— MEMORIA —

Métodos Lineales de Regresión y Clasificación
PRÁCTICA 4

CURSO 2023/2024

Simón Alonso (821038)
Óscar Brizuela (820773)
Jan Carlos Franco (806955)

Índice

1. Calidad de la cerveza

1.1. Valores estadísticos

Para cada uno de los modelos de regresión lineal que vamos a utilizar en este ejercicio, realizamos un análisis de sus valores estadísticos mediante la función *summary()* de *R*. Esta función nos permite conocer los valores estadísticos que vamos a explicar en este apartado, tanto aquellos correspondientes a los coeficientes de los predictores utilizados en los modelos como aquellos intrínsecos del modelo en cuestión.
La nomenclatura utilizada en R para cada uno de los modelos es la siguiente:

model_<prueba>_<lista_predictores>

De esta forma, el nombre del modelo almacena el nombre de la prueba que va a tratar de explicar el modelo (variable objetivo) y la lista de predictores (que, en caso de ser más de uno, estarán separados entre sí por el carácter "_"). Por ejemplo, el modelo que trata de explicar la prueba "ASA" y utiliza los componentes químicos *TPC* y *TSO2* como predictores se llamará *model_asa_tpc_tso2*.

Cabe destacar también que para cada una de las 5 pruebas se han realizado los 7 modelos lineales correspondientes (con los 3 componentes químicos por separado, por parejas y con los 3 componentes juntos), realizándose así un total de 35 pruebas (5 pruebas * 7 modelos cada prueba). Sin embargo, se ha decidido considerar los 2 modelos más interesantes para cada prueba teniendo en cuenta la influencia de los predictores.

1.1.1. Estadísticos de los coeficientes calculados

- *Estimate*: En primer lugar, podemos ver los valores de los coeficientes β_i por los cuales tienen que ser multiplicados los respectivos valores de los predictores en el modelo lineal.

- *Std. Error*: En segundo lugar, nos encontramos con los errores estándar ε_i, los cuales son independientes para cada predictor y siguen una distribución normal de media 0. Todos tienen la misma varianza. Estos errores sirven para determinar cómo de bueno son sus correspondientes coeficientes calculados por el modelo lineal. Por tanto, cuanto más cercano a 0, mejor serán los coeficientes calculados.
  Sin embargo, el error del β_0 no es relevante, puesto que no depende de ningún predictor en concreto.

- *t-value*: En tercer lugar, encontramos los valores *t*, que son resultado de la división entre el valor calculado del coeficiente del predictor correspondiente y su error estándar. Básicamente, nos dice a cuántos errores estándar el valor estimado del coeficiente está de 0.

- *Pr(>|t|)*: En cuarto lugar, encontramos los *p-valores* de cada predictor. Estos *p-valores* determinan si el predictor correspondiente realmente influye en la variable objetivo a predecir / explicar (en este caso, si un determinado componente químico influye en la prueba). Por tanto, cuanto más cerca de 0 mejor. Para una mayor precisión, a lo largo de este ejercicio se utilizará un nivel de **significancia de 0.01** como valor discriminante para el *p-valor* a la hora de analizar la influencia de los predictores. En realidad, el *p-valor* es la probabilidad de encontrarnos con el valor *t* para ese predictor en una distribución normal con media 0 y desviación típica σ siempre y cuando se considere la hipótesis nula (es decir, β_i = 0, por lo que el predictor no tiene influencia para la variable objetivo). De esta forma, si el *p-valor* es muy bajo, se puede rechazar la hipótesis nula (la hipótesis de que el predictor en cuestión no influye en la variable objetivo), determinando así que el predictor sí influye.
  De la misma manera que con los errores estándar de los predictores, el *p-valor* del β_0 no es relevante.

1.1.2. Estadísticos del modelo

- *Residual standard error*: el error estándar residual es calculado en función de la suma cuadrática de residuos de todos los coeficientes (*RSS*, considerando residuo como la diferencia entre el valor real y el predicho de la variable de un ejemplo), el número de ejemplos y el número de predictores de la siguiente manera:

  $$RSE = \sqrt{\frac{RSS}{(n-p-1)}}$$

  Al fin y al cabo, es una estimación de la desviación típica del error irreducible del modelo (epsilon), por lo que cuanto menor sea este error estándar residual, mejor será el modelo.

- *Multiple R-squared*: el coeficiente de determinación o *R-cuadrado* se calcula en función del *RSS* y de la suma cuadrática total de errores (*TSS*) de la siguiente manera:

  $$R^2 = \frac{TSS - RSS}{TSS}$$

  Por tanto, indica cuánta variabilidad (en porcentaje) puede explicar nuestro modelo lineal, es decir, la proporción de la varianza explicada por el modelo. De esta forma, cuanto más cerca esté este valor del 1, mejor será el modelo, ya que más varianza es explicada por este. No obstante, cabe mencionar que a mayor número de predictores introducidos en el modelo, mayor será siempre el *R-cuadrado*. Por tanto, el *R-cuadrado* sólo nos será útil cuando queramos comparar distintos modelos con el mismo número de predictores.

- *Adjusted R-squared*: debido a que introducir nuevas variables en el modelo siempre aumenta el valor del *R-cuadrado* del modelo (aunque este aumento sea mínimo), hacemos uso del coeficiente de determinación ajustado o *R-cuadrado ajustado*. Este

es calculado en función del *RSS*, del *TSS*, del número de ejemplos y del número de predictores utilizados de la siguiente manera:

$$R^2\ ajustado\ =\ 1\ -\ \frac{RSS\ /\ (n-p-1)}{TSS\ /\ (n-1)}$$

Por tanto, el *R-cuadrado ajustado* se encuentra corregido por el número de predictores, puesto que si el aumento del *R-cuadrado* procedente de introducir un nuevo predictor en el modelo es pequeño, el valor del *R-cuadrado ajustado* decrementa. De esta forma, se penaliza de cierta manera ese predictor sin aparente influencia.

- *F-statistic*: el valor del *F-estadístico* sirve para realizar el test de hipótesis ya que, bajo las condiciones del modelo lineal, y suponiendo que la hipótesis nula es cierta (que ninguno de los predictores influye en la respuesta), este valor sigue una distribución *F.* Este valor, de la misma forma que el *R-cuadrado ajustado*, se calcula en función del *TSS*, el *RSS*, el número de predictores y el número de ejemplos de la siguiente manera:

  $$F\ =\ \frac{(TSS-RSS)\ /\ p}{RSS\ /\ (n-p-1)}$$

  De esta forma, cuanto más cerca esté el *F-estadístico* de 1, más probable es la hipótesis nula, es decir, más probable es que ninguno de los predictores utilizados en el modelo tenga influencia.

- *p-value*: el *p-valor* del *f-estadístico* muestra si existe algún predictor que tenga influencia dentro de la variable objetivo. Para ser consistentes con los *p-valores* de cada predictor vamos a utilizar el mismo nivel de significancia (0.01). De esta forma, por la misma razón que antes, si este valor es muy pequeño (< 0.01) podemos determinar que existe al menos un predictor que influye en la variable objetivo (puesto que este valor es la probabilidad de que se cumpla la hipótesis nula, es decir, de que ninguno de los predictores tenga influencia)

  Se ha determinado un **valor de significancia 0.01** en lugar del comúnmente usado 0.05.
  .

1.2. Prueba DSA

```
Call:
lm(formula = dsa ~ ma, data = adv.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36787 -0.09035  0.00824  0.07952  0.33450

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.36788    0.09640   3.816 0.000485 ***
ma           0.02336    0.00903   2.587 0.013646 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.16 on 38 degrees of freedom
Multiple R-squared:  0.1497,    Adjusted R-squared:  0.1273
F-statistic: 6.691 on 1 and 38 DF,  p-value: 0.01365
```

Figura 1: prueba DSA con componente MA como predictor

En primer lugar, tenemos que fijarnos en el *p-valor* del *F-estadístico*. A pesar de que está por encima del nivel de significancia que tendremos en cuenta para todas las pruebas, podemos realizar cierto análisis del predictor, puesto que justo se encuentra entre los 2 niveles de significancia (*0.01365* $\in$ *[0.01, 0.05]*) utilizados por convención.
Como se puede observar en el *p-valo*r, el componente *MA* no tendría influencia en la prueba *DSA* en caso de utilizarse de manera aislada y de ser consistentes con el nivel de significancia previamente mencionado (0.01). Sin embargo, sí tendría influencia en caso de escoger un nivel de significancia de 0.05. Además, el error estándar es relativamente pequeño (0.00903).

Por otro lado, tenemos que este modelo no es el mejor de aquellos que utilizan un solo predictor. A pesar de que no está presente la imágen en este documento (por cuestiones de espacio), el modelo que utiliza un solo predictor para esta prueba y que tiene mayor R-cuadrado es el que utiliza como único predictor el componente *TPC*, con un *R-cuadrado* de 0.692 (que es mayor que 0.1497, y también mayor que 0.0018, correspondiente al modelo que solo utiliza TSO2 como predictor). En este caso, no hace falta mirar el valor del *R-cuadrado ajustado*, ya que se está analizando los modelos que solo utilizan un componente químico como predictor.

Este modelo tampoco es el que tiene menor error residual, sino que es aquel que usa *TPC* como único predictor, con un error residual estándar de 0.096 (menor que 0.16). Por tanto, podríamos decir que, dentro de los modelos que solo utilizan un predictor para esta prueba, el mejor es aquel que utiliza el componente químico *TPC*, y no *MA* ni *TSO2*.

```
Call:
lm(formula = dsa ~ tpc + ma + tso2, data = adv.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11466 -0.04983 -0.01744  0.03028  0.36833

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0418869  0.0824789  -0.508   0.6147
tpc          0.0034765  0.0004144   8.390 5.42e-10 ***
ma           0.0032586  0.0065274   0.499   0.6207
tso2         0.0036574  0.0020140   1.816   0.0777 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09466 on 36 degrees of freedom
Multiple R-squared:  0.7181,    Adjusted R-squared:  0.6946
F-statistic: 30.56 on 3 and 36 DF,  p-value: 5.3e-10
```

Figura 2: prueba DSA con componentes TPC, MA y TSO2 como predictores

Echando un rápido vistazo al *p-valor* del *F-estadístico*, podemos concluir que al menos uno de los predictores tiene influencia sobre la prueba *DSA*. Gracias a la notación que nos ofrece *R*, encontramos que este componente influyente es *TPC*, con un valor muy cercano a 0 (y, por tanto, muy por debajo de cualquiera de los dos niveles de significancia). A pesar de que los 3 errores estándar correspondientes a los coeficientes estimados son muy bajos, tan solo *TPC* influiría en caso de escoger todos los componentes como predictores. El error residual estándar también es relativamente bajo (0.09466).
En este caso, al ser este modelo el único que integra todos los predictores, no sería necesario fijarse en el valor del *R-cuadrado* del modelo, sino en el valor del *R-cuadrado ajustado* con el objetivo de compararlo con el de aquellos modelos que utilizan un número distinto de predictores. A pesar de que no se encuentran en imagen todos los resúmenes de *R* de todos los modelos, las pruebas indican que aquel modelo con mayor *R-cuadrado ajustado* es el que utiliza tanto *TPC* como *TSO2* como predictores, con un valor de *R-cuadrado ajustado* de 0.7008 (mayor que el modelo que estamos analizando, con un valor de *R-cuadrado ajustado* de 0.6946 a pesar de utilizar todos los predictores). Cabe mencionar que para cada uno de los modelos que utilizan el mismo número de predictores (aquellos que usan 1 ó 2) se ha utilizado el *R-cuadrado* para decidir qué modelo es el mejor entre ellos.

1.3. Prueba ASA

```
Call:
lm(formula = asa ~ tpc + ma, data = adv.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09710 -0.13768  0.07892  0.17627  0.94328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.473614   0.253542   1.868   0.0697 .
tpc         0.005032   0.001513   3.326   0.0020 **
ma          0.001439   0.022261   0.065   0.9488
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3459 on 37 degrees of freedom
Multiple R-squared:  0.2837,    Adjusted R-squared:  0.245
F-statistic: 7.327 on 2 and 37 DF,  p-value: 0.002086
```

Figura 3: prueba ASA con componentes TPC y MA como predictores

Al fijarnos en el *p-valor* del *F-estadístico*, observamos que es muy probable que exista algún predictor de los dos seleccionados para nuestro modelo que influya en la respuesta (0.002086 está por debajo del nivel de significancia considerado). Por tanto, al observar los *p-valores* de cada predictor deducimos que este componente que influye es *TPC*, con un *p-valor* de 0.0020. Observamos también que los errores estándar para los valores de los coeficientes estimados son relativamente pequeños, por lo que en un principio estos valores están bien calculados.
Si hacemos referencia al valor del *R-cuadrado*, observamos que no es muy grande (al menos en comparación con los valores de la prueba anterior), ni es el más grande de los modelos que utilizan dos predictores para esta prueba (0.2837 es mayor que 0.0708, *R-cuadrado* del modelo que utiliza *MA* y *TSO2*, pero es menor que 0.284, *R-cuadrado* del modelo que utiliza *TPC* y *TSO2*). Por tanto, no podemos decir que, dentro de los modelos que utilizan dos predictores para esta prueba, el que utiliza *TPC* y *MA* sea el mejor.
El error residual no es tan pequeño como los errores vistos en los modelos de las pruebas anteriores.

```
Call:
lm(formula = asa ~ tpc + ma + tso2, data = adv.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10193 -0.13994  0.07935  0.17842  0.95427

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4927224  0.3055306   1.613  0.11555
tpc          0.0050250  0.0015350   3.274  0.00235 **
ma           0.0004340  0.0241799   0.018  0.98578
tso2        -0.0008627  0.0074604  -0.116  0.90858
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3506 on 36 degrees of freedom
Multiple R-squared:  0.284,     Adjusted R-squared:  0.2243
F-statistic: 4.759 on 3 and 36 DF,  p-value: 0.006774
```

Figura 4: prueba ASA con componentes TPC, MA y TSO2 como predictor

Al fijarnos de nuevo en el *p-valor* del *F-estadístico* observamos que es bastante probable que alguno de los predictores utilizados en el modelo (que, en este caso, son todos) tengan influencia en la prueba. Por tanto, observando ahora los *p-valores* de los predictores, podemos deducir que el componente *TPC* tiene influencia, pero no los demás componentes, puesto que sus *p-valores* están muy por encima del nivel de significancia considerado. Sin embargo, los errores estándar de los coeficientes calculados son relativamente pequeños, por lo que podemos deducir que están bien estimados. De todas formas, llama la atención el valor del coeficiente del *TSO2*, que implica que…
El error residual estándar es bastante parecido a los modelos anteriores para esta misma prueba.
Por otro lado, el valor del *R-cuadrado ajustado* no es el mayor en relación con los modelos que utilizan un distinto número de predictores. Contraintuitivamente, el modelo con mayor *R-cuadrado ajustado* es aquel que tan solo utiliza el componente *TPC* como predictor, con un valor de 0.2648, que es mayor al 0.2243 procedente de aquel que utiliza todos los predictores, o al 0.2453 que, además de utilizar el *TPC*, también hace uso del *TSO2*. Por tanto, podemos concluir que el modelo que solo utiliza *TPC* es el mejor en cuanto a modelo lineal.

1.4. Prueba ORAC

```
Call:
lm(formula = orac ~ tpc, data = adv.data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2722 -1.5866  0.5834  1.6174  4.9489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.057208   1.530578   0.037   0.9704
tpc         0.020884   0.008837   2.363   0.0233 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.304 on 38 degrees of freedom
Multiple R-squared:  0.1281,    Adjusted R-squared:  0.1052
F-statistic: 5.585 on 1 and 38 DF,  p-value: 0.02333
```

Figura 5: prueba ORAC con componente TPC como predictor

Se observa un *p-valor* del *F-estadístico* superior al nivel de significancia considerado, por lo que no se rechaza la hipótesis nula (es decir, que el predictor considerado en el modelo no influya en la respuesta). Por tanto, si miramos el valor del *p-valor* del predictor, observamos que también se encuentra por encima de este nivel de significancia.
También se puede observar que el valor del *R-cuadrado* es el más alto de los modelos que utilizan un solo predictor para esta prueba (a pesar de que no haya imagen de ello). Sin embargo, sigue siendo relativamente bajo, lo que indica que se puede explicar poca varianza de los datos con el predictor. Asimismo, el *R-cuadrado* ajustado también es muy bajo.
Además, el error residual es hasta ahora el más grande de las pruebas realizadas.
De esta manera, podemos concluir que el componente *TCP* no es un buen predictor para nuestro nivel de significancia, al menos cuando se utiliza de manera aislada.

```
Call:
lm(formula = orac ~ tpc + tso2, data = adv.data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1111 -1.7350  0.5597  1.4095  4.9622

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.369413   1.723158  -0.214   0.8314
tpc          0.022080   0.009171   2.407   0.0212 *
tso2         0.025781   0.046158   0.559   0.5798
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.325 on 37 degrees of freedom
Multiple R-squared:  0.1354,    Adjusted R-squared:  0.0887
F-statistic: 2.898 on 2 and 37 DF,  p-value: 0.06773
```

Figura 6: prueba ORAC con componentes TPC y TSO2 como predictores

Se observa un *p-valor* del *F-estadístico* superior al nivel de significancia que se ha indicado, por lo que podría descartarse de forma prácticamente automática que haya algún predictor que tenga influencia en la prueba realizada.
Tanto el *R-cuadrado* como el *R-cuadrado ajustado* son muy pequeños, con lo que escoger como predictores los componentes *TPC* y *TSO2* no sirven para explicar el modelo con precisión. El error residual estándar, de la misma manera que en el modelo anterior, es también muy elevado.
Por tanto podemos concluir que, independientemente del nivel de significancia considerado, estos dos predictores no serían buenos para explicar el modelo lineal en cuestión.

1.5. Prueba RP

```
Call:
lm(formula = rp ~ tpc + tso2, data = adv.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23257 -0.05460 -0.01241  0.06315  0.32396

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2259917  0.0890036   2.539   0.0154 *
tpc          0.0030284  0.0004737   6.393 1.85e-07 ***
tso2        -0.0051425  0.0023841  -2.157   0.0376 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1201 on 37 degrees of freedom
Multiple R-squared:  0.5976,    Adjusted R-squared:  0.5759
F-statistic: 27.48 on 2 and 37 DF,  p-value: 4.846e-08
```

Figura 7: prueba RP con componentes TPC y TSO2 como predictores

En este caso se han considerado los componentes *TPC* y *TSO2* para observar la relación que tienen con la prueba *RP*.
Se observa un *p-valor* del *F-estadístico* muy cercano a 0, por lo que se considera que alguno de los predictores tiene influencia sobre los datos. Además, el *F-estadístico* nos indica un valor que no es muy cercano a 1, por lo que es probable que se rechace la hipótesis nula, tal y como nos ha indicado el *p-valor*.
Se observa además que el error estimado de los datos es diminuto , por lo que podemos asegurar que nuestras predicciones no se desviarán mucho del valor real.
Por otro lado, dentro de los modelos que tan solo utilizan dos predictores, este es el que obtiene un mayor *R-cuadrado*, siendo el que más variabilidad de los datos es capaz de explicar con dos predictores. Esto se debe a que el componente *TPC* cuenta con un *p-valor* que hace rechazar la hipótesis nula para este predictor, aunque no suceda lo mismo o¡con el predictor *TSO2* (aunque sí podría rechazarse la hipótesis nula para este último predictor en caso de escoger un nivel de significancia de 0.05). Además, el error residual estándar es muy pequeño.

```
Call:
lm(formula = rp ~ tpc + ma + tso2, data = adv.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23739 -0.05559 -0.00693  0.06502  0.32773

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2557022  0.1056551   2.420   0.0207 *
tpc          0.0031515  0.0005308   5.937 8.43e-07 ***
ma          -0.0044724  0.0083616  -0.535   0.5960
tso2        -0.0056384  0.0025799  -2.186   0.0354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1213 on 36 degrees of freedom
Multiple R-squared:  0.6008,    Adjusted R-squared:  0.5675
F-statistic: 18.06 on 3 and 36 DF,  p-value: 2.552e-07
```

Figura 8: prueba RP con componentes TPC, MA y TSO2 como predictores

Si nos fijamos en el *p-valor* del *F-estadístico*, podemos concluir que al menos uno de los predictores tiene influencia sobre la prueba *RP*. Encontramos que este componente se trata de *TPC*, con un valor muy cercano a 0.
Por otro lado, al fijarse en el valor del *R-cuadrado ajustado* con el objetivo de compararlo con el de aquellos modelos que utilizan un número distinto de predictores, las pruebas indican que aquel modelo con mayor *R-cuadrado ajustado* es el que utiliza tanto *TPC* como *TSO2* como predictores (explicado en la figura 7), con un valor de *R-cuadrado ajustado* de 0.5759, mayor al presente en esta figura.

1.6. Prueba MCA

```
Call:
lm(formula = mca ~ ma + tso2, data = adv.data)

Residuals:
    Min      1Q  Median      3Q     Max
-26.604 -12.741  -5.995   3.002  50.117

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.92300   16.31026   0.302    0.764
ma           2.43569    1.34366   1.813    0.078 .
tso2        -0.03477    0.45964  -0.076    0.940
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.62 on 37 degrees of freedom
Multiple R-squared:  0.1004,    Adjusted R-squared:  0.0518
F-statistic: 2.065 on 2 and 37 DF,  p-value: 0.1411
```

Figura 9: prueba MCA con componentes MA y TSO2 como predictores

Al fijarnos en el *p-valor* del *F-estadístico*, se puede observar que está muy por encima del umbral de aceptación (0.1411 > 0.01). Esto sugiere que podría haber suficiente evidencia para aceptar la hipótesis nula y, por lo tanto, no se puede afirmar que ninguno de los predictores tenga un efecto significativo en el resultado. Observamos también que los errores estándar para los valores de los coeficientes estimados son relativamente altos, por lo que en un principio estos valores están bien calculados.
Si hacemos referencia al valor del *R-cuadrado*, observamos que es el más pequeño de los modelos que utilizan dos predictores para esta prueba (0.1004 es menor que 0.3872, *R-cuadrado* del modelo que utiliza *MA* y *TPC*, y que 0.3867, *R-cuadrado* del modelo que utiliza *TPC* y *TSO2*). Por tanto, podemos decir que, dentro de los modelos que utilizan dos predictores para esta prueba, el que utiliza *MA* y *TPC* es el mejor.
No obstante, se puede observar de un rápido vistazo que el modelo presente en el resumen de esta figura es bastante impreciso a la hora de explicar la respuesta, no solo por los valores tanto de *R-cuadrado* como de *R-cuadrado ajustado* previamente comentados, sino también por el error residual estándar, que es bastante alto, así como los errores estándar de los coeficientes estimados.

```
Call:
lm(formula = mca ~ tpc + ma + tso2, data = adv.data)

Residuals:
    Min      1Q  Median      3Q     Max
-42.011 -12.269  -0.798  10.558  45.058

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.46215   15.76295  -1.742 0.090013 .
tpc           0.32509    0.07919   4.105 0.000222 ***
ma            0.21620    1.24749   0.173 0.863379
tso2          0.02938    0.38490   0.076 0.939579
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.09 on 36 degrees of freedom
Multiple R-squared:  0.3873,    Adjusted R-squared:  0.3362
F-statistic: 7.584 on 3 and 36 DF,  p-value: 0.0004689
```

Figura 10: prueba MCA con componentes TPC, MA y TSO2 como predictores

En este caso, observamos rápidamente tras echar un vistazo al *p-valor* del *F-estadístico* y de los predictores que, cuando cogemos todos los predictores para la prueba *MCA*, es el componente *TPC* el único que tiene influencia en ella. Sin embargo, no cuenta con el mayor valor de *R-cuadrado ajustado* de todos los valores para esta prueba, puesto que este lo tiene aquel que solo utiliza el propio componente *TPC*. Esto se debe a que la adición de predictores irrelevantes se transforma en cierta forma de ruido a la hora de explicar e influir en la prueba *MCA*. Además, el valor estándar residual es alto, y ligeramente mayor que aquel que solo usa *TCP* como predictor (18.09 frente a 17.62).
Por otro lado, ninguno de los errores estándar de los coeficientes calculados destaca por ser muy pequeño (de hecho, el de *MA* es considerablemente grande).

2. Un problema de clasificación

2.1. Obtención de los puntos más al norte y más al sur

La playa se encuentra modelada, tal y como dice el enunciado, mediante un segmento que une el punto que se encuentra más al norte de los almacenados en el conjunto de datos con el que se encuentra más al sur. Para calcular dichos puntos, se han extraído los puntos con una latitud (columna “latitud.y” en el dataset) mayor, para el punto más al norte, y menor, para el punto más al sur.
En R, estos puntos pueden ser extraídos de manera sencilla mediante las funciones *which.max* y *which.min*, que devuelven los valores máximos y mínimos de una columna especificada (en este caso, aquella con el predictor “latitud.y”).
Por ejemplo, se puede obtener el punto más al norte simplemente con la siguiente línea de código:

```
max_latitude_row <- data[which.max(data$latitud.y), c("longitud.x",
                         "latitud.y")]
```

Una vez se obtienen los dos puntos, se puede generar una línea que los une, representando la playa.

2.2. Obtención de los puntos que separan las zonas norte y centro, y centro y sur

Para obtener los dos puntos solicitados, ha sido necesario crear dos subconjuntos del dataset original.

- *Subconjunto 1*: aquellos fardos que han acabado en la cala 0 junto con aquellos que han acabado en la cala 1 (zona norte y zona centro).
- *Subconjunto 2*: aquellos fardos que han acabado en la cala 1 junto con aquellos que han acabado en la cala 2 (zona centro y zona sur).

Una vez se han obtenido los correspondientes subconjuntos de datos, se han realizado dos regresiones logísticas diferentes, una por cada par de zonas a ser divididas. Ambas regresiones logísticas utilizan tanto la longitud como la latitud de los puntos como predictores. Cada modelo genera una recta que divide, de la mejor forma posible, los datos pertenecientes a una clase frente a los de la otra, convirtiéndose así en dos problemas de regresión logística binomial. Por tanto, se busca la mejor función logística que se adapte a los datos.

Ahora, para encontrar los puntos (y no solo las rectas) que separan las respectivas zonas, necesitamos hallar los puntos de corte entre las líneas procedentes de los modelos de regresión logística y el segmento que modela la playa. Por tanto, el primer paso es encontrar la fórmula de la recta que pasa por los dos puntos que modelan la línea de playa (el que está más al norte y el que está más al sur) que, hasta ahora, es un segmento.
Para encontrar la recta de la forma $y = a + b * x$ dados dos puntos *A(x1, y1)* y *B(x2, y2)*, se aplica un poco de geometría básica mediante la siguiente ecuación:

$$\frac{x - x1}{x2 - x1} = \frac{y - y1}{y2 - y1}$$

Tras sustituir las correspondientes coordenadas de los puntos *A* y *B* y dejar la ecuación de la forma $y = a + b * x$, ya tenemos la recta que pasa por dichos puntos. Hemos pasado de un segmento, dado por dos puntos, a una recta, que pasa por infinitos puntos. Este procedimiento matemático es llevado a cabo por la función *find_line_coefficients()*, la cual acepta como parámetros los dos puntos, con sus respectivas coordenadas, por los cuales pasa la recta que queremos calcular.

A continuación, debemos deducir que los puntos solicitados serán los puntos de corte de la recta que modela la playa (que antes era un segmento) y cada una de las rectas dadas por los modelos de regresión logística. Debido a que las dos rectas se encuentran ahora de la misma forma (y = a + b * x), tenemos los coeficientes a y b de cada una de las rectas, necesarios para encontrar los puntos de corte.

La función *find_intersection_point()* del código acepta como parámetros los coeficientes *a* y *b* de las dos rectas (*a_recta1, b_recta1, a_recta2, b_recta2*), y halla el punto de corte entre ellas.

Por tanto, el punto que separa la zona norte y centro será el punto de corte entre la recta que modela la playa y la recta obtenida por el modelo de regresión logística que separa la zona norte de la zona centro. Concretamente, este punto es (*9.558354, 1.100006*).

De la misma manera, el punto que separa la zona centro y la zona sur será el punto de corte entre la recta que modela la playa y la recta obtenida por el modelo de regresión logística que separa la zona centro de la zona sur. Concretamente, este punto es (*9.564314, 1.089023*).

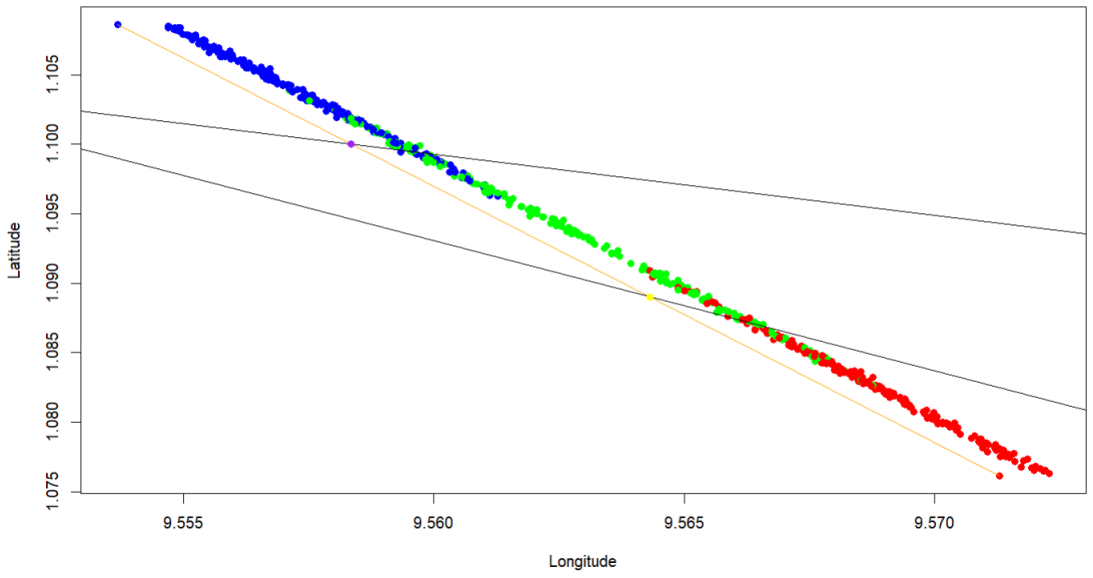

Figura 11: gráfico con toda la información relevante del ejercicio 2

En el gráfico se observan, en negro, las rectas de los modelos de regresión logística. La superior es la correspondiente al modelo que divide los fardos de la zona norte y la zona centro, y la inferior es la correspondiente al modelo que divide los fardos de la zona centro y la zona sur.

Por otro lado, en naranja se aprecia el segmento que modela la playa, uniendo el punto más al sur de la misma con el punto más al norte.

Por último, en morado se identifica el punto que separa la zona norte de la zona centro, y en amarillo, el punto que separa la zona centro de la zona sur.

2.3. Funciones de predicción

2.3.1. Función 1

La función 1, llamada en el código *predict_cala()*, devuelve la cala a la que es más probable que llegue un fardo dadas las coordenadas del punto en el que fue tirado. Al contar con dos modelos de regresión logística, primero debemos saber cuál de los dos utilizar en base a las coordenadas proporcionadas como entrada. El modelo a utilizar será aquel cuyo punto de corte con la línea que modela la playa (los puntos obtenidos en el *apartado 2.2*) se encuentra más cerca (según la distancia euclídea) del punto pasado como parámetro. Este modelo, por tanto, devolverá la cala a la que es más probable que llegue dicho fardo.

2.3.2. Función 2

La función 2, llamada en el código *predict_calas_vector()*, devuelve las probabilidades, en forma de vector, de que el fardo llegue a cada una de las calas dadas las coordenadas del punto en el que fue tirado. Se ha decidido proceder de la misma forma que en la función anterior: para determinar las probabilidades, se escoge el modelo correspondiente al par de zonas cuyo punto que las separa se encuentra más cerca del punto pasado como parámetro. De esta forma si, por ejemplo, el punto se encuentra más cerca del punto que separa las zonas centro y sur que del punto que separa las zonas centro y norte, se utilizará el modelo que separa las zonas centro y sur. Este dará las probabilidades de que el fardo acabe en la cala 0 (sur) y en la cala 1 (centro), y dará una probabilidad de 0 a que el fardo acabe en la cala 2 (norte).

2.3.3. Función 3

Esta función, llamada *predict_cala_dataset()*, es muy similar a la función *predict_cala()* (*punto 2.3.1*) pero, en vez de devolver la cala más probable para un solo punto, devuelve la cala más probable para todos los puntos del dataset. Esta función será necesaria para obtener la matriz de confusión de los datos de entrenamiento del punto *2.4*.

2.4. Matriz de confusión

Para la matriz de confusión se han utilizado todos los puntos con sus etiquetas (clases), así como la función de predicción 1, que devuelve la cala a la que es más probable que llegue un fardo dadas las coordenadas en las que fue lanzado. Pasándole como parámetros cada uno de los puntos a esta función, se ha obtenido la siguiente matriz de confusión:

predictions	0	1	2
0	179	22	0
1	21	144	28
2	0	34	172

Figura 12: matriz de confusión con todos los fardos del ejercicio 2

El hecho de que los valores de las celdas *(0, 2)* y *(2, 0)* valgan 0 implica que no ha habido ningún fardo que, perteneciendo a la cala 0, haya sido clasificado como perteneciente a la cala 2, y viceversa. Esto tiene sentido, ya que ambas calas se encuentran opuestas en la playa, separadas por la cala 1. Sin embargo, sigue habiendo bastantes fardos en las zonas intermedias mal clasificados por nuestro modelo. No obstante, la diagonal principal de la matriz contiene la mayoría de los datos, por lo que la mayoría de los fardos han sido bien clasificados por nuestro modelo.

2.5. Predicciones de los fardos solicitados

```
> print(paste("Cala más probable en la que acabe el fardo 1:", fardo1_cala))
[1] "Cala más probable en la que acabe el fardo 1: 2"
> print(paste("Probabilidades de que el fardo 1 acabe en cada cala: Cala 0:", fardo1_calas_vector[1],
+           "Cala 1:", fardo1_calas_vector[2], "Cala 2:", fardo1_calas_vector[3]))
[1] "Probabilidades de que el fardo 1 acabe en cada cala: Cala 0: 0 Cala 1: 0.499381994581967 Cala 2: 0.500618005418033"
>
> print(paste("Cala más probable en la que acabe el fardo 2:", fardo2_cala))
[1] "Cala más probable en la que acabe el fardo 2: 0"
> print(paste("Probabilidades de que el fardo 2 acabe en cada cala: Cala 0:", fardo2_calas_vector[1],
+           "Cala 1:", fardo2_calas_vector[2], "Cala 2:", fardo2_calas_vector[3]))
[1] "Probabilidades de que el fardo 2 acabe en cada cala: Cala 0: 0.503058544439607 Cala 1: 0.496941455560393 Cala 2: 0"
>
> print(paste("Cala más probable en la que acabe el fardo 3:", fardo3_cala))
[1] "Cala más probable en la que acabe el fardo 3: 0"
> print(paste("Probabilidades de que el fardo 3 acabe en cada cala: Cala 0:", fardo3_calas_vector[1],
+           "Cala 1:", fardo3_calas_vector[2], "Cala 2:", fardo3_calas_vector[3]))
[1] "Probabilidades de que el fardo 3 acabe en cada cala: Cala 0: 0.994127090881929 Cala 1: 0.0058729091180709 Cala 2: 0"
```

Respecto a la última pregunta del ejercicio, los protagonistas de esta historia podrían vivir perfectamente en España, concretamente en los municipios de la Línea de la Concepción o Algeciras, ambos en Cádiz.

3. La liga nacional de Kárate

3.1. Manipulación de los datos

Se ha llevado a cabo una serie de pasos mediante manipulación de datos para que, al final de esta tarea, el *dataframe* solicitado quede lo más organizado y completo posible.
En primer lugar, se extraen todas las hojas de cada libro Excel facilitado por el enunciado

A continuación, se ha creado una función que, dado un nombre de hoja *Excel* (compuesto por nombre y año), es capaz de extraer los valores de las siguientes columnas:

- *Jornada*: se ha hecho uso de una expresión regular, que obtiene los dígitos posteriores al carácter “J”.
- *Modalidad*: se extrae el último carácter del nombre del fichero
- *Pasa*: se comprueba si en la columna *estadoSigRonda* figura la cadena “pasa” o no.

Una vez extraídos los datos mencionados anteriormente, se crea en el *dataframe* una nueva columna para cada uno de estos campos. Esta secuencia de acciones se repite para todas las hojas, iterando sobre ellas y concatenado su contenido al *dataframe* mediante la función *rbind()*.

3.2. Realización del análisis

El objetivo puede simplificarse en obtener la media del primer y último grupo junto con la proporción de clasificados para saber si el orden influye en la clasificación. Cabe mencionar que, debido a las razones que indica en el enunciado, se ha restringido el estudio a los datos de la ronda 1. Se ha utilizado también $k = 3$ para determinar los participantes considerados como “primeros”.
Los resultados obtenidos del estudio de dicha ronda han sido los siguientes:

- Puntuación media de los participantes del primer grupo: 21,72
- Puntuación media de los participantes del último grupo: 22,72
- Proporción de clasificados de los que pasan del primer grupo: 37,63%
- Proporción de clasificados de los que pasan del último grupo: 53,37%

Observando los resultados, se puede ver que la media es ligeramente peor en los del primer grupo respecto a los del último grupo. Esta diferencia supone un 4% de un grupo respecto al otro.
Por otro lado, si observamos la proporción de clasificados se puede observar que los que participan primero tienen una probabilidad del 37% de pasar de ronda, lo que implica que un 63% de los que participan en el primer grupo no pasan de ronda. Sin embargo, los que participan en el último grupo tienen una probabilidad más alta de clasificarse, 53%. De modo que los del último grupo sólo un 47% no pasan de ronda.

En conclusión, sí se puede decir que el orden influye en la clasificación, puesto que los datos demuestran que los participantes que compiten primero tienen menos probabilidades de pasar de ronda, a pesar de que la nota no sea muy diferente respecto a las de aquellos que compiten saliendo últimos.

3.3. Pregunta realizada por el estudiantado

Se ha optado por formular la siguiente pregunta de interés:

¿Son los jueces más compasivos dependiendo de la categoria y el sexo?

En otras palabras…

¿Los jueces son imparciales dada la edad de los participantes y el género, o están sesgados para dar mejores puntuaciones a los más pequeños (benjamines) o a un género en concreto?

Por tanto, en términos estadísticos, se puede reformular la pregunta anterior de la siguiente manera:

¿La puntuación media de todos los participantes es similar, independientemente del género y la categoría?

Cabe destacar que el kárate en modalidad kata consta de una serie de movimientos y técnicas, sin llegar al combate físico. Este apunte es importante porque indica que, en principio, las puntuaciones de los participantes masculinos y femeninos no deberían ser muy distintas entre sí. Si esta modalidad del karate tuviera una fuerte componente, por ejemplo, de flexibilidad, las mujeres podrían obtener mejores resultados (como sucede en gimnasia). Lo mismo sucedería si tuviera una componente de fuerza, donde los hombres se verían más beneficiados.

Para resolver esta cuestión se ha planteado obtener la media de la puntuación total de todos los participantes de cada categoría, separadas por el género.
Los resultados obtenidos han sido los siguientes:

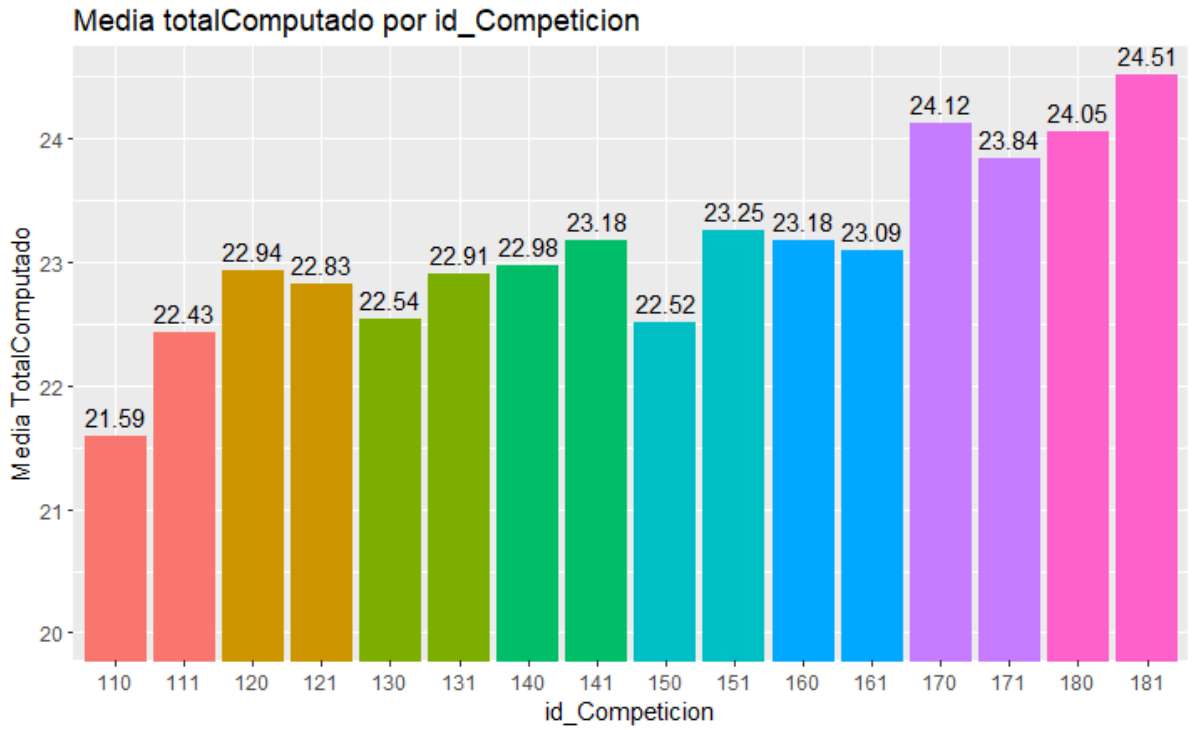

Figura 13: puntuación media de cada categoría

Leyenda:
Categoría: <Número de dos digitos><0(m) o 1(f)>
11->Benjamin, 12->Alevín, 13->Infantil, 14->Juvenil, 15->Cadete,
16->Junior, 17->Sub21, 18->Senior

La gráfica muestra la media total de cada competición (*id_Competicion*). Por ejemplo, la primera barra permite visualizar la media de los participantes masculinos en la categoría Benjamin, mientras que la segunda es la misma categoría, pero con participantes femeninas.

Podemos ver cómo hay una ligera diferencia en la media de las puntuaciones siendo la modalidad femenina superior en casi un punto respecto a los masculinos. Al ser tan poca la diferencia no podemos decir que está sesgado hacia un género u otro ya que, además, esto no ocurre en todas las categorías (las categorías “Alevín”, “Junior” y “Sub21”, un total de 3 sobre 8, cuenta con mejores calificaciones para los hombres).

Lo que sí que se observa es que,por regla general, a medida que se aumenta de categoría, las medias de las puntuaciones suelen aumentar. Esto podría implicar una justa imparcialidad de los jueces ya que, cuanto más mayor es un participante, más práctica y mejores habilidades debería tener (aunque esto no tiene por qué ser así), lo que suele resultar en unos mejores resultados. Además, como se ha mencionado anteriormente, no se observa ninguna preferencia por algún género en ninguna de las categorías, ya que en ninguna la diferencia de las medias en la misma categoría para los distintos géneros es superior a 1.

En conclusión, se puede determinar que los jueces no son más compasivos dependiendo de la edad. Tampoco se encuentran sesgados hacía ningún género, siendo, *a priori*, bastante imparciales en estos aspectos.

4. Conclusiones

Con el primer problema hemos aprendido a utilizar y comparar modelos de regresión lineal sobre datos reales, para hallar conclusiones relevantes que puedan servir a la hora de realizar informes.
Con el segundo problema hemos aprendido a combinar varios modelos de regresión logística para resolver un problema planteado.
Con el último problema hemos comprendido la importancia de la media de los datos en la obtención de conclusiones significativas. Esto se puede observar cuando se reformulan las preguntas de manera más simple para obtener las mismas conclusiones pero aplicando métodos más sencillos.

Tarea	**Jan**	**Simón**	**Óscar**
Problema 1	3		
Problema 2	5		
Problema 3	3		
Realización de la memoria	4		
Total: 45 horas	15		