Pr(>|z|)  
## (Intercept)      -1.116e+00  1.798e-02 -62.083 < 2e-16 ***  
## DAY_OF_WEEK_REF_2 -2.300e-01  2.321e-02 -9.911 < 2e-16 ***  
## DAY_OF_WEEK_REF_3 -3.623e-01  2.361e-02 -15.348 < 2e-16 ***  
## DAY_OF_WEEK_REF_4 -8.811e-02  2.260e-02 -3.899 9.66e-05 ***  
## DAY_OF_WEEK_REF_5 -1.444e-01  2.286e-02 -6.317 2.67e-10 ***  
## DAY_OF_WEEK_REF_6 -6.698e-01  2.677e-02 -25.021 < 2e-16 ***  
## DAY_OF_WEEK_REF_7 -8.999e-02  2.290e-02 -3.929 8.52e-05 ***  
## DISTANCE         -1.124e-05  7.213e-06 -1.559    0.119  
## ---
```

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 149532  on 145951  degrees of freedom
## Residual deviance: 148655  on 145944  degrees of freedom
## AIC: 148671
##
## Number of Fisher Scoring iterations: 4

```

El p-value de DISTANCE no es significativo. Sin embargo chequeemos el OR:

```
exp(coefficients(model_22c))
```

```

## (Intercept) DAY_OF_WEEK_REF_2 DAY_OF_WEEK_REF_3 DAY_OF_WEEK_REF_4
## 0.3276020    0.7945386    0.6960534    0.9156643
## DAY_OF_WEEK_REF_5 DAY_OF_WEEK_REF_6 DAY_OF_WEEK_REF_7      DISTANCE
## 0.8655449    0.5117976    0.9139416    0.9999888

```

El OR de distancia es casi igual a 1 lo cual implica, ademas de no ser significativa, que no existe asociación entre la variable respuesta y la covariable. O sea entre el Retraso y la Distancia.

Pero esto no quiere decir que sea 100% concluyente, se podria realizar agrupamientos por distancias, como hemos visto con las regresiones lineales, donde para distancias cortas y distancias largas se podian obtener modelos distintos. Se podria hasta realizar agrupaciones por dias de la semana y fines de semana, se podria analizar aerolineas low-cost vs aerolineas consideradas de primera linea, etc..

2.2.d. Seleccion de las variables mas significativas

Se creará un nuevo modelo con la variable dependiente y tomando como variables explicativas, aquéllas que han sido significativas en los apartados anteriores, y además se añadirá la variable ARRIVAL_DELAY. ¿Se observa una mejora con referencia a los anteriores? Realizad el cálculo de las OR.

Usamos solo estas dos que fueron significativas y no la distancia, por lo que sumando el arrival_delay obtenemos los siguientes resultados

```
model_22d=glm(formula=DELAY_SFO~DAY_OF_WEEK_REF_+AIRLINE_REF_+ARRIVAL_DELAY,
               family=binomial(link=logit), data=vuelos)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model_22d)
```

```

##
## Call:
## glm(formula = DELAY_SFO ~ DAY_OF_WEEK_REF_ + AIRLINE_REF_ + ARRIVAL_DELAY,
##       family = binomial(link = logit), data = vuelos)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -6.4560   -0.3203   -0.1759   -0.0681    3.9342
##
```

```

## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.8691554  0.0498514 -57.554 < 2e-16 ***
## DAY_OF_WEEK_REF_2   -0.0421141  0.0397209 -1.060  0.2890
## DAY_OF_WEEK_REF_3    0.0044199  0.0393901  0.112  0.9107
## DAY_OF_WEEK_REF_4   -0.0734021  0.0388849 -1.888  0.0591 .
## DAY_OF_WEEK_REF_5   -0.0569314  0.0392853 -1.449  0.1473
## DAY_OF_WEEK_REF_6   -0.0878411  0.0442668 -1.984  0.0472 *
## DAY_OF_WEEK_REF_7   -0.0165176  0.0398777 -0.414  0.6787
## AIRLINE_REF_AS      -0.2213846  0.0779260 -2.841  0.0045 **
## AIRLINE_REF_B6       0.4745195  0.0716104  6.626 3.44e-11 ***
## AIRLINE_REF_DL       0.1038926  0.0619854  1.676  0.0937 .
## AIRLINE_REF_F9       -0.5881464  0.1061365 -5.541 3.00e-08 ***
## AIRLINE_REF_HA       -2.5285097  0.2180803 -11.594 < 2e-16 ***
## AIRLINE_REF_OO       0.0558507  0.0475452  1.175  0.2401
## AIRLINE_REF_UA       0.8107268  0.0455949  17.781 < 2e-16 ***
## AIRLINE_REF_US       -0.5855628  0.1062483 -5.511 3.56e-08 ***
## AIRLINE_REF_VX       -0.1219613  0.0541320 -2.253  0.0243 *
## AIRLINE_REF_WN       0.5355286  0.0540037  9.917 < 2e-16 ***
## ARRIVAL_DELAY        0.1352095  0.0008587 157.450 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 148916  on 145490  degrees of freedom
## Residual deviance: 58403  on 145473  degrees of freedom
## (461 observations deleted due to missingness)
## AIC: 58439
##
## Number of Fisher Scoring iterations: 7

```

Al combinar estas variables, se evidencia un cambio tanto en los coeficientes de cada variable como tambien su nivel de significancia.

Veamos que ocurre con los Odds

```
exp(coefficients(model_22d))
```

```

## (Intercept) DAY_OF_WEEK_REF_2 DAY_OF_WEEK_REF_3 DAY_OF_WEEK_REF_4
## 0.05674683     0.95876039     1.00442971     0.92922713
## DAY_OF_WEEK_REF_5 DAY_OF_WEEK_REF_6 DAY_OF_WEEK_REF_7  AIRLINE_REF_AS
## 0.94465891     0.91590644     0.98361807     0.80140838
## AIRLINE_REF_B6  AIRLINE_REF_DL  AIRLINE_REF_F9  AIRLINE_REF_HA
## 1.60724170     1.10948132     0.55535572     0.07977782
## AIRLINE_REF_OO  AIRLINE_REF_UA  AIRLINE_REF_US  AIRLINE_REF_VX
## 1.05743981     2.24954225     0.55679242     0.88518262
## AIRLINE_REF_WN  ARRIVAL_DELAY
## 1.70835106     1.14477653

```

Esta combinacion de variables muestra que en su conjunto son un factor de riesgo del retraso.

Por ultimo comparemos la bondad del ajuste de los 4 modelos a partir del coeficiente AIC:

```
model_22a$aic
```

```
## [1] 148671.7
```

```
model_22b$aic
```

```
## [1] 148513.4
```

```
model_22c$aic
```

```
## [1] 148671.3
```

```
model_22d$aic
```

```
## [1] 58438.71
```

Como vemos, con este ultimo modelo (model_22d) al incluir las 3 variables mencionadas obtenemos el valor mas bajo de AIC, o sea, es el modelo que mejor se ajusta a los datos. Los 3 previos poseen valores muy similares de AIC entre ellos.

2.3. Predicción

Según el modelo del apartado c), calcula la probabilidad de retraso en el vuelo, si nuestro destino está a 1500 millas y viajamos en jueves.

```
predict_2.3 = data.frame(DAY_OF_WEEK_REF_ = "4", DISTANCE=1500)  
  
predicted_2.3 = predict(model_22c, newdata=predict_2.3, type ="response")  
predicted_2.3
```

```
## 1  
## 0.2277729
```

De acuerdo con el modelo, se obtiene una probabilidad de retraso de 0.23.

2.4. Bondad del ajuste

Usa el test de Hosman-Lemeshow para ver la bondad de ajuste, tomando el modelo del apartado c). En la librería ResourceSelection hay una función que ajusta el test de Hosmer- Lemeshow.

Se dice que un modelo presenta un buen ajuste a los datos si los valores que predice reflejan de manera adecuada los valores observados. Si el modelo presenta un mal ajuste, este no puede ser utilizado para extraer conclusiones ni efectuar predicciones.

Un modo de medir la adecuación de un modelo es proporcionando medidas globales de **bondad de ajuste** mediante test estadísticos.

Existen varias medidas de ajuste global para comparar la diferencia entre valores predichos y valores observados. Tres de las más utilizadas son:

- 1- el test basado en la devianza D,
- 2- el estadístico 2 (chi cuadrado) de Pearson y
- 3- el test de Hosmer-Lemeshow.

Los dos primeros se basan en los patrones de las covariables y pueden ser usados en los modelos lineales generalizados (MLG) en general.

El tercero se basa en probabilidades estimadas y se aplica en el caso de un MLG con distribución binomial, es decir, un modelo de regresión logística, que es justamente nuestro caso. Si una de las variables explicativas es continua (DISTANCE), no deben usarse los test 1 y 2, sino el test de Hosmer-Lemeshow. Este test consiste en comparar los valores previstos (esperados) por el modelo con los valores observados.

Veamoslo:

```
hoslem.test(vuelos$DELAY_SFO,fitted(model_22c))

## Warning in Ops.factor(1, y): '-' not meaningful for factors

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: vuelos$DELAY_SFO, fitted(model_22c)
## X-squared = 145952, df = 8, p-value < 2.2e-16
```

Este test se basa en las siguientes hipótesis:

H0: no hay diferencias entre las frecuencias observadas y las predichas (buen ajuste).

H1: sí hay diferencias (mal ajuste).

Por lo tanto dado que nuestro p-value es significativo, lo que implica el rechazo de H0 y por lo tanto que el modelo no ajusta bien a los datos.

2.5. Curva ROC

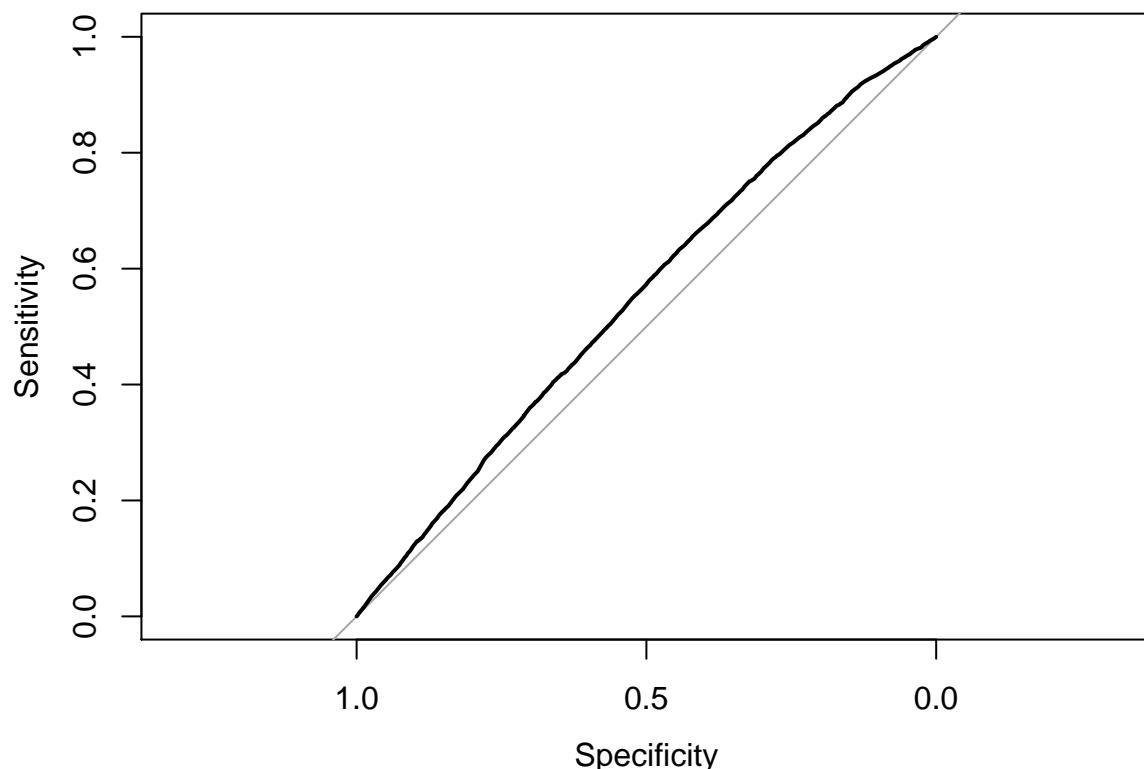
Dibujar la curva ROC, y calcular el área debajo de la curva con los modelos de los apartados c) y d). Discutir el resultado.

```
DELAY_SFO_DATA = vuelos[, c('DAY_OF_WEEK', 'DISTANCE')]
prob = predict(model_22c, newdata=DELAY_SFO_DATA, type ="response")
r=roc(vuelos$DELAY_SFO,prob, data=DELAY_SFO_DATA)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot (r)
```



```
auc(r)
```

```
## Area under the curve: 0.5505
```

El análisis ROC proporciona un modo de seleccionar modelos posiblemente óptimos y subóptimos basado en la calidad de la clasificación a diferentes niveles o umbrales. Para tener una regla objetiva de comparación de las curvas ROC, se calcula el área bajo la curva, simplemente llamada AUROC (area under the ROC). El modelo cuya área sea superior es el preferido.

En general:

- Si AUROC = 0,5, el modelo no ayuda a discriminar.
- Si $0,6 \leq \text{AUROC} < 0,8$, el modelo discrimina de manera adecuada.
- Si $0,8 \leq \text{AUROC} < 0,9$, el modelo discrimina de forma excelente.
- Si $\text{AUROC} \geq 0,9$, el modelo discrimina de modo excepcional.

Por lo tanto el valor obtenido para el modelo 2.2.c esta al límite entre discriminar de manera adecuada a no ayudar a discriminar. Situación que concuerda con lo visto en el punto 2.4 al momento de analizar la bondad del ajuste.

3. Conclusiones del análisis

Tanto con modelos lineales como con logaritmos podemos predecir de forma continua como dicotómica el retraso de los vuelos. Obtuvimos mejores o peores modelos y predictores según las diferentes alternativas

testeadas. Pero claro esta que este proceso solo fue el inicio porque hemos visto que segun que combinacion de variables podemos obtener mejores o peores ajustes, o segun que transformacion apliquemos a ciertas variables podemos mejorar la prediccion del modelo. Incluso segun que tipo de variables utilicemos y las relaciones que hubiera entre ellas sera mas optimo utilizar regresion lineal o logistica.

Referencias

Regresión lineal simple, Josep Gibergans Bàguena, P08/75057/02311

<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/plot.lm>

<https://r-coder.com/tabla-contingencia-r/>

<https://rpubs.com/osoramirez/111403>

<https://data.library.virginia.edu/diagnostic-plots/>