Práctica 4: Métodos Lineales de Regresión y Clasificación

Almacenes y Minería de Datos

Pedro Allué Tamargo (758267) Cristina Oriol García (755922)
Alejandro Paricio García (761783)

16 de diciembre de 2020

Índice

1. Introducción

En esta práctica se van a utilizar los conceptos de regresiones lineales y logísticas vistos en clase. Estos conceptos son muy útiles de cara al cálculo de predicciones. En el caso de las regresiones lineales se van a utilizar para estudiar qué compuestos son importantes en los análisis de los antioxidantes para 5 variables de las cervezas. En el caso de las regresiones logísticas se utilizarán para estudiar el comportamiento de unos fardos de droga que son tirados al mar por los narcotraficantes y van a parar a 3 calas distintas dependiendo de la posición desde la que se lanzan.

2. Problema 1

2.1. Análisis de los ensayos

2.1.1. Análisis variable *DSA*

Se ha creado un modelo de regresión lineal utilizando las variables *tpc*, *tso2* y *ma*. Se puede observar que en la salida obtenida para este modelo (Figura 1) el *p-valor* del *F estadístico* tiene un valor por debajo de *0.05* y por lo tanto implica que $\exists\hat{\beta}_i \neq 0$. Esto implica que alguno de los estimadores β es influyente en el modelo. Utilizando las 3 variables se obtiene un *adjusted* R^2 que explica el 69 % de la varianza.

Se va a proceder a realizar otro modelo de regresión lineal descartando la variable *ma*. Se ha descartado esta variable porque su *p-valor* es muy alto con respecto al de los demás predictores. La salida obtenida con este modelo se puede observar en la Figura 2. Se puede observar que el *p-valor* del *F estadístico* sigue siendo inferior al umbral (*0.05*) y por lo tanto existe algún predictor que es influyente en el modelo. Tambíen se puede observar que el *adjusted* R^2 ha aumentado (*60 %*). Ha subido al eliminar el preditor *ma* debido a la penalización asociada de aumentar el número de predictores no influyentes.

Tras concluir que al menos un elemento es influyente en el modelo se va a analizar con únicamente el predictor *tpc*. Los resultados de este modelo se pueden encontrar en la Figura 3. Se puede observar que el *adjusted* R^2 baja respecto al análisis anterior a explicar solamente el 68 % de la varianza. Esto implica que la variable *tso2* era influyente en cierta medida pero menos que la variable *tpc*. El *p-valor* sigue siendo menor que 0,05 y por lo tanto implica que la variable *tpc* es influyente ($\hat{\beta_{tpc}} \neq 0$).

Para concluir, se va a proceder a crear un modelo de regresión lineal utilizando la variable *tso2*. Los resultados se pueden observar en la Figura 4. Se puede observar que el *p-valor* tiene un valor superior a 0,05 y por lo tanto se acepta la hipótesis nula que dice que $\hat{\beta_{tso2}} = 0$. La conclusión de este análisis es que la variable *tso2* influye muy poco sobre el modelo con respecto a la variable *tpc*.

Partiendo de los modelos anteriores se ha llegado a la conclusión de que los predictores *ma* y *tso2* no son suficientemente influyentes, por lo que la variable *dsa* quedaría explicada únicamente con *tpc*. El descarte de *ma* es claro, no obstante, el de *tso2* no lo es tanto. Primero se observó que influía en cierta medida, ya que al incluirlo se pasaba de explicar el 68 % al 70 % de la variabilidad. Pese a ello, al observar cómo explicaba por si mismo *tso2* a *dsa*, el análisis de los resultados daba a entender que no era influyente. El balance entre ambas posibilidades llevó a la eliminación del predictor anterior en este tipo de examen de propiedades antioxidantes.

2.1.2. Análisis variable RP

Se ha realizado un primer análisis sobre la variable *RP* utilizando las 3 variables (Figura 5). Se puede observar que el *p-valor* de la variable *ma* es superior a los anteriores y eso puede significar que no sea influyente en el análisis. El modelo tiene un *p-valor* en el F estadístico muy bajo (inferior a *0.01*) y por lo tanto implica que $\exists\hat{\beta}_i \neq 0$, es decir, alguno de los estimadores parece tener influencia en el modelo. El *adjusted* R^2 presenta un valor de *56 %*. Esto implica que el modelo con los predictores actuales explica un 56 % de la varianza.

Dado que la variable *ma* ha mostrado un *p-valor* alto se va a proceder a realizar otro modelo de regresión lineal sin utilizar esta variable. En la Figura 6 se puede observar el resultado de este modelo. Se puede observar que el modelo sigue presentando un *p-valor* en el *F estadístico* muy pequeño, por lo tanto implica que $\exists\hat{\beta}_i \neq 0$, es decir, alguno de los estimadores tiene influencia en el modelo. Se puede observar que el *adjusted* R^2 presenta un valor mínimamente superior al obtenido en el modelo de 3 variables (Figura 5). Esto se debe a que al eliminar una variable que no influía sobre el modelo se reduce la penalización del número de variables en el modelo.

Puesto que la variable *tso2* presenta un *p-valor* superior al presentado por la variable *tpc* pero inferior al umbral *0.05* se va a proceder a realizar un modelo con el cual solo se tenga en cuenta la variable *tpc*. El resultado de este modelo se puede observar en la Figura 7. Se puede observar que el *p-valor* del *F estadístico* tiene un valor muy próximo a *0* y menor que el umbral *0.01* y por lo tanto existe un estimador que es influyente en el modelo. En este caso, al solo contar con la variable *tpc* se puede llegar a la conclusión de que esta variable es influyente en el modelo. También se puede observar que el valor de *adjusted* R^2 se ha reducido ligeramente comparado con el modelo de la

Figura 6 ya que ha pasado del valor *57 %* a un *53 %*. Esto implica que la variable *tso2* era influyente en el modelo aunque no tanto como *tpc*.

Para comprobar la hipótesis de que la variable *tso2* es influyente en el modelo se va a realizar un modelo utilizando solo esa variable para intentar predecir el valor de la variable *RP*. Los resultados de este modelo se pueden observar en la Figura 8. Se puede observar que el *p-valor* del *F estadístico* es menor que un umbral de *0.05* y próximo al umbral de *0.01* y por lo tanto la variable *tso2* es influyente en el modelo. También se puede observar que el *adjusted* R^2 indica que esta variable es capaz de explicar un *13 %* de la varianza.

Por lo tanto, se puede concluir que las variables *tpc* y *tso2* son influyentes en el modelo de regresión lineal para el antioxidante *RP*. La variable *tso2* no es tan influyente en el mismo como la variable *tpc*.

2.1.3. Análisis variable ORAC

El primer paso es llevar a cabo la regresión lineal sobre la varrable ORAC utilizando *tsp*, *tso2* y *ma*. La figura 9 muestra los resultados. Se puede observar que el R^2 adquiere un valor muy bajo, explicando únicamente el 20 % de la varianza. Además, el p–valor del F–estadístico tiene un valor de 0.042, bastante superior a 0.01, lo que nos indica que es muy probable que no haya ningún predictor realmente influyente en el modelo. Si estudiamos los p–valores de cada uno de los predictores podemos descartar directamente *tso2* y *ma*. Debido a ello, se va a analizar el modelo con únicamente la variable tpc, los resultados se pueden observar en la figura 10.

Vuelven a reportarse p–valores cuestionables para *tpc*. El p–valor del F–estadístico no es suficientemente bajo como para determinar que con alta probabilidad alguno de los predictores es influyente, y la varianza explicada ronda el 12 %. Se puede concluir que, pese a que *tpc* es la que más contribuye a la explicación de *orac* de entre las tres, no se puede garantizar que sea suficientemente influyente como para llegar a ser relevante sino que, más bien contrario, su p–valor parece indicar que no es lo suficientemente bueno. Es el único predictor restante, eliminarlo daría lugar a tener en cuenta únicamente el interceptor, que tomaría como valor la media de los datos. Los resultados de esto se muestran en la figura 11. Estos confirman la hipótesis anterior de que no era lo suficientemente relevante y apoya la idea extraída a partir del primer p–valor del F–estadístico, que indicaba que era muy probable que ninguno de los predictores fuera distinto de cero.

Se concluye que *tpc* influye muy poco en el resultado y que en ningún caso se explica suficientemente bien la varianza de los datos. Esto nos indica que éstos no son predictores del todo apropiados del test ORAC, lo que puede abrir la pregunta acerca de la corrección del estudio del problema, puesto que podrían estar pasándose por alto otros factores que si que sean influyentes en el anterior y que deberían considerarse.

2.1.4. Análisis variable ASA

Primero llevamos a cabo una regresión lineal utilizando las tres variables *tpc, tso2, ma*. Como se muestra en la figura 12 podemos ver que el R^2 tiene un valor bastante lo que significa que solo explica un 22 %de la varianza. Observamos que el p–valor de F–estadístico tiene un valor de 0.0067 por lo que podemos suponer que alguna de los predictores influye en el modelo. Analizando los valores de estos predictores podemos observar que el p–valor de *ma, tso2* es excesivamente alto por lo que podemos descartarlos de este modelo.

Por otro lado el p–valor de *tpc* es lo suficientemente bajo como para concluir que es influyente en nuestro modelo. Procedemos a hacer un análisis únicamente sobre la variable *tpc* por los datos comentados anteriormente. El resultado lo observamos en la figura 13. Como se puede observar según el R^2 se explica un 26 % de la varianza, subiendo respecto al modelo anterior únicamente un 4 %. Tanto el p–valor del F–estadístico como el p–valor del predictor *tpc* mantienen que este predictor influye en nuestro modelo. Aunque sigue sin explicar correctamente la varianza de los datos.

Para comprobar las hipótesis sobre los predictores realizamos un último análisis sobre las variables previamente descartadas, podemos ver el resultado en la figura 14 . De este análisis comprobamos que el F–estadístico es 1,41 suficientemente cercano a 1 y su p–valor 0.257 suficientemente grande para concluir que efectivamente los dos predictores utilizados no influyen en nuestro modelo, aceptando la hipótesis nula habiendo optado correctamente por

su descarte.

De estos análisis podemos concluir que únicamente *tpc* influye en los resultados obtenidos aunque no explica suficientemente la varianza de los datos.

2.1.5. Análisis variable MCA

Como podemos observar en la figura 15 el p–valor del F–estadístico nos indica que al menos uno de los predictores influye en el modelo, con la información proporcionada por los p–valor de cada predictor podemos descartar los predictores *ma, tso2* por tener un valor excesivamente alto, por lo que el único con un valor aceptable sería *tpc*. Realizamos un análisis con únicamente este predictor, como se ve reflejado en la figura **??** los p–valores continúan siendo suficientemente pequeños como para determinar que este predictor efectivamente influye en nuestra variable y como nos indica el R^2 explica un 37 % de la varianza de esta variable.

2.2. Conclusiones del análisis de antioxidantes

Tras la realización de los análisis de los antioxidantes estudiados anteriormente se puede observar que el compuesto *tpc* es muy influyente en los cinco ensayos. Otro compuesto influyente es *tso2* aunque en menor medida que el *tpc*. El compuesto *ma* no es influyente en ninguno de los ensayos.

No obstante, en los ensayos *ORAC*, *ASA*, *MCA* se puede observar que *tpc* da parte de la explicación de la varianza pero no toda y por lo tanto puede existir, o no, algún compuesto más que no se esté teniendo en cuenta en este conjunto de datos.

3. Problema 2

3.1. Parte 1: ¿A dónde va el fardo?

Para la realización de este ejercicio se ha utilizado la interpretación geométrica de las regresiones logísticas. Si se interpreta de que la playa es una línea recta que se describe entre los puntos de mayor y menor latitud entonces los puntos que dividen las zonas de las 3 calas se corresponden con las intersecciones de las rectas creadas por las regresiones logísticas con la recta de la playa. Una representación de esto se puede observar en la Figura 17.
Se puede observar que la recta que describe la playa se corresponde con la recta $y = -1{,}842551x + 18{,}711764$.

Como nos encontramos ante una recta el uso de los dos parámetros no es necesario ya que al conocer la ecuación de esta teniendo uno de los dos podríamos calcular el otro así que es posible prescindir de uno de los dos, no importa cual para el cálculo del modelo. Si se opta por tener en cuenta los dos se obtiene el modelo que se puede observar en la figura: 16 como se muestra los p–valores de los dos predictores son excesivamente altos por la relación entre ellos previamente comentada, por eso se ha decidido utilizar exclusivamente el parámetro *latitud.y*

Para hallar estas dos rectas se han creado dos regresiones logísticas con el objetivo de hallar qué fardos terminarán en la cala 0 y qué fardos terminarán en la cala 2. Por lo tanto dado un punto si estas regresiones predicen una probabilidad cercana a 0 de que el fardo termine en la cala 0 y la cala 2 significará que el fardo terminará en la cala 1.

Las rectas correspondientes a los modelos de regresión logística son:

- Regresión de *Cala 0*: $y = 0{,}001563063x + 1{,}087117$
- Regresión de *Cala 2*: $y = -0{,}001635438x + 1{,}099433$

El código correspondiente a la creación de los modelos se puede encontrar en el *Listing 1.*

Los puntos que dividen la playa en las 3 calas se pueden averiguar calculando la intersección de las 3 rectas (regresiones logísticas y playa). El punto de corte situado más al norte (latitud mayor) es: *(9.557243, 1.102056)* y el punto de corte situado más al sur es: *(9.5671585, 1.0837867).*

Para estas funciones se han utilizado los modelos anteriores y mediante una normalización de las probabilidades se han hallado las 3 componentes. Ha sido necesaria la normalización de estos debido a que al utilizar dos regresiones logísticas binomiales las probabilidades para la cala *"pivote"* no se podía obtener el valor realizando una resta con la unidad. La fórmula utilizada para la normalización de las probabilidades de cada una de las variables es[1]:

$$p(cala = 0) = \frac{\frac{P(cala=0|modelo=1)}{P(cala=1|modelo=1)}}{1 + \frac{P(cala=0|modelo=1)}{P(cala=1|modelo=1)} + \frac{P(cala=2|modelo=2)}{P(cala=1|modelo=2)}} \tag{1}$$

$$p(cala = 2) = \frac{\frac{P(cala=2|modelo=2)}{P(cala=1|modelo=2)}}{1 + \frac{P(cala=0|modelo=1)}{P(cala=1|modelo=1)} + \frac{P(cala=2|modelo=2)}{P(cala=1|modelo=2)}} \tag{2}$$

$$p(cala = 1) = 1 - (p(cala = 0) + p(cala = 2)) \tag{3}$$

Para el cálculo de la matriz de confusión de este modelo se ha realizado una partición de los datos utilizando el 80 % para entrenarlo y el resto como test obteniendo los resultados que se pueden encontrar en la figura 18.
Para calcular la probabilidad de que dadas unas coordenadas un fardo aparezca en cada cala se ha combinado el resultado obtenido de las dos regresiones logísticas en una función *calcularProbCala* que devuelve un vector con las tres probabilidades correspondientes. El cálculo de la cala para una coordenada se hace en *caclularCala* utilizando esta función previamente comentada devolviendo como resultado la calla cuya probabilidad sea más alta.

Los resultados obtenidos para los tres fardos pedidos han sido:

- Primer fardo con coordenadas *(9,558359, 1,1)*: Llega a la cala 2 con una probabilidad: *(0.0001090677 0.414166 0.5857249)* para las respectivas calas.

[1] `https://en.wikipedia.org/wiki/Multinomial_logistic_regression#As_a_set_of_independent_binary_regressions`

- Segundo fardo con coordenadas *(9,564329, 1,089)* : Llega a la cala 1 con una probabilidad: *(0.2303593 0.7683376 0.001303152)* para las respectivas calas.
- Tercer fardo con coordenadas *(9,568671, 1,081)*: Llega a la cala 0 con una probabilidad: *(0.9804216 0.01957819 2.493425e-07)* para las respectivas calas.

3.2. Parte 2: ¿Podría hacerse con menos fardos?

Se pretende saber con cuántos fardos podría haberse llevarse a cabo el experimento. Para determinar que era posible se seguirá cualquiera de los siguientes dos criterios sobre 1000 experimentos:

- La media de las 100 peores distancias desde el punto norte (sur) calculado en ese modelo hasta el punto norte (sur) calculado inicialmente con todos los datos es menor que 100 metros.
- La media de los 100 peores porcentajes de acierto para la predicción de todos los datos difiere en menos de un punto porcentual de el porcentaje de acierto del modelo con todos los fardos.

Para obtener las medidas anteriores se han empleado las funciones de la parte anterior. Primero, se han obtenido las rectas y tasa de acierto con la totalidad de los datos y, acto seguido, para cada porcentaje de los datos a analizar, se llevan a cabo 1000 iteraciones recalculando los modelos y los puntos de intersección entre las rectas generadas (punto norte y sur), así como la distancia de los puntos a los originales con todos los datos. Con ello se obtienen las peores distancias y porcentajes de acierto, lo que permite determinar si se podría haber llevado a cabo, o no, con ese número de fardos.
La condición de las distancias entre los puntos norte y sur se satisface con 420, 480 y 520 fardos. Por otro lado, la condición de los porcentajes de acierto se satisface con todos las cantidades probadas. Por tanto, si nuestra restricción de aceptación es que se cumpla alguno de los anteriores, con 60 fardos se cumple una de ellas y podrían considerarse suficientes. Por otra parte, si se exige el cumplimiento de ambas, serían necesarios como mínimo 420 fardos. La salida del programa generado puede observarse al final de la memoria.

Para cada una de las 1000 iteraciones se ha seguido un procedimiento parecido al del apartado anterior. Primero, se calculan dos rectas que cortan a la línea que define la playa. Acto seguido se encuentra el punto en el que intersecta cada una de ellas con la anterior, es decir, los puntos norte y sur. A continuación se predice la cala para todos los resultados y se calcula el porcentaje de aciertos que va almacenándose en una lista. De la misma forma, para cada punto norte y sur se calcula la distancia con los puntos norte y sur originales (calculados también en el código) y se almacena. Al finalizar las 1000 iteraciones se ordenan las listas y se escogen los 100 peores. Se calcula la media y se evalúa la condición para saber si podría hacerse con ese número de fardos.

4. Conclusiones

En esta práctica se han estudiado las utilidades de los modelos de regresión lineal y regresión logística. Estos modelos aunque parecen simples tienen multitud de aplicaciones tales como la predicción de valores influyentes en estudios como el de los antioxidantes de la cerveza. En el caso de las regresiones logísticas la utilización de estos modelos para obtener las probabilidades de que un atributo pertenezca a una clase o a otra.

En la primera parte de la práctica se ha utilizado la regresión lineal para determinar que componentes tienen una gran influencia en la aparición de antioxidantes. Este experimento ha sido interesante de cara a obtener una mejor visión de cómo funciona el modelo de regresión lineal. Respecto al segundo problema elegimos utilizar dos regresiones logísticas en vez de una multinomial lo que supuso una mayor complicación a la hora de calcular las probabilidades y la correspondiente matriz de confusión asociada al modelo por lo que se podría contemplar la alternativa de haber utilizado la multinomial. Ello habría simplificado el proceso, pero se decidió restringirse a los tipos de modelos estudiados en la asignatura.

Uno de los aspectos más interesantes de la práctica ha sido el estudio del número de muestras que habría sido necesario lanzar al mar para poder llevar a cabo el experimento con resultados similares. Supone una puesta en práctica del uso de modelos de regresión logística a un problema que podría llegar a ser una adaptación a uno real, y ha nos ha obligado al estudio de las distintas técnicas.

4.1. Control de esfuerzos

- Pedro Allué Tamargo: análisis DSA, análisis RP, memoria de los análisis anteriores, cálculo de los puntos de intersección del problema 2 parte 1, memoria problema 2 parte 1, figuras.
  - Horas invertidas: 13
- Cristina Oriol García: análisis ASA, análisisMCA y sus correspondientes apartados en la memoria, cálculo matriz de confusión, probabilidad y calas de los tres puntos pedidos del problema 2 parte 1, memoria.
  - Horas invertidas: 12
- Alejandro Paricio García: análisis $ORAC$ y DSA y sus apartados de la memoria. Apartado 2.2, decisión de si podría, o no, hacerse con menos fardos. Memoria del anterior.
  - Horas invertidas: 13.5

5. Anexo 1: Figuras

5.1. Análisis DSA

```
> beer.model = lm(dsa~tpc+ma+tso2, data=beer.data)
> summary(beer.model)

Call:
lm(formula = dsa ~ tpc + ma + tso2, data = beer.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 1: Captura de pantalla de los resultados del análisis DSA utilizando 3 variables

```
> beer.two.model = lm(dsa~tpc+tso2, data=beer.data)
> summary(beer.two.model)

Call:
lm(formula = dsa ~ tpc + tso2, data = beer.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.10807 -0.06153 -0.01832  0.03494  0.37108

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0202398  0.0694448  -0.291   0.7723
tpc          0.0035661  0.0003696   9.648 1.21e-11 ***
tso2         0.0032960  0.0018602   1.772   0.0847 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09369 on 37 degrees of freedom
Multiple R-squared:  0.7161,    Adjusted R-squared:  0.7008
F-statistic: 46.67 on 2 and 37 DF,  p-value: 7.644e-11
```

Figura 2: Captura de pantalla de los resultados del análisis DSA sin la variable *ma*

```
> beer.three.model = lm(dsa~tpc, data=beer.data)
> summary(beer.three.model)

Call:
lm(formula = dsa ~ tpc, data = beer.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13223 -0.05551 -0.01483  0.02167  0.36938

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0343018  0.0639781   0.536    0.595
tpc         0.0034132  0.0003694   9.240 2.93e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09629 on 38 degrees of freedom
Multiple R-squared:  0.692,     Adjusted R-squared:  0.6839
F-statistic: 85.38 on 1 and 38 DF,  p-value: 2.926e-11
```

Figura 3: Captura de pantalla de los resultadolabel=lst:labels del análisis DSA solo con la variable *tpc*

```
> beer.four.model = lm(dsa~tso2, data=beer.data)
> summary(beer.four.model)

Call:
lm(formula = dsa ~ tso2, data = beer.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37150 -0.08752  0.00168  0.06868  0.35616

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6163238  0.0400909  15.373   <2e-16 ***
tso2        -0.0008949  0.0033467  -0.267    0.791
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1734 on 38 degrees of freedom
Multiple R-squared:  0.001878,  Adjusted R-squared:  -0.02439
F-statistic: 0.07151 on 1 and 38 DF,  p-value: 0.7906
```

Figura 4: Captura de pantalla de la salida del análisis DSA solo con la variable *tso2*

5.2. Análisis RP

```
> beer.model = lm(rp~tpc+ma+tso2, data = beer.data)
> summary(beer.model)

Call:
lm(formula = rp ~ tpc + ma + tso2, data = beer.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 5: Captura de pantalla de la salida del análisis RP utilizando las 3 variables

```
> beer.model2 = lm(rp~tpc+tso2, data = beer.data)
> summary(beer.model2)

Call:
lm(formula = rp ~ tpc + tso2, data = beer.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 6: Captura de pantalla de la salida del análisis RP utilizando 2 variables (*tpc* y *tso2*)

```
> beer.model3 = lm(rp~tpc, data = beer.data)
> summary(beer.model3)

Call:
lm(formula = rp ~ tpc, data = beer.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32256 -0.03819 -0.01002  0.05190  0.32661

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1408956  0.0835284   1.687   0.0998 .
tpc         0.0032670  0.0004823   6.774 4.97e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1257 on 38 degrees of freedom
Multiple R-squared:  0.547,     Adjusted R-squared:  0.5351
F-statistic: 45.89 on 1 and 38 DF,  p-value: 4.973e-08
```

Figura 7: Captura de pantalla de la salida del análisis RP utilizando la variable *tpc*

```
> beer.model4 = lm(rp~tso2, data = beer.data)
> summary(beer.model4)

Call:
lm(formula = rp ~ tso2, data = beer.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33967 -0.08243 -0.01739  0.07839  0.53343

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.766571   0.039754  19.283   <2e-16 ***
tso2        -0.008701   0.003319  -2.622   0.0125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1719 on 38 degrees of freedom
Multiple R-squared:  0.1532,    Adjusted R-squared:  0.1309
F-statistic: 6.875 on 1 and 38 DF,  p-value: 0.0125
```

Figura 8: Captura de pantalla de la salida del análisis RP utilizando la variable *tso2*

5.3. Análisis ORAC

```
> beer.model = lm(orac~tpc+ma+tso2, data = beer.data)
> summary(beer.model)

Call:
lm(formula = orac ~ tpc + ma + tso2, data = beer.data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1864 -1.4553  0.1254  1.4146  5.1889

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.419372   1.973807   0.719  0.47672
tpc          0.029489   0.009916   2.974  0.00522 **
ma          -0.269269   0.156208  -1.724  0.09333 .
tso2        -0.004080   0.048196  -0.085  0.93300
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.265 on 36 degrees of freedom
Multiple R-squared:  0.2014,    Adjusted R-squared:  0.1348
F-statistic: 3.025 on 3 and 36 DF,  p-value: 0.04199
```

Figura 9: Captura de pantalla de la salida del análisis ORAC utilizando las tres variables.

```
> beer.model = lm(orac~tpc, data = beer.data)
> summary(beer.model)

Call:
lm(formula = orac ~ tpc, data = beer.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 10: Captura de pantalla de la salida del análisis ORAC utilizando únicamente tpc.

```
> beer.model = lm(orac~1, data = beer.data)
> summary(beer.model)

Call:
lm(formula = orac ~ 1, data = beer.data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5605 -2.0705 -0.2805  1.3595  5.5495

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.5705     0.3851   9.272 2.08e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.435 on 39 degrees of freedom
```

Figura 11: Captura de pantalla del análisis ORAC sin predictores.

5.4. Análisis ASA

```
> beer.model=lm(asa~tpc+ma+tso2, data=beer.data)
> summary(beer.model)

Call:
lm(formula = asa ~ tpc + ma + tso2, data = beer.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 12: Captura de pantalla de la salida del análisis ASA utilizando las tres variables.

```
> beer.model=lm(asa~tpc, data=beer.data)
> summary(beer.model)

Call:
lm(formula = asa ~ tpc, data = beer.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09745 -0.13653  0.08454  0.17441  0.94192

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.480533   0.226812   2.119 0.040716 *
tpc         0.005079   0.001310   3.879 0.000404 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3414 on 38 degrees of freedom
Multiple R-squared:  0.2836,    Adjusted R-squared:  0.2648
F-statistic: 15.04 on 1 and 38 DF,  p-value: 0.000404
```

Figura 13: Captura de pantalla de la salida del análisis ASA utilizando únicamente tpc.

```
> beer.model=lm(asa~ma+tso2, data=beer.data)
> summary(beer.model)

Call:
lm(formula = asa ~ ma + tso2, data = beer.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15777 -0.20266  0.05921  0.20237  0.96461

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.993308   0.297225   3.342  0.00191 **
ma           0.034741   0.024486   1.419  0.16432
tso2        -0.001854   0.008376  -0.221  0.82601
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.394 on 37 degrees of freedom
Multiple R-squared:  0.0708,    Adjusted R-squared:  0.02058
F-statistic:  1.41 on 2 and 37 DF,  p-value: 0.257
```

Figura 14: Captura de pantalla del análisis ASA utilizando ma y tso2

5.5. Análisis MCA

```
> beer.model=lm(mca~ma+tpc+tso2, data=beer.data)
> summary(beer.model)

Call:
lm(formula = mca ~ ma + tpc + tso2, data = beer.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 .
ma            0.21620    1.24749   0.173 0.863379
tpc           0.32509    0.07919   4.105 0.000222 ***
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 15: Captura de pantalla de la salida del análisis MCA utilizando las tres variables.

5.6. Análisis Modelo con longitud y latitud

```
> fardos.model= glm(cala~longitud.x+latitud.y, data = fardos.data.2, family = binomial)
> summary(fardos.model)

Call:
glm(formula = cala ~ longitud.x + latitud.y, family = binomial,
    data = fardos.data.2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.37618  -0.07127  -0.00824   0.17644   2.05014

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4262.2    16532.5  -0.258    0.797
longitud.x     353.5     1627.8   0.217    0.828
latitud.y      803.1      885.6   0.907    0.364

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 763.82  on 599  degrees of freedom
Residual deviance: 221.62  on 597  degrees of freedom
AIC: 227.62

Number of Fisher Scoring iterations: 8
```

Figura 16: Captura de pantalla de la salida del análisis del modelo de la playa utilizando longitud y latitud.

5.7. Interpretación geométrica de la división de la playa

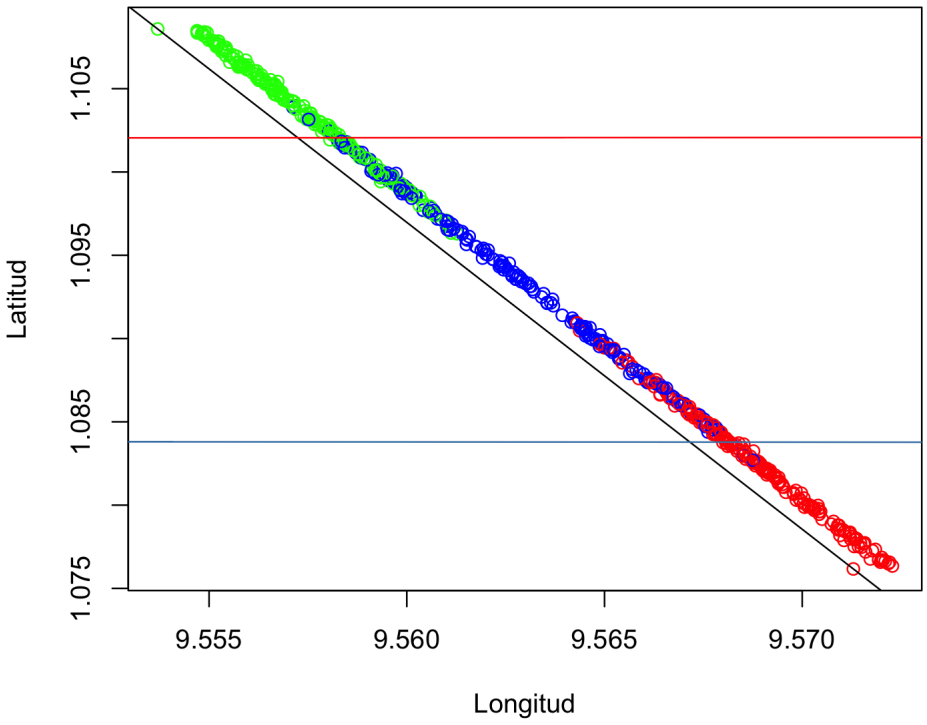

Figura 17: Interpretación geométrica de la división de la playa. En rojo: cala 2, en azul: cala 1 y en verde: cala 0. La playa queda dividida por dos rectas que intersectan con la línea formada por los fardos. Esas dos intersecciones dan como resultados los puntos norte y sur.

5.8. Matriz de confusión Parte 1

```
matrizConf  0  1  2
         0 38  6  0
         1  2 23  3
         2  0  5 43
```

Figura 18: Matriz de confusión para el modelo de la playa.

6. Anexo 2: Códigos

6.1. Código de creación de modelos

Listing 1: Fragmento de código de la definición de los modelos de regresión logística

```
# Lectura del fichero de fardos
fardos.data = read.csv("/home/cris/Escritorio/Universidad/almacenes_datos/Practica4
    /datacala.csv")
fardos.data$X = NULL
# Creaci n del plot de la playa
plot(fardos.data$longitud.x, fardos.data$latitud.y,
      xlab = "Longitud", ylab = "Latitud",
      col = ifelse(fardos.data$cala == 0, "red", ifelse(fardos.data$cala == 1, "blue
          ", "green")))
# Calculo de la recta (y=ax+b)
rectaPlaya = recta(fardos.data)
# Primer argumento -> intercept, segundo argumento -> slope
abline(rectaPlaya[2], rectaPlaya[1])
# Utilizando una regresi n log stica
fardos.data.1 = fardos.data
fardos.data.2 = fardos.data
# Los 1 apuntan al centro
fardos.data.1$cala = ifelse(fardos.data$cala == 0, 1, 0)
fardos.data.2$cala = ifelse(fardos.data$cala == 2, 1, 0)
# Se puede ignorar uno de ellos porque son combinaciones lineales (rectas)
fardos.model1 = glm(cala~latitud.y, data = fardos.data.1, family = binomial)
fardos.model2 = glm(cala~latitud.y, data = fardos.data.2, family = binomial)
```

6.2. Resultados ejecución parte 2

```
[1] 0.0005866667
Se podria haber hecho con 60 fardos.
Media de las peores distancias del primer punto de separacion:  501.3619
Media de las peores distancias del segundo punto de separacion:  464.8893
Media de los peores porcentajes de acierto:  0.8222533
[1] 0.00349
Se podria haber hecho con 120 fardos.
Media de las peores distancias del primer punto de separacion:  297.9804
Media de las peores distancias del segundo punto de separacion:  299.8049
Media de los peores porcentajes de acierto:  0.8251567
[1] 0.004413333
Se podria haber hecho con 180 fardos.
Media de las peores distancias del primer punto de separacion:  233.2526
Media de las peores distancias del segundo punto de separacion:  221.0558
Media de los peores porcentajes de acierto:  0.82608
[1] 0.004885
Se podria haber hecho con 240 fardos.
Media de las peores distancias del primer punto de separacion:  178.2154
Media de las peores distancias del segundo punto de separacion:  165.3858
Media de los peores porcentajes de acierto:  0.8265517
[1] 0.005021667
Se podria haber hecho con 300 fardos.
Media de las peores distancias del primer punto de separacion:  142.3178
Media de las peores distancias del segundo punto de separacion:  136.9945
Media de los peores porcentajes de acierto:  0.8266883
[1] 0.0047
Se podria haber hecho con 360 fardos.
Media de las peores distancias del primer punto de separacion:  121.6917
Media de las peores distancias del segundo punto de separacion:  109.2879
Media de los peores porcentajes de acierto:  0.8263667
[1] 0.00473
Se podria haber hecho con 420 fardos.
Media de las peores distancias del primer punto de separacion:  97.99388
Media de las peores distancias del segundo punto de separacion:  94.66309
Media de los peores porcentajes de acierto:  0.8263967
[1] 0.004541667
Se podria haber hecho con 480 fardos.
Media de las peores distancias del primer punto de separacion:  74.27477
Media de las peores distancias del segundo punto de separacion:  69.41352
Media de los peores porcentajes de acierto:  0.8262083
[1] 0.003846667
Se podria haber hecho con 540 fardos.
Media de las peores distancias del primer punto de separacion:  46.11319
Media de las peores distancias del segundo punto de separacion:  44.92015
Media de los peores porcentajes de acierto:  0.8255133
```