

# Actividad 3: Modelos predictivos

Solución

Semestre 2020.2

## Índice

<b>1. Modelo de regresión lineal</b>	<b>1</b>
1.1. Modelo de regresión lineal (regresores cuantitativos) . . . . .	1
1.2. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos) . . . . .	4
1.3. Diagnosis del modelo. . . . .	4
1.4. Predicció del model . . . . .	6
<b>2. Modelo de regresión logística.</b>	<b>6</b>
2.1. Estudio de relaciones entre variables. . . . .	6
2.2. Modelo de regresión logística. . . . .	10
2.3. Predicción . . . . .	14
2.4. Bondad del ajuste . . . . .	14
2.5. Curva ROC . . . . .	15
2.6. Conclusiones . . . . .	17

## 1. Modelo de regresión lineal

### 1.1. Modelo de regresión lineal (regresores cuantitativos)

- a) 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.

NOTA: En la base de datos los nombres de las variables están en mayúsculas.

- b) Se añadirá al modelo anterior la variable independiente DISTANCE. ¿Existe una mejora del ajuste?. Razonar.
- c) Posteriormente, se procederá 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. Razonar los resultados.

Solución:

a)

#Estimacion del modelo

```
Model.1.1<- lm(DEPARTURE_DELAY~ARRIVAL_DELAY, data=dat_SFO )
summary(Model.1.1)
```

```
## 
## Call:
## lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY, data = dat_SFO)
```

```

## 
## 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
cor(x = DEPARTURE_DELAY, y = ARRIVAL_DELAY, method = "pearson", use="pairwise.complete.obs")
## [1] 0.9529701

```

Como era de esperar existe una relación lineal positiva alta, entre ambas variables. Se observa que el coeficiente de determinación ajustado es: 0.9082. Si se calcula el coeficiente de correlación obtenemos un valor de 0.9529. La relación entre ambos es que el coeficiente de determinación es el cuadrado del coeficiente de correlación. Nota: Ya que el modelo sólo tiene una variable explicativa, el coeficiente de determinación coincide con el coeficiente de determinación ajustado.

b)

#### #Estimacion del modelo

```

Model.1.2<- lm(DEPARTURE_DELAY~ARRIVAL_DELAY+DISTANCE, data=dat_SF0 )
summary(Model.1.2)

```

```

## 
## Call:
## lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY + DISTANCE, data = dat_SF0)
## 
## Residuals:
##      Min       1Q    Median       3Q      Max
## -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

```

En el modelo del apartado a), el coeficiente de determinación es de 0.9082, y en este de 0.9127, por lo que el valor es muy parecido. No hay evidencia de mejora del modelo.

c)

```
#Estimacion del modelo

selected_BigD <- which(DISTANCE> 600)
data1=dat_SF0[selected_BigD,]
selected_LowD <- which(DISTANCE<= 600 )
data2=dat_SF0[selected_LowD,]
dim(data1)

## [1] 82581    28
dim(data2)

## [1] 63371    28
Model.1.1.1<- lm(DEPARTURE_DELAY~DISTANCE, data=data1 )
summary(Model.1.1.1)

##
## Call:
## lm(formula = DEPARTURE_DELAY ~ DISTANCE, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -44.85  -15.72  -12.52   -2.28 1484.62
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.808780  0.404317  24.26 < 2e-16 ***
## DISTANCE    0.001075  0.000206   5.22  1.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.29 on 82579 degrees of freedom
## Multiple R-squared:  0.0003298, Adjusted R-squared:  0.0003177
## F-statistic: 27.24 on 1 and 82579 DF, p-value: 1.797e-07

Model.1.1.2<- lm(DEPARTURE_DELAY~DISTANCE, data=data2)
summary(Model.1.1.2)

##
## Call:
## lm(formula = DEPARTURE_DELAY ~ DISTANCE, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -42.11  -15.61  -12.52   -1.32 762.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.406419  0.471962 19.930 <2e-16 ***
## DISTANCE    0.002698  0.001200  2.248  0.0246 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.84 on 63369 degrees of freedom
```

```

## Multiple R-squared:  7.976e-05, Adjusted R-squared:  6.398e-05
## F-statistic: 5.054 on 1 and 63369 DF, p-value: 0.02457

```

A la vista de los resultados, se puede concluir que en los vuelos que recorren más de 600 millas, la variable DISTANCE es más significativa, comparado con los vuelos más cortos. De todas formas, si se observan los coeficientes de determinación, la asociación entre la distancia y el retraso en la salida del vuelo es prácticamente nula.

## 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 en mayor y menor o igual a 15 minutos.

```

#Estimacion del modelo
late.air <- (LATE_AIRCRAFT_DELAY > 15)
late.air_RE <- ifelse(late.air==TRUE, 1, 0)
table(late.air_RE)

## late.air_RE
##      0      1
## 38739 13514

Model.1.3 = lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY + late.air_RE, data = dat_SFO)
summary(Model.1.3)

##
## Call:
## lm(formula = DEPARTURE_DELAY ~ ARRIVAL_DELAY + late.air_RE, data = dat_SFO)
##
## Residuals:
##       Min        1Q        Median        3Q       Max
## -167.250   -6.857    0.953     8.269    71.978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.933194  0.073492 12.70 <2e-16 ***
## ARRIVAL_DELAY 0.962026  0.001197 803.63 <2e-16 ***
## late.air_RE   6.565464  0.155841  42.13 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.96 on 52137 degrees of freedom
## (93812 observations deleted due to missingness)
## Multiple R-squared:  0.9419, Adjusted R-squared:  0.9419
## F-statistic: 4.227e+05 on 2 and 52137 DF, p-value: < 2.2e-16

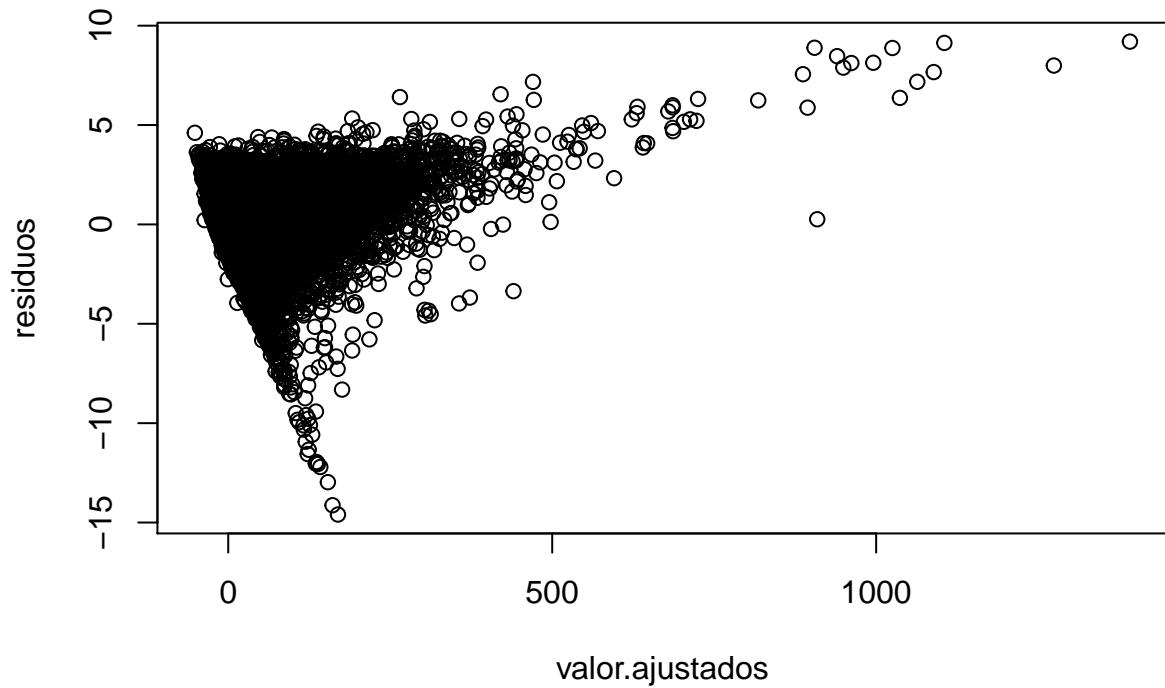
```

Se puede observar que las dos variables explicativas son significativas para el modelo. Por otro lado, se tiene que el coeficiente de determinación ajustado es 0,9419, por lo que el ajuste es muy bueno.

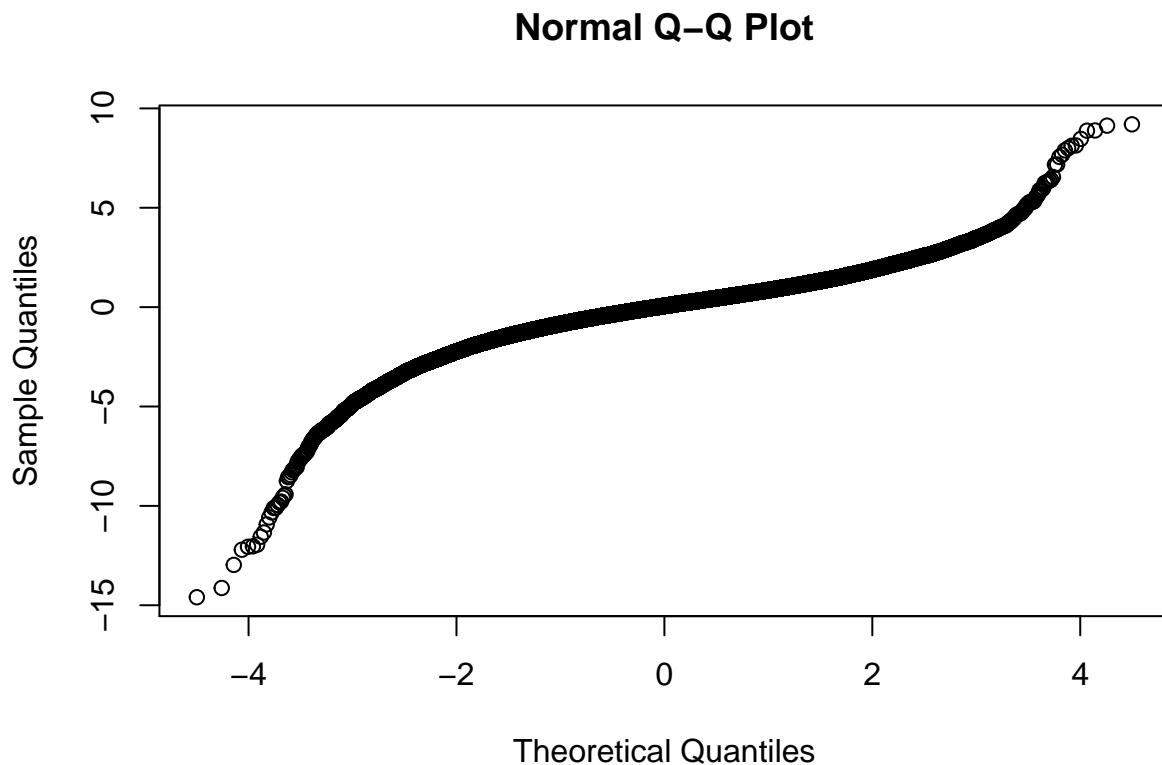
## 1.3. Diagnosis del modelo.

Para la diagnosis se escoge el modelo construido en el apartado 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). Interpretar los resultados.

```
residuos <- rstandard(Model.1.2)
valor.ajustados <- fitted(Model.1.2)
plot(valor.ajustados, residuos)
```



```
qqnorm(residuos)
qqnorm(residuos)
```



A la vista del gráfico se observa un patrón de dispersión irregular. Es decir no es un patrón aleatorio de los residuos. Esto indica que no se cumple el supuesto de varianza constante en los errores del modelo.

Por otro lado el Q\_Q plot, muestra que los datos no se ajustan bien a una normal.

#### 1.4. Predicció del model

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.

```
newdata = data.frame(DISTANCE = 2500, ARRIVAL_DELAY=30)
predict(Model.1.2, newdata)
```

```
##          1
## 37.34989
```

Se obtiene un valor de 37 minutos.

## 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.

- a) Visualizar la relación entre delay\_SFO y las variables independientes:DAY\_OF\_WEEK y AIRLINE. Calcular las frecuencias relativas por fila y columna. Interpretar el significado. (Visualizar con barplot).
- b) Para comprobar si existe asociación entre la 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. Interpretar.

Solución:

Se seleccionan los vuelos con retraso superior o igual a 15 minutos

```
delay.yes <- (dat_SFO$DEPARTURE_DELAY >= 15)
table(delay.yes)
```

```
## delay.yes
## FALSE    TRUE
## 115488  30464
```

Recodificar la variable

```
delay_SF0 <- ifelse(delay.yes==TRUE, 1, 0)
table(delay_SF0)
```

```
## delay_SF0
##      0      1
## 115488 30464
```

- a) Visualizar la relación

```
tw<-table(delay_SF0, DAY_OF_WEEK)
tw
```

```
##          DAY_OF_WEEK
## delay_SF0      1      2      3      4      5      6      7
##      0 16431 16911 17669 16747 16734 15068 15928
##      1  5311  4343  3975  4956  4681  2493  4705
```

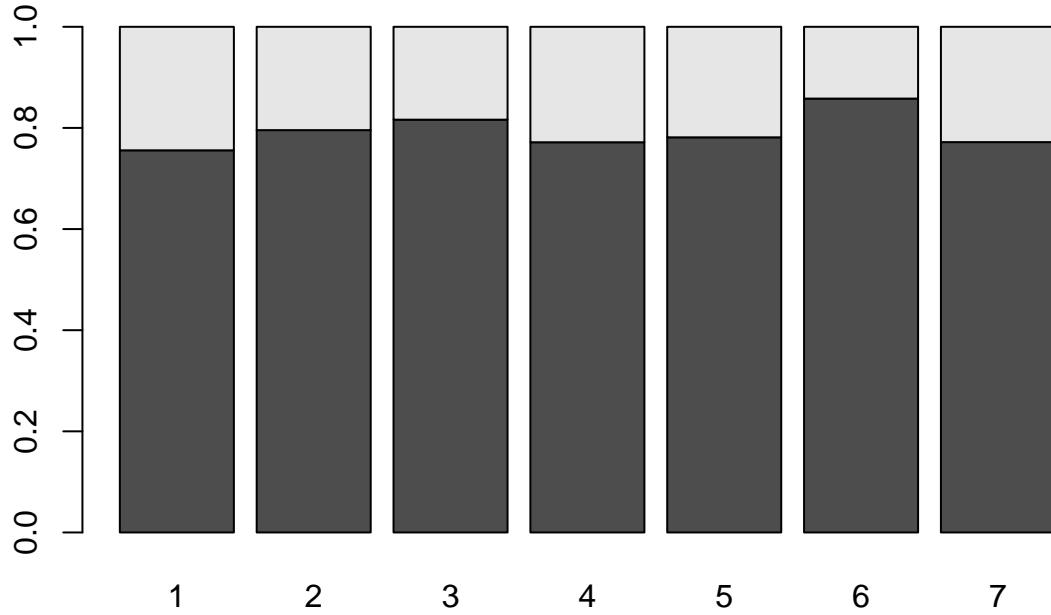
```
tw_rel<-prop.table(tw,1)
tw_rel
```

```
##          DAY_OF_WEEK
## delay_SF0      1      2      3      4      5      6      7
##      0 0.1422745 0.1464308 0.1529943 0.1450107 0.1448982 0.1304724 0.1379191
##      1 0.1743369 0.1425617 0.1304819 0.1626838 0.1536568 0.0818343 0.1544446
```

```
tw_rel_col<-prop.table(tw,2)
tw_rel_col
```

```
##          DAY_OF_WEEK
## delay_SF0      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
```

```
barplot(tw_rel_col)
```



```
ta<-table(delay_SFO,AIRLINE)  
ta
```

```
##          AIRLINE  
## delay_SFO    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
```

```
ta_rel<-prop.table(ta,1)  
ta_rel
```

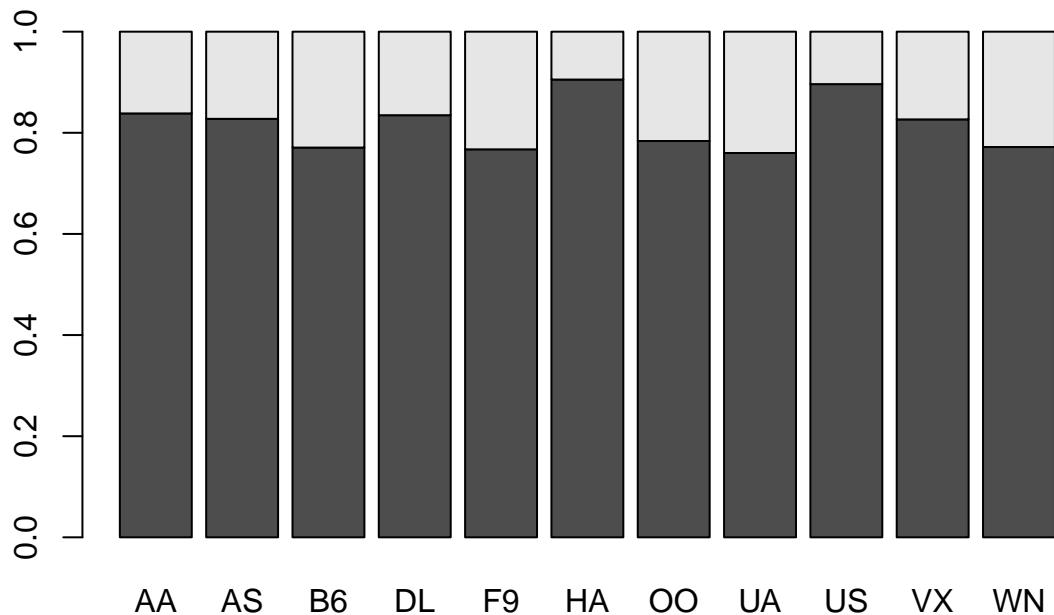
```
##          AIRLINE  
## delay_SFO      AA          AS          B6          DL          F9  
##           0 0.087558881 0.036869631 0.032635425 0.069816778 0.012901773  
##           1 0.064009979 0.029116334 0.036797532 0.052389706 0.014837185  
##          AIRLINE  
## delay_SFO      HA          OO          UA          US          VX  
##           0 0.005212663 0.232266556 0.297035190 0.020123303 0.113414381  
##           1 0.002068015 0.242876838 0.355534401 0.008830095 0.090270483  
##          AIRLINE  
## delay_SFO      WN  
##           0 0.092165420  
##           1 0.103269433
```

```
ta_rel_col<-prop.table(ta,2)  
ta_rel_col
```

```

##          AIRLINE
## delay_SFO      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
##          AIRLINE
## delay_SFO      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
barplot(ta_rel_col)

```



b) delay\_SFO y DAY\_OF\_WEEK, AIRLINE

```

chi.test<-chisq.test(tw)
print(chi.test)

##
##  Pearson's Chi-squared test
##
##  data:  tw
##  X-squared = 834.95, df = 6, p-value < 2.2e-16
chi.test<-chisq.test(ta)
print(chi.test)

##
##  Pearson's Chi-squared test

```

```

## 
## data: ta
## X-squared = 986.78, df = 10, p-value < 2.2e-16

a)

```

En el caso de la variable DAY\_OF\_WEEK, se observa que no hay mucha diferencia entre el número de vuelos totales por días, por lo que la comparación por filas, como por columnas tendría sentido.(Nótese que el sábado es el día con menor número de vuelos).

En cambio, en las aerolíneas (AIRLINE), el número de vuelos varía bastante, según sea una aerolínea u otra, por lo tanto si se comparase el porcentaje por filas, se podría caer en el error de afirmar que las aerolíneas con mayor retraso son OO y UA, con un 24% y 35% de retrasos. El motivo de estos valores es que son las compañías con mayor número de vuelos. Por lo tanto, en este caso, se tendría que comparar sólamente los porcentajes por columna.

Con referencia a DAY\_OF\_WEEK, y observando el porcentaje por columnas, se puede apreciar que los miércoles y sábado, el porcentaje de vuelos con retraso es menor al resto de los días (porcentaje inferior al 20%). De estos, es el sábado es el día con menos retrasos. En cambio, el lunes, sería el día con más retrasos con un 24,4%, sobre el total de vuelos que han salido en lunes. Si nos fijamos en el porcentaje por fila, también se ve que el sábado sería el día con menor retraso, con un 8% de los vuelos, sobre el total de retrasados.

Respecto a AIRLINE, y observando el porcentaje por columnas, se puede apreciar que las aerolíneas con porcentaje de vuelos retrasados superior al 20% son B6, F9, OO, UA y WN, siendo UA (United Airlines), la de mayor porcentaje, con un 24%, seguida de F9 (Frontier Airlines), con un 23,2%. Obsérvese, que si nos hubiésemos basado en el porcentaje de fila, F9, aparecería como una de las líneas con menos retraso, debido a que el número de vuelos anual es mucho menor que el de las aerolíneas más grandes.

Al visualizar los diagramas de barra, se puede llegar a la misma conclusión.

b)

A la vista de los resultados, se puede observar que el p\_value es prácticamente 0 en ambas variables. Por lo tanto se rechaza la hipótesis nula de independencia de variables, aceptando que existe una relación significativa entre delay\_SFO y DAY\_OF\_WEEK y delay\_SFO y AIRLINE.

## 2.2. Modelo de regresión logística.

- 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.
- 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? Justifica tu respuesta.
- 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? Explicar.
- 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? Explicar. Realizad el cálculo de las OR.

Solución:

- Modelo 1

```

DAY_OF_WEEK<-as.factor(as.numeric(DAY_OF_WEEK))
week_Rel=relevel(DAY_OF_WEEK, ref = '1')

```

```

logit_model_1 <- glm(formula=delay_SFO~week_Rel, data=dat_SFO, family=binomial)
summary(logit_model_1)

##
## Call:
## glm(formula = delay_SFO ~ week_Rel, family = binomial, data = dat_SFO)
##
## 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 ***
## week_Rel2   -0.23001  0.02321 -9.911 < 2e-16 ***
## week_Rel3   -0.36240  0.02361 -15.351 < 2e-16 ***
## week_Rel4   -0.08823  0.02260 -3.904 9.44e-05 ***
## week_Rel5   -0.14454  0.02286 -6.323 2.56e-10 ***
## week_Rel6   -0.66970  0.02677 -25.017 < 2e-16 ***
## week_Rel7   -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

```

Se obtiene un p-valor inferior a 0.05, por lo que podemos concluir que existe relación entre la variable delay\_SFO y DAY\_OF\_WEEK. Se puede considerar día de la semana como un factor de riesgo, ya que dependiendo el día que viajemos tendremos más probabilidad de sufrir un retraso en la salida del avión. Se observa que los coeficientes son negativos, por lo que la probabilidad de sufrir un retraso el resto de los días de la semana, con respecto al lunes, es menor, coincidiendo con lo comentado en el apartado a) anterior. Se interpretarán los resultados de las OR, en el modelo del apartado d).

b) Modelo 2

```

airline<-factor(AIRLINE)
airline_Rel=relevel(airline, ref = 'AA')
logit_model_2 <- glm(formula=delay_SFO~airline_Rel, data=dat_SFO, family=binomial)
summary(logit_model_2)

##
## Call:
## glm(formula = delay_SFO ~ airline_Rel, family = binomial, data = dat_SFO)
##
## 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_RelAS 0.07718  0.04443  1.737  0.08235 .
## airline_RelB6 0.43330  0.04206 10.302 < 2e-16 ***
## airline_RelDL 0.02611  0.03691  0.707  0.47934
## airline_RelF9 0.45304  0.05912  7.663 1.82e-14 ***
## airline_RelHA -0.61123  0.13469 -4.538 5.68e-06 ***
## airline_RelOO 0.35794  0.02800 12.782 < 2e-16 ***
## airline_RelUA 0.49304  0.02708 18.209 < 2e-16 ***
## airline_RelUS -0.51044  0.06899 -7.399 1.37e-13 ***
## airline_RelVX 0.08503  0.03243  2.622  0.00874 **
## airline_RelWN 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

```

Al igual que en el apartado anterior, también se puede concluir que dependiendo de la aerolínea, podremos tener un mayor o menor retraso.

### c) Modelo 3

```
logit_model_3 <- glm(formula=delay_SFO~week_Rel+DISTANCE, data=dat_SFO, family=binomial)
summary(logit_model_3)
```

```

##
## Call:
## glm(formula = delay_SFO ~ week_Rel + DISTANCE, family = binomial,
##      data = dat_SFO)
##
## 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 ***
## week_Rel2   -2.300e-01  2.321e-02 -9.911 < 2e-16 ***
## week_Rel3   -3.623e-01  2.361e-02 -15.348 < 2e-16 ***
## week_Rel4   -8.811e-02  2.260e-02 -3.899 9.66e-05 ***
## week_Rel5   -1.444e-01  2.286e-02 -6.317 2.67e-10 ***
## week_Rel6   -6.698e-01  2.677e-02 -25.021 < 2e-16 ***
## week_Rel7   -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
exp(coefficients(logit_model_3))

## (Intercept) week_Rel2 week_Rel3 week_Rel4 week_Rel5 week_Rel6
## 0.3276020 0.7945386 0.6960534 0.9156643 0.8655449 0.5117976
## week_Rel7 DISTANCE
## 0.9139416 0.9999888

```

En base al indicador  $AIC = 145983$ , no se ve una mejora en el modelo, con referencia a los anteriores. Por otro lado, la OR de DISTANCE es prácticamente 1, por lo que no existiría relación entre los retrasos de los vuelos y la distancia recorrida. A esta misma conclusión se puede llegar si se observa el  $p\_value$  asociado a DISTANCE.

d) Modelo 4

```

logit_model_4 <- glm(formula=delay_SFO~week_Rel+airline_Rel+ARRIVAL_DELAY, data=dat_SFO, family=binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
summary(logit_model_4)

##
## Call:
## glm(formula = delay_SFO ~ week_Rel + airline_Rel + ARRIVAL_DELAY,
##      family = binomial, data = dat_SFO)
##
## 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 ***
## week_Rel2   -0.0421141 0.0397209 -1.060  0.2890
## week_Rel3    0.0044199 0.0393901  0.112  0.9107
## week_Rel4   -0.0734021 0.0388849 -1.888  0.0591 .
## week_Rel5   -0.0569314 0.0392853 -1.449  0.1473
## week_Rel6   -0.0878411 0.0442668 -1.984  0.0472 *
## week_Rel7   -0.0165176 0.0398777 -0.414  0.6787
## airline_RelAS -0.2213846 0.0779260 -2.841  0.0045 **
## airline_RelB6  0.4745195 0.0716104  6.626 3.44e-11 ***
## airline_RelDL  0.1038926 0.0619854  1.676  0.0937 .
## airline_RelF9 -0.5881464 0.1061365 -5.541 3.00e-08 ***
## airline_RelHA -2.5285097 0.2180803 -11.594 < 2e-16 ***
## airline_RelOO  0.0558507 0.0475452  1.175  0.2401
## airline_RelUA  0.8107268 0.0455949 17.781 < 2e-16 ***
## airline_RelUS -0.5855628 0.1062483 -5.511 3.56e-08 ***
## airline_RelVX -0.1219613 0.0541320 -2.253  0.0243 *
## airline_RelWN  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
exp(coefficients(logit_model_4))

## (Intercept) week_Rel2 week_Rel3 week_Rel4 week_Rel5
## 0.05674683 0.95876039 1.00442971 0.92922713 0.94465891
## week_Rel6 week_Rel7 airline_RelAS airline_RelB6 airline_RelDL
## 0.91590644 0.98361807 0.80140838 1.60724170 1.10948132
## airline_RelF9 airline_RelHA airline_RelOO airline_RelUA airline_RelUS
## 0.55535572 0.07977782 1.05743981 2.24954225 0.55679242
## airline_RelVX airline_RelWN ARRIVAL_DELAY
## 0.88518262 1.70835106 1.14477653

```

En vista a los resultados de los apartados anteriores, se toman como variables explicativas aquéllas que son significativas: week\_Rel, airline\_Rel y ARRIVAL\_DELAY. El indicador AIC= 58439 es menor que en los otros modelos, por lo que existe una mejora en el ajuste.

En base a las OR ajustadas se tiene:

Las OR de week\_Rel, ajustadas por airline\_Rel y ARRIVAL\_DELAY, tienen valores menores que 1, excepto el miércoles. Esto se debe a que viajar en un día comparando con el lunes, ajustando con airline\_Rel y ARRIVAL\_DELAY, el retraso o bien sería el mismo o bien un poco menor. De todas formas, se observa que al ajustar el modelo por Airline y ARRIVAL\_DELAY, el día de la semana no parece tan significativo, exceptuando el sábado.

Las OR de airline\_Rel ajustadas por week\_Rel y ARRIVAL\_DELAY, nos indican, por ejemplo que la aerolínea WN tiene una probabilidad 1,7 veces mayor de sufrir un retraso con respecto a la línea AA (American Airlines), siendo UA la aerolínea con mayor retraso, con 2,25 veces más probabilidad de retraso en la salida de un vuelo.

Por otro lado, la OR ajustada de ARRIVAL\_DELAY es de 1.14, lo que nos indica que a medida que aumenta en una unidad el retraso en destino, el retraso en la salida aumenta 1.14 veces.

### 2.3. Predicción

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

```

pred<-predict(logit_model_3, data.frame(week_Rel="4",DISTANCE=1500),type = "response")
pred

## 1
## 0.22777729

```

El modelo del apartado c) nos predice una probabilidad del 22,7 % de poder sufrir un retraso, si viajamos en jueves a un destino que está a 1500 millas.

### 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.

```

library(ResourceSelection)

## ResourceSelection 0.3-5 2019-07-22
hoslem.test(delay_SF0,fitted(logit_model_3))

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: delay_SF0, fitted(logit_model_3)
## X-squared = 16.805, df = 8, p-value = 0.03221

```

La probabilidad es de 0.0001343 lo que indicaría, como era de esperar que el modelo del apartado c) no ajusta correctamente los datos. Recordad que se había comprobado que la variable DISTANCE no es significativa.

## 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.

```

library(pROC)

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

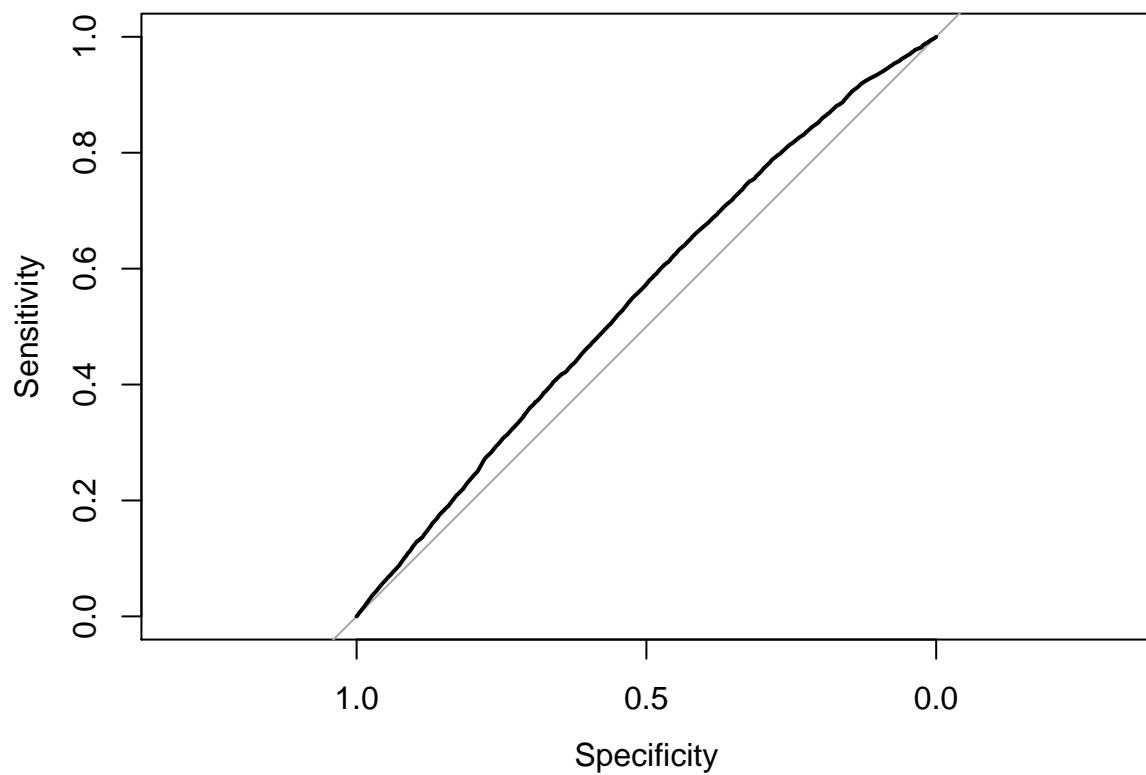
##
## Attaching package: 'pROC'

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

prob_low=predict(logit_model_3, dat_SF0, type="response")
r=roc(delay_SF0,prob_low, data=dat_SF0)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
plot(r)

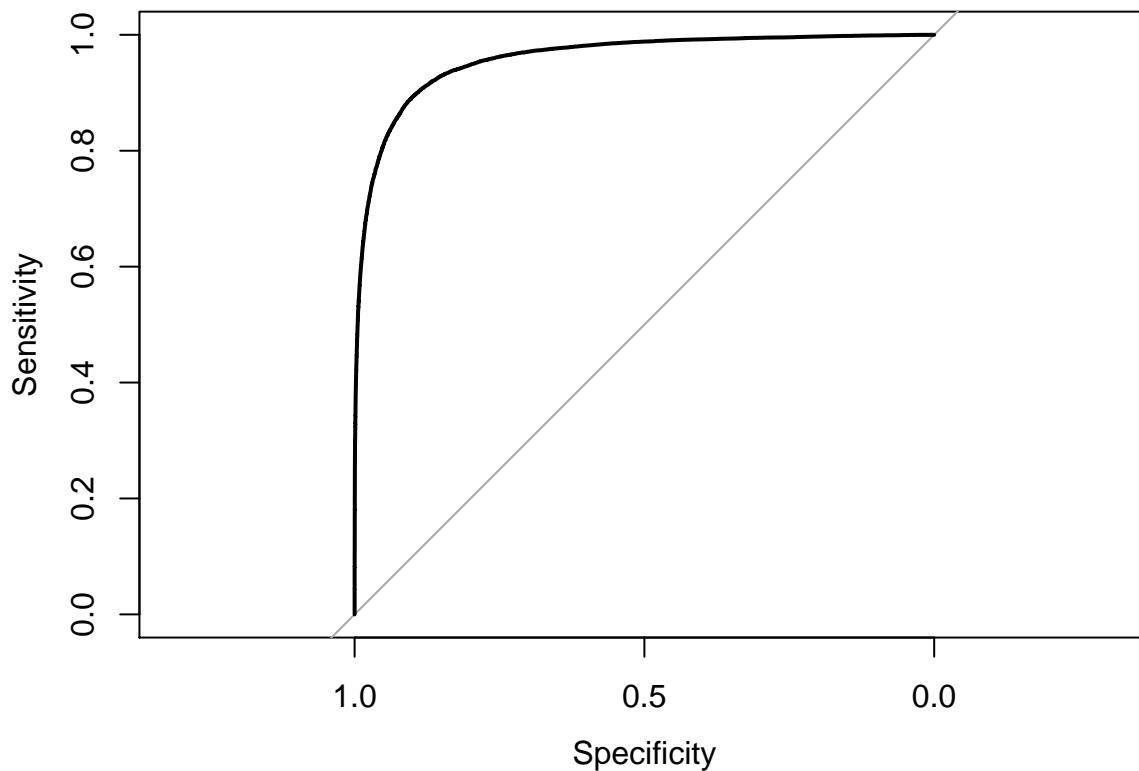
```



```
auc(r)

## Area under the curve: 0.5505
prob_low2=predict(logit_model_4, dat_SF0, type="response")
r=roc(delay_SF0,prob_low2, data=dat_SF0)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
plot(r)
```



```
auc(r)
```

```
## Area under the curve: 0.9594
```

En el primer modelo, el área por debajo de esa curva toma el valor de 0.55, por lo que la habilidad del modelo para predecir retrasos no es muy buena. En cambio, en el modelo 4, el área por debajo de esa curva toma el valor de 0.96, por lo que la habilidad del modelo para predecir retrasos es muy buena, como era de esperar.

## 2.6. Conclusiones

En la primera parte se han estudiado las posibles asociaciones lineales entre la variable DEPARTURE\_DELAY y diferentes covariables. Como era de esperar y a la vista de los resultados, se ve la existencia de una relación lineal positiva muy alta entre dicha variable y ARRIVAL\_DELAY. También podemos concluir que existe relación lineal entre dicha variable dependiente y LATE\_AIRCRAFT\_DELAY, siendo muy débil con la variable DISTANCE.

Una vez ajustado el último modelo, se ha obtenido la recta de regresión:  $y = 0.933194 + 0.962026 * \text{ARRIVAL\_DELAY} + 6.565464 * \text{LATE\_AIRCRAFT\_DELAY}$ , con un coeficiente de determinación ajustado de 0,9419, por lo que el ajuste es muy bueno.

En vista a los resultados obtenidos con los modelos de regresión logística, las variables explicativas DAY\_OF\_WEEK, AIRLINE y ARRIVAL\_DELAY, pueden considerarse factores de riesgo a la hora de que la salida de un vuelo sea o no puntual. En este caso se ve que el día más conflictivo sería el lunes, siendo las líneas aéreas UA (United Airlines), WN (Southwest Airlines) y B6 (JetBlue Airlines), las que más retrasos tienen.

En base al modelo logístico, se tiene que la variable DISTANCE es no significativa, por lo que viajar a un destino más alejado, no parece ser causa de un retraso superior o igual a 15 minutos en la salida del vuelo.

Por otro lado, del estudio de la curva ROC, se puede deducir que la habilidad del modelo del apartado d) para predecir retrasos en la salida de un vuelo es muy buena.

En este estudio, se ha demostrado la importancia de la elección de variables, pudiendo pasar de un modelo poco efectivo a otro con una probabilidad de diagnóstico alta.

Podemos concluir que los modelos de regresión son una buena herramienta para predecir la probabilidad de retraso de un vuelo.