La aplicación de la regresión logística en el estudio de las cerámicas griegas: Hacia una arqueología clásica más objetiva

Álvaro Miranda García - Aplicaciones Informáticas en Arqueología y Estadística

2023-05-28

Lekythos y Kylix

Las cerámicas griegas que recorrieron el Mediterráneo antiguo han llamado la atención a estudiosos e investigadores por su gran interés, tanto artístico en sus imágenes y belleza; como arqueológico en su valor para datar contextos con extrema precisión y documentar los numerosos y variados contactos transmediterráneos entre comunidades a lo largo del I milenio a. C. Sin embargo, su investigación se encuentra arraigada en la tradición histórico-artística, y se halla reacia a utilizar nuevas herramientas como la que hoy nos ocupa. La estadística y las matemáticas en R nos ofrecen muchas posibilidades para un análisis más objetivo de estas cerámicas griegas.

Mediante el uso de R para realizar análisis estadísticos como la regresión logística, podremos predecir el resultado de ciertas variables asignadas a nuestro registro cerámico; en nuestro caso,veremos la relación entre la forma tipológica o la técnica, y alguna variable morfométrica como altura o diámetro de borde o base, utilizando como ejemplos la Kylix y el Lekythos, es decir, una forma abierta, como es la Kylix, y una cerrada, como es el lekythos. Analizando las medidas mediante una regresión logística utilizando la forma como predictor, podremos predecir si un ejemplar pertenece a un grupo tipológico o a otro, de forma científica y objetiva, mediante la estadística y las matemáticas. La arqueología y los estudios cerámicos a menudo pecan de resultados e interpretaciones poco científicas; desde luego, nuestra disciplina se puede beneficiar de estas metodologías y de la aplicación de la estadística y las matemáticas.

La base de datos que se ha elaborado para este trabajo se ha hecho con registros de kylikes y lekythoi procedentes de varios yacimientos occidentales. Para los lekythoi: Ampurias, el pecio de El Sec, Castellones de Ceal, El Pajarillo y, sobre todo, Ibiza, cuya necrópolis ha aportado uno de los mayores conjuntos de lekythoi griegos con excelentes estados de conservación. Para las kylikes: Cancho Roano, Mértola, Castro Marim, Cerro del Castillo (Fuengirola), Cástulo, Ampurias, el pecio de El Sec, así como varias piezas de Mesas de Asta, cuyo rico conjunto de cerámicas griegas estamos estudiando en el Museo Arqueológico de Jerez de la Frontera para el Trabajo de Fin de Grado.

En la base de datos se incluyen 83 muestras, con las siguientes variables:

- Yacimiento
- Forma: aquí, distinguimos entre Kylix, copas abiertas usadas para el consumo de vino; y el Lekythos, una forma cerrada utilizada para aceites y ungüentos.
- Técnica: Distinguimos entre cerámicas Figuradas, y cerámicas de Barniz Negro, sin imágenes.
- Datos morfométricos, como diámetro de borde y base, altura y grosor de pared.

Regresión Logística

El método que vamos a aplicar en el presente trabajo es la regresión logística, esta fue desarrollada por David Cox en 1958. La regresión logística se utiliza para predecir la clase (o categoría) de los individuos en función de una o varias variables predictoras (x). Se utiliza para modelizar un resultado binario, es decir, una variable que solo puede tener dos valores posibles: 0 o 1, sí o no etc. La regresión logística pertenece a la familia llamada: Generalized Linear Model (GLM).

En nuestro caso, hemos utilizado 0 y 1 para sustituir las categorias Figurada y Barniz Negro dentro de la variable “Técnica”, y Lekythos y Kylix dentro de “Forma”. Con esto, podremos estimar la probabilidad de que una vasija con ciertas medidas morfométricas sea de una u otra forma, o tienda a presentar una decoración realizada con una u otra técnica. El potencial y la posibilidad de aplicaciones de esta metodología es enorme, y es una herramienta excelente para construir interpretaciones históricas de forma científica.

La función para la regresión logística estándar, para predecir el resultado de una observación dada una variable predictora (x), es una curva en forma de S definida como:

$$p = exp(y)/[1 + exp(y)]$$

Simplificada también como:

$$p = 1/[1 + exp(-y)]$$

$$p = 1/[1 + exp(-y)]$$

Manipulando un poco la función, se puede demostrar que:

$$p/(1 - p) = exp(b0 + b1 * x)$$

$$p/(1 - p) = exp(b0 + b1 * x)$$

Tomando el logaritmo de ambos lados, la fórmula se convierte en una combinación lineal de predictores:

$$log[p/(1 - p)] = b0 + b1 * x$$

$$log[p/(1 - p)] = b0 + b1 * x$$

Pero, ¿Por qué regresión logística?

La existencia de una relación entre una variable cualitativa con dos niveles y una variable continua se puede estudiar también mediante otros tests estadísticos, como t-test. Sin embargo, la regresión logística permite también calcular la probabilidad de que la variable dependiente pertenezca a cada una de las dos categorías

en función del valor que adquiera la variable independiente o predictora, que es lo que nos interesa en este caso.

La regresión lineal por mínimos cuadrados también permite crear un modelo para una variable cualitativa binomial codificada como 0 y 1, pero con ella, al ser para valores extremos del predictor, se obtienen valores de Y menores que 0 o mayores que 1, y no nos da como resultado una probabilidad dentro del rango [0,1], por ello utilizamos la regresión logística. La regresión logística transforma el valor devuelto por la regresión lineal ($\beta 0+\beta 1X$) empleando una función cuyo resultado está siempre comprendido entre 0 y 1.

Antes de empezar, hay que tratar ciertos aspectos, como los logit o “log of the odds”. El logit es el coeficiente proporcionado por una regresión logística en R.

Para calcular la regresión logística, utilizamos la función glm(), de Generalized Linear Model. Es necesario especificar la opción family = binomial, que indica a R que queremos ajustar la regresión logística.

Desarrollo

Lo primero que cabe hacer es introducir nuestra tabla, y convertirla a dataframe para poder trabajar eficazmente con ella. Los valores “Kylix”, “Lekythos”, “Figurada” y “Barniz Negro” ya se han renombrado como 0 y 1 para facilitar el workflow del trabajo.

```
#Introducimos la tabla
library(readxl)
copas <- read_excel("E:\\trabajo\\copas2.xlsx")
```

```
#Convertimos a dataframe. Los paquetes que vamos a necesitar son Tidyverse, para una sencilla manipulación y visualización de los datos; Caret, que facilita el machine learning y el uso de métodos complejos de clasificación y regresión, e ISLR.
ceramicas = as.data.frame(copas)
is.data.frame(ceramicas)
```

```
## [1] TRUE
```

```
View (ceramicas)


library(ISLR)
library(tidyverse)
library(caret)

#Con el paquete DT podemos visualizar de una forma más dinámica nuestra base de datos.
library(DT)
datatable(ceramicas)
```

Show 10 entries Search:

	N	Yacimiento	Forma	Tecnica	Diam. Base	Diam. Borde	Altura	Gros. Pared
1	1	Cancho Roano	1	0	8.8	16	4.6	0.4

	N	Yacimiento	Forma	Tecnica	Diam. Base	Diam. Borde	Altura	Gros. Pared
2	2	Cancho Roano	1	0	9	NA	2.2	0.8
3	3	Cancho Roano	1	0	9.5	17	3.2	0.4
4	4	Cancho Roano	1	0	8	16.2	4.3	0.4
5	5	Cancho Roano	1	0	7.5	15	4.5	0.5
6	6	Cancho Roano	1	0	15	NA	3	0.5
7	7	Cancho Roano	1	1	9	11.4	NA	0.5
8	8	Cancho Roano	1	1	9.5	16	5.4	0.5
9	9	Cancho Roano	1	1	8.2	14.5	4.2	0.5
10	10	Cancho Roano	1	1	7.4	14.5	4.6	0.6

Showing 1 to 10 of 83 entries

Previous 1 2 3 4 5 … 9 Next

```
#Ahora, creamos objetos para las variables que vamos a estudiar, así facilitamos su manejo.

forma = ceramicas$Forma
tecnica = ceramicas$Tecnica
borde = na.omit (ceramicas$`Diam. Borde`)
base = na.omit (ceramicas$`Diam. Base`)
altura = na.omit (ceramicas$Altura)


#Nos aseguramos que los vectores sean numéricos y no se registran como factor o character, lo
s convertimos con as.numeric().


borde <- as.numeric(borde)
base <- as.numeric(base)
altura <- as.numeric(altura)


#Con la función glm () ajustamos los modelos. Primero, vamos a usar la forma como variable pr
edictora, y el diámetro del borde como variable dependiente.


modelo <- glm (forma ~ borde, data = ceramicas, family = binomial)

#Con summary () obtenemos los resultados del modelo.
summary (modelo)
```

```
##
## Call:
## glm(formula = forma ~ borde, family = binomial, data = ceramicas)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -3.281e-05   1.100e-08   2.100e-08   2.100e-08   3.876e-05
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -71.636  62932.966  -0.001    0.999
## borde           8.823   7134.947   0.001    0.999
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.9974e+01  on 79  degrees of freedom
## Residual deviance: 2.5791e-09  on 78  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 4
##
## Number of Fisher Scoring iterations: 25
```

```
#Ahora, utilizamos la altura como variable dependiente

modelo1 <- glm (forma ~ altura, data = ceramicas, family = binomial)
summary(modelo1)
```

```
##
## Call:
## glm(formula = forma ~ altura, family = binomial, data = ceramicas)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.06737  -0.01902   0.14586   0.20504   1.59190
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  13.3811     4.3410   3.082  0.00205 **
## altura       -2.1055     0.7133  -2.952  0.00316 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 63.510  on 47  degrees of freedom
## Residual deviance: 15.465  on 46  degrees of freedom
##   (35 observations deleted due to missingness)
## AIC: 19.465
##
## Number of Fisher Scoring iterations: 8
```

La tabla de coeficientes muestra las estimaciones de los coeficientes β y sus niveles de significación. Las columnas son:

Estimate: el intercept (b0) y las estimaciones del coeficiente β asociadas a cada variable predictora.

Standard error: el error estándar de las estimaciones de los coeficientes. Representa la precisión de los coeficientes.

Z value: el estadístico z, que es la estimación del coeficiente dividida por el error estándar de la estimación.

Pr(>|z|): El valor p correspondiente al estadístico z.

Un concepto importante que hay que entender para interpretar los coeficientes β es el **odds ratio**. Un odds ratio mide la asociación entre una variable predictora (x) y la variable de resultado (y). Representa la proporción de probabilidades de que se produzca un suceso (suceso = 1) dada la presencia del predictor x (x = 1), comparada con las probabilidades de que se produzca el suceso en ausencia de ese predictor (x = 0). En nuestro caso, x = 0 corresponde a Lekythos en la variable Forma, y x = 1 corresponde a Kylix.

El coeficiente de regresión logística β asociado a un predictor X es el cambio esperado en las log odss de tener el resultado por cambio unitario en X. Así, aumentar el predictor en 1 unidad multiplica las probabilidades de tener el resultado por eβ.

GLM ajusta una función de enlace. En el modelo logit esta función es:

$$\eta = log(p/1-p)$$

$$\eta = log(p/1-p)$$

El cociente p/1−p es el odds ratio. Entonces, los coeficientes del modelo logit se interpretan como el logaritmo del odds ratio. El coeficiente de la variable Borde es 8.823, esto indica que el logaritmo del odds ratio de ser x = 1, aumenta 8.823 unidades por cada unidad que aumenta el diámetro del borde. Entonces, un aumento del diámetro del borde es un aumento de la probabilidad de que esa cerámica corresponda a una Kylix.

En el caso de la variable Altura, vemos que esta es significativa por tener un p-value <0.05. Su coeficiente es -2,1055. Aquí, un aumento de unidad de altura corresponde a una disminución de 2.1055 unidades en la probabilidad de x = 1. Dicho de otra forma, una disminución en altura es un aumento de la probabilidad de que sea una Kylix.

Esto se aplica al resto de los modelos que hagamos, ya sea simples, o múltiples junto a otras variables.

Ahora, podemos analizar los modelos con otras funciones como anova (), confint () o predict ().

```
#Con anova() analizamos la tabla de desviación.
anova(modelo, test ='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: forma
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     79     89.974
## borde  1   89.974        78      0.000 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Con confint () obtenemos los intervalos de confianza correspondientes a los coeficientes de correlación del modelo.
confint(object = modelo, level = 0.95)
```

```
##                   2.5 %   97.5 %
## (Intercept) -37372.8750 5792.000
## borde         -655.9591 4237.805
```

```
#Con la función predict() podemos, incluso, obtener una respuesta (con la opción type ="response") acerca de a qué forma pertenece cada ejemplar según su diámetro de borde.

newdataborde <- data.frame(borde)
probabilities <- modelo %>% predict(newdataborde, type = "response")
predicted.classes1 <- ifelse(probabilities > 0.5, "Kylix", "Lekythos")
predicted.classes1
```

```
##          1          2          3          4          5          6          7
##    "Kylix"         NA    "Kylix"    "Kylix"    "Kylix"         NA    "Kylix"
##          8          9         10         11         12         13         14
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         15         16         17         18         19         20         21
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         22         23         24         25         26         27         28
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         29         30         31         32         33         34         35
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         36         37         38         39         40         41         42
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         43         44         45         46         47         48         49
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         50         51         52         53         54         55         56
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         57         58         59         60         61         62         63
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix" "Lekythos"
##         64         65         66         67         68         69         70
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"
##         71         72         73         74         75         76         77
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"         NA "Lekythos"
##         78         79         80         81         82         83
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"
```

Ahora, con la Altura:

```
confint(object = modelo1, level = 0.95)
```

```
##                2.5 %    97.5 %
## (Intercept)  7.027539 25.060056
## altura      -3.982328 -1.053857
```

```
anova(modelo1, test ='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: forma
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      47     63.510
## altura  1   48.045        46     15.465 4.165e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
newdataaltura <- data.frame(altura)
probabilities <- modelo %>% predict(newdataaltura, type = "response")
predicted.classes2 <- ifelse(probabilities > 0.5, "Kylix", "Lekythos")
predicted.classes2
```

```
##          1          2          3          4          5          6          7
##    "Kylix"         NA    "Kylix"    "Kylix"    "Kylix"         NA    "Kylix"
##          8          9         10         11         12         13         14
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         15         16         17         18         19         20         21
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         22         23         24         25         26         27         28
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         29         30         31         32         33         34         35
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         36         37         38         39         40         41         42
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         43         44         45         46         47         48         49
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         50         51         52         53         54         55         56
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         57         58         59         60         61         62         63
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix" "Lekythos"
##         64         65         66         67         68         69         70
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"
##         71         72         73         74         75         76         77
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"         NA "Lekythos"
##         78         79         80         81         82         83
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"
```

Con Anova (), vemos que la altura y el borde son significativos para la forma.

Ahora que hemos hecho regresiones logísticas simples, vamos a probar con una múltiple.

```
modelo2 <- glm (forma ~ altura + borde, data = ceramicas, family = binomial)
summary(modelo2)
```

```
##
## Call:
## glm(formula = forma ~ altura + borde, family = binomial, data = ceramicas)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -9.006e-06  -5.402e-07   2.110e-08   2.110e-08   2.131e-05
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -31.363 267585.015       0        1
## altura          -1.358  29794.508       0        1
## borde            5.800  12488.337       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.1578e+01  on 45  degrees of freedom
## Residual deviance: 7.3882e-10  on 43  degrees of freedom
##   (37 observations deleted due to missingness)
## AIC: 6
##
## Number of Fisher Scoring iterations: 25
```

```
anova(modelo2, test ='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: forma
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      45     61.578
## altura  1   46.115        44     15.463 1.115e-11 ***
## borde   1   15.463        43      0.000 8.415e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(object = modelo1, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept)  7.027539 25.060056
## altura      -3.982328 -1.053857
```

```
newdatamult <- data.frame(altura + borde)
probabilities <- modelo2 %>% predict(newdatamult, type = "response")
predicted.classes2 <- ifelse(probabilities > 0.5, "Kylix", "Lekythos")
predicted.classes2
```

```
##          1          2          3          4          5          6          7
##    "Kylix"         NA    "Kylix"    "Kylix"    "Kylix"         NA         NA
##          8          9         10         11         12         13         14
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         15         16         17         18         19         20         21
##    "Kylix"    "Kylix"         NA         NA         NA         NA         NA
##         22         23         24         25         26         27         28
##         NA         NA         NA         NA         NA         NA         NA
##         29         30         31         32         33         34         35
##         NA         NA         NA         NA         NA         NA         NA
##         36         37         38         39         40         41         42
##         NA         NA         NA         NA         NA         NA         NA
##         43         44         45         46         47         48         49
##         NA    "Kylix"         NA         NA         NA         NA    "Kylix"
##         50         51         52         53         54         55         56
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         57         58         59         60         61         62         63
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix" "Lekythos"
##         64         65         66         67         68         69         70
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"
##         71         72         73         74         75         76         77
## "Lekythos" "Lekythos" "Lekythos"         NA         NA         NA "Lekythos"
##         78         79         80         81         82         83
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"
```

En este predict () vemos como, al igual que en el primer caso, predice correctamente la tipología en función de la altura y el borde. Los ejemplares que dan como resultado NA es debido a que para ellos desconocemos alguna de las dos variables, altura o diámetro de borde. Vamos a probar sumando diámetro de base.

```
modelo3 <- glm (forma ~ altura + borde + base, data = ceramicas, family = binomial)
summary(modelo3)
```

```
##
## Call:
## glm(formula = forma ~ altura + borde + base, family = binomial,
##     data = ceramicas)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.025e-05  -2.110e-08   2.110e-08   2.110e-08   1.584e-05
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.249e+00  3.057e+05       0        1
## altura      -5.723e-01  2.476e+04       0        1
## borde        9.819e+00  3.900e+04       0        1
## base        -1.249e+01  9.890e+04       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5.5051e+01  on 39  degrees of freedom
## Residual deviance: 4.9687e-10  on 36  degrees of freedom
##   (43 observations deleted due to missingness)
## AIC: 8
##
## Number of Fisher Scoring iterations: 25
```

```
summary(modelo3)
```

```
##
## Call:
## glm(formula = forma ~ altura + borde + base, family = binomial,
##     data = ceramicas)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.025e-05  -2.110e-08   2.110e-08   2.110e-08   1.584e-05
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.249e+00  3.057e+05       0        1
## altura      -5.723e-01  2.476e+04       0        1
## borde        9.819e+00  3.900e+04       0        1
## base        -1.249e+01  9.890e+04       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5.5051e+01  on 39  degrees of freedom
## Residual deviance: 4.9687e-10  on 36  degrees of freedom
##   (43 observations deleted due to missingness)
## AIC: 8
##
## Number of Fisher Scoring iterations: 25
```

```
#Con coef () sacamos solamente los coeficientes.
coef(modelo3)
```

```
## (Intercept)      altura       borde        base
##  -7.2488468  -0.5723287   9.8186307 -12.4922573
```

```
anova(modelo3, test ='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: forma
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      39     55.051
## altura  1   40.509        38     14.542 1.957e-10 ***
## borde   1   14.542        37      0.000 0.0001371 ***
## base    1    0.000        36      0.000 0.9999916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
newdatamult2 <- data.frame(altura + borde + base)
probabilities <- modelo3 %>% predict(newdatamult2, type = "response")
predicted.classes3 <- ifelse(probabilities > 0.5, "Kylix", "Lekythos")
predicted.classes3
```

```
##          1          2          3          4          5          6          7
##    "Kylix"         NA    "Kylix"    "Kylix"    "Kylix"         NA         NA
##          8          9         10         11         12         13         14
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"
##         15         16         17         18         19         20         21
##    "Kylix"    "Kylix"         NA         NA         NA         NA         NA
##         22         23         24         25         26         27         28
##         NA         NA         NA         NA         NA         NA         NA
##         29         30         31         32         33         34         35
##         NA         NA         NA         NA         NA         NA         NA
##         36         37         38         39         40         41         42
##         NA         NA         NA         NA         NA         NA         NA
##         43         44         45         46         47         48         49
##         NA    "Kylix"         NA         NA         NA         NA         NA
##         50         51         52         53         54         55         56
##         NA         NA         NA         NA         NA    "Kylix"    "Kylix"
##         57         58         59         60         61         62         63
##    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix"    "Kylix" "Lekythos"
##         64         65         66         67         68         69         70
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"
##         71         72         73         74         75         76         77
## "Lekythos" "Lekythos" "Lekythos"         NA         NA         NA "Lekythos"
##         78         79         80         81         82         83
## "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos" "Lekythos"
```

Aquí, vemos que el diámetro de base no es significativo. En cuanto al predict (), para aquellos ejemplares que se encuentran completos y disponemos de altura y diámetro de borde y base, la predicción es, nuevamente, acertada.

A continuación, podemos usar la variable Tecnica como predictora, con Borde como dependiente.

```
modelo_tec <- glm (tecnica ~ borde, data = ceramicas, family = binomial)

summary(modelo_tec)
```

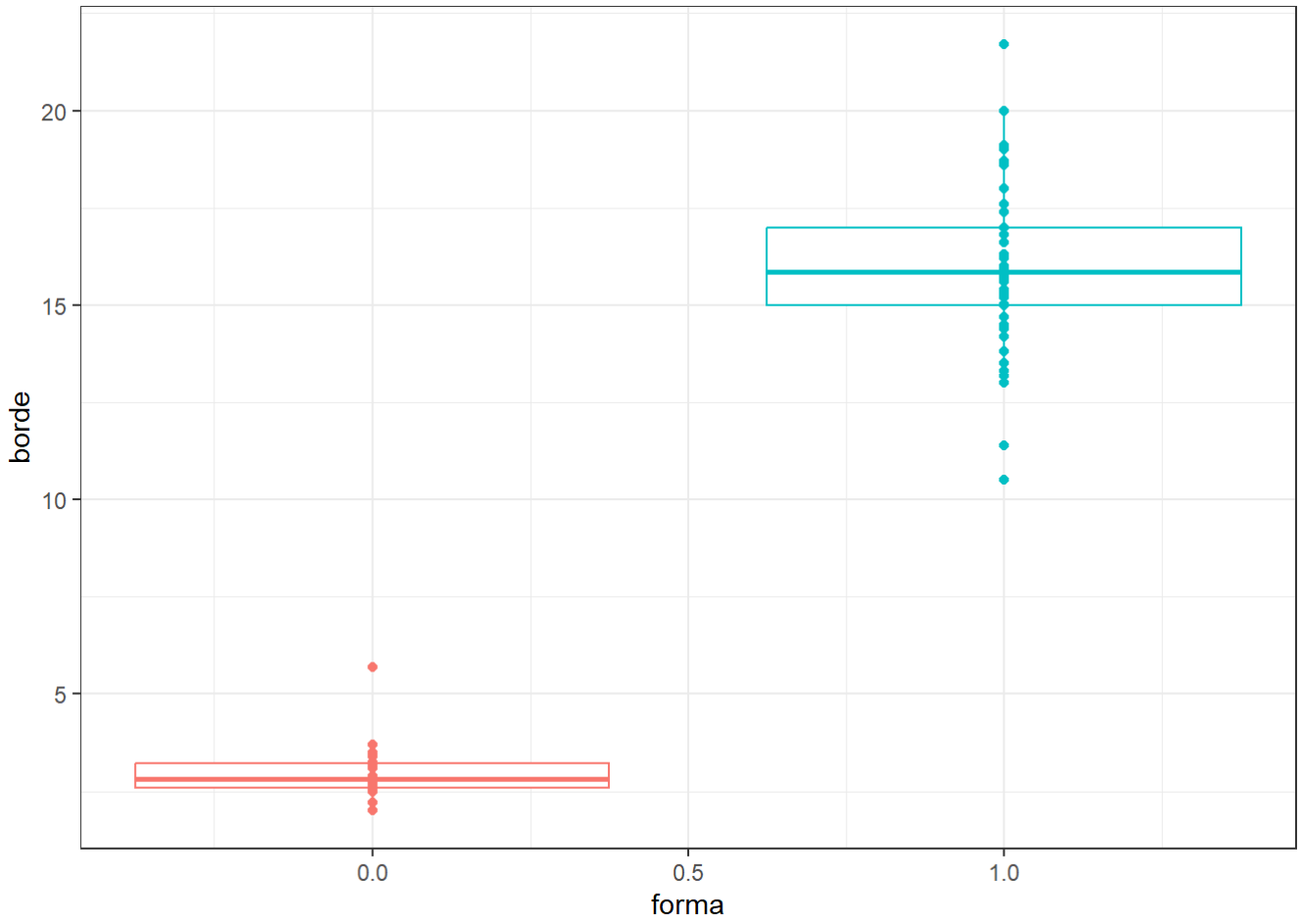
```
##
## Call:
## glm(formula = tecnica ~ borde, family = binomial, data = ceramicas)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1886  -0.4427   0.6365   0.8547   2.1754
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.14582    0.93282  -3.372 0.000745 ***
## borde        0.25093    0.06307   3.979 6.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.453  on 79  degrees of freedom
## Residual deviance:  83.382  on 78  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 87.382
##
## Number of Fisher Scoring iterations: 4
```

Dentro de nuestra base de datos, un aumento de unidad de diámetro de borde corresponde a un aumento de 0.25093 unidades en la probabilidad de que la cerámica sea de barniz negro, sin imágenes.

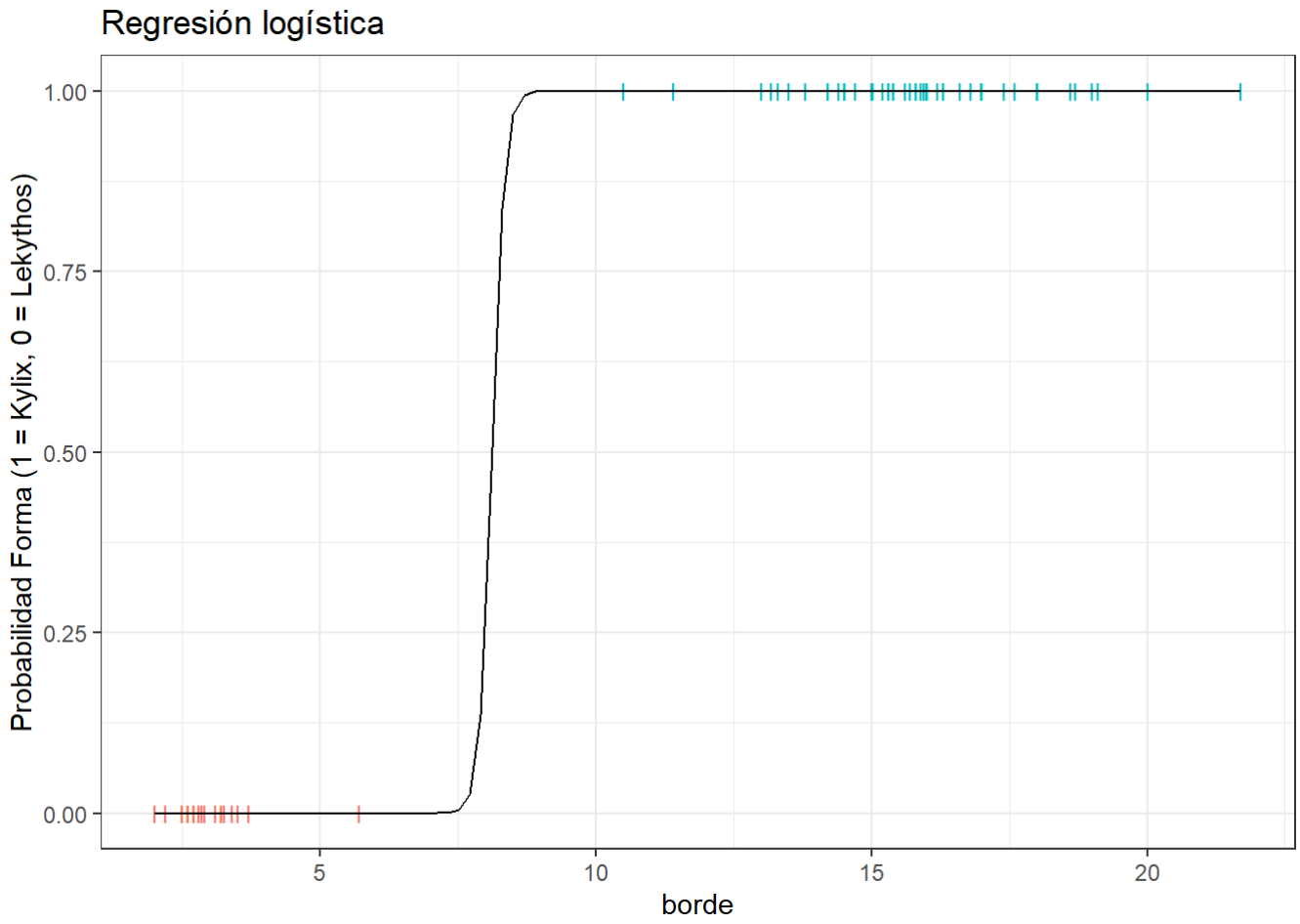
Representación Gráfica

Ahora, podemos representar la regresión logística con boxplot para un diagrama de caja y bigote, o con plot o ggplot para visualizar la curva en forma de S. Con la función predict(), añadiendo type = “response”, obtenemos directamente las probabilidades, las cuales podemos representar como una curva para incluir en los gráficos.

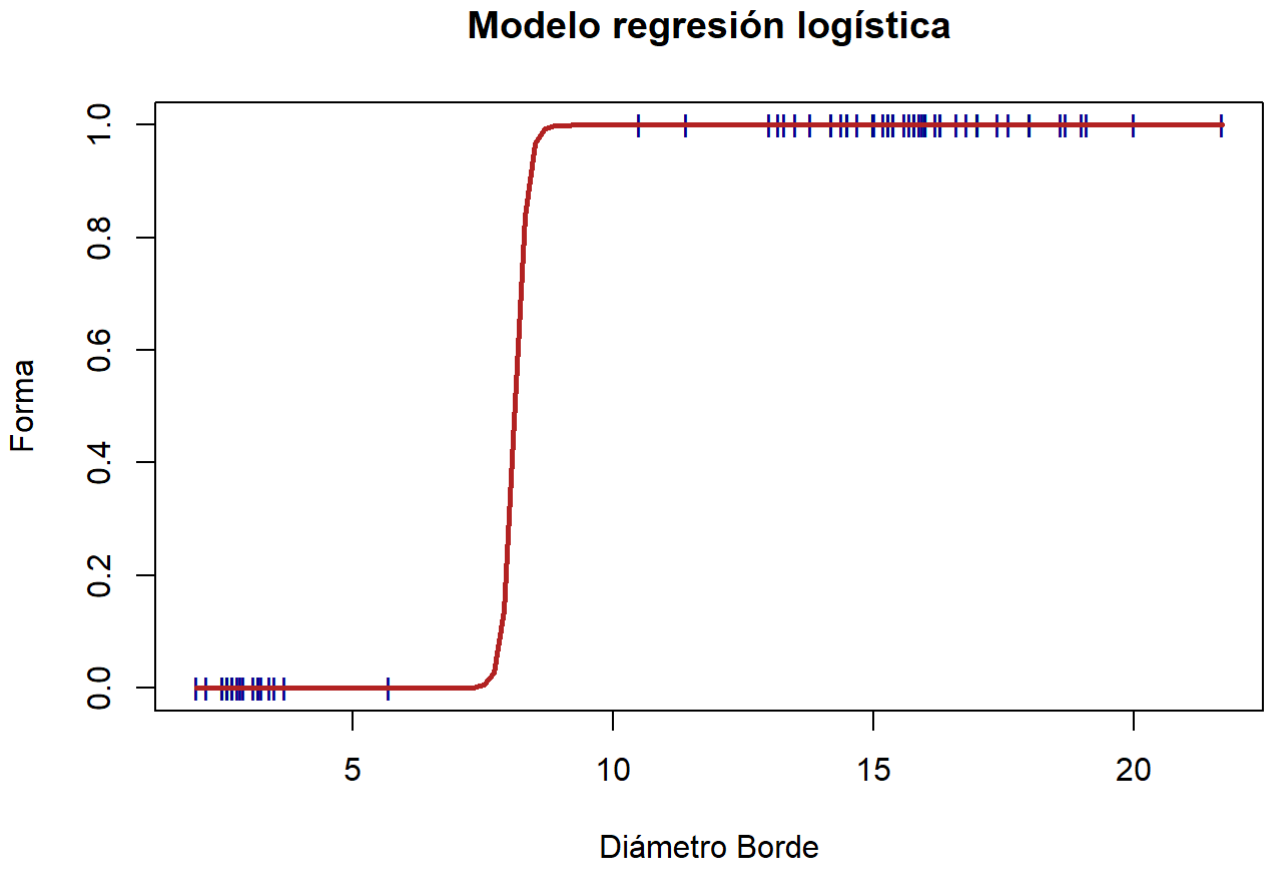
```
library(ggplot2)

#diagrama de caja y bigote

ggplot(data = ceramicas, mapping = aes(x = forma, y = borde)) +
geom_boxplot(aes(color = as.factor (forma))) +
geom_point(aes(color = as.factor (forma))) +
theme_bw() +
theme(legend.position = "null")
```




```
#Con ggplot ()

ggplot(data = ceramicas, aes(x = borde, y = forma)) +
  geom_point(aes(color = as.factor(forma)), shape = "I", size = 3) +
  stat_function(fun = function(x){predict(modelo,
                                          newdata = data.frame(borde = x),
                                          type = "response")}) +
  theme_bw() +
  labs(title = "Regresión logística",
       y = "Probabilidad Forma (1 = Kylix, 0 = Lekythos)") +
  theme(legend.position = "none")
```

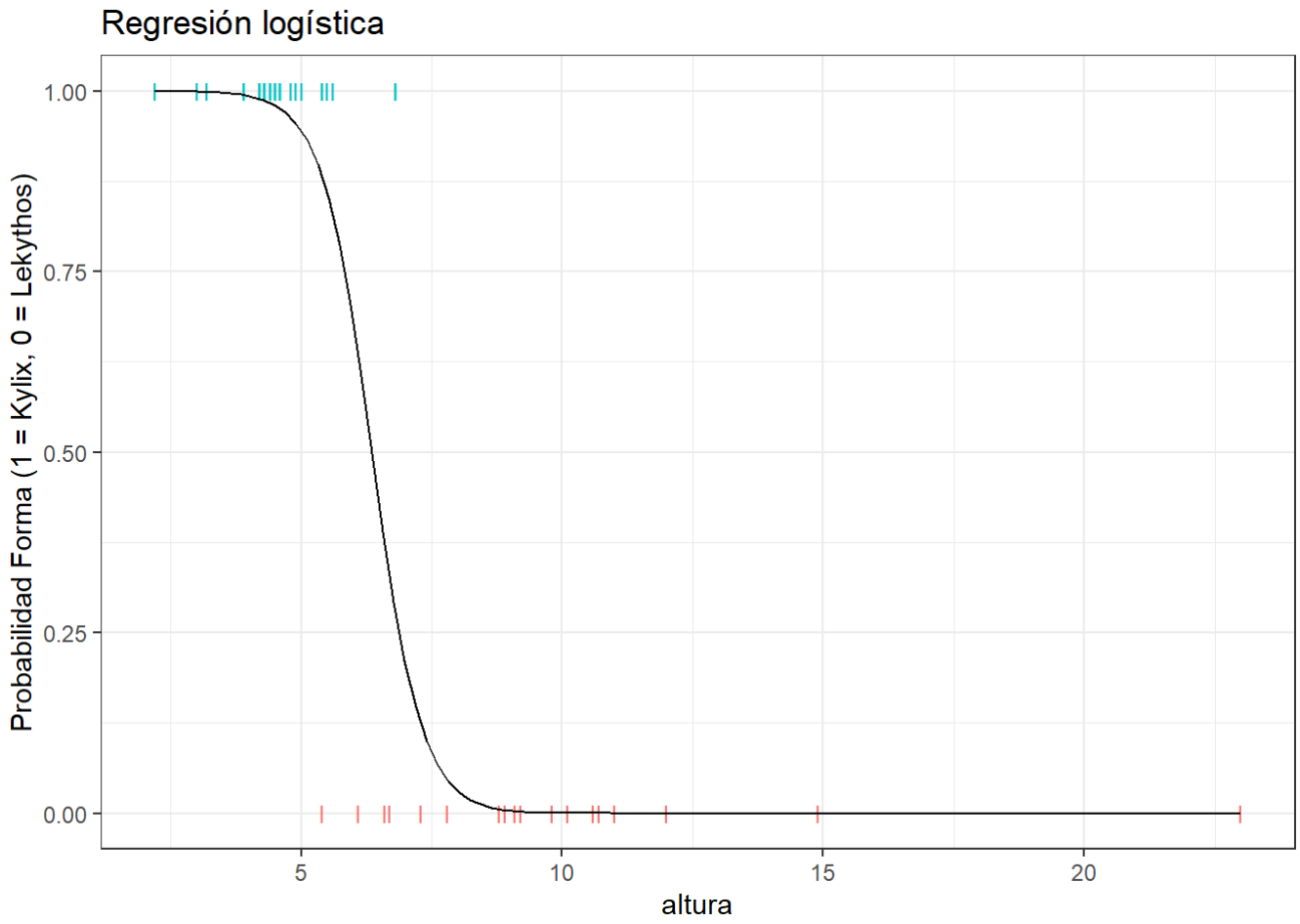



```
#Con plot ()

plot(forma ~ borde, ceramicas, col = "darkblue",
     main = "Modelo regresión logística",
     ylab = "Forma",
     xlab = "Diámetro Borde", pch = "I")
# type = "response" devuelve las predicciones en forma de probabilidad en lugar de en log_ODD
s
curve(predict(modelo, data.frame(borde = x), type = "response"),
      col = "firebrick", lwd = 2.5, add = TRUE)
```

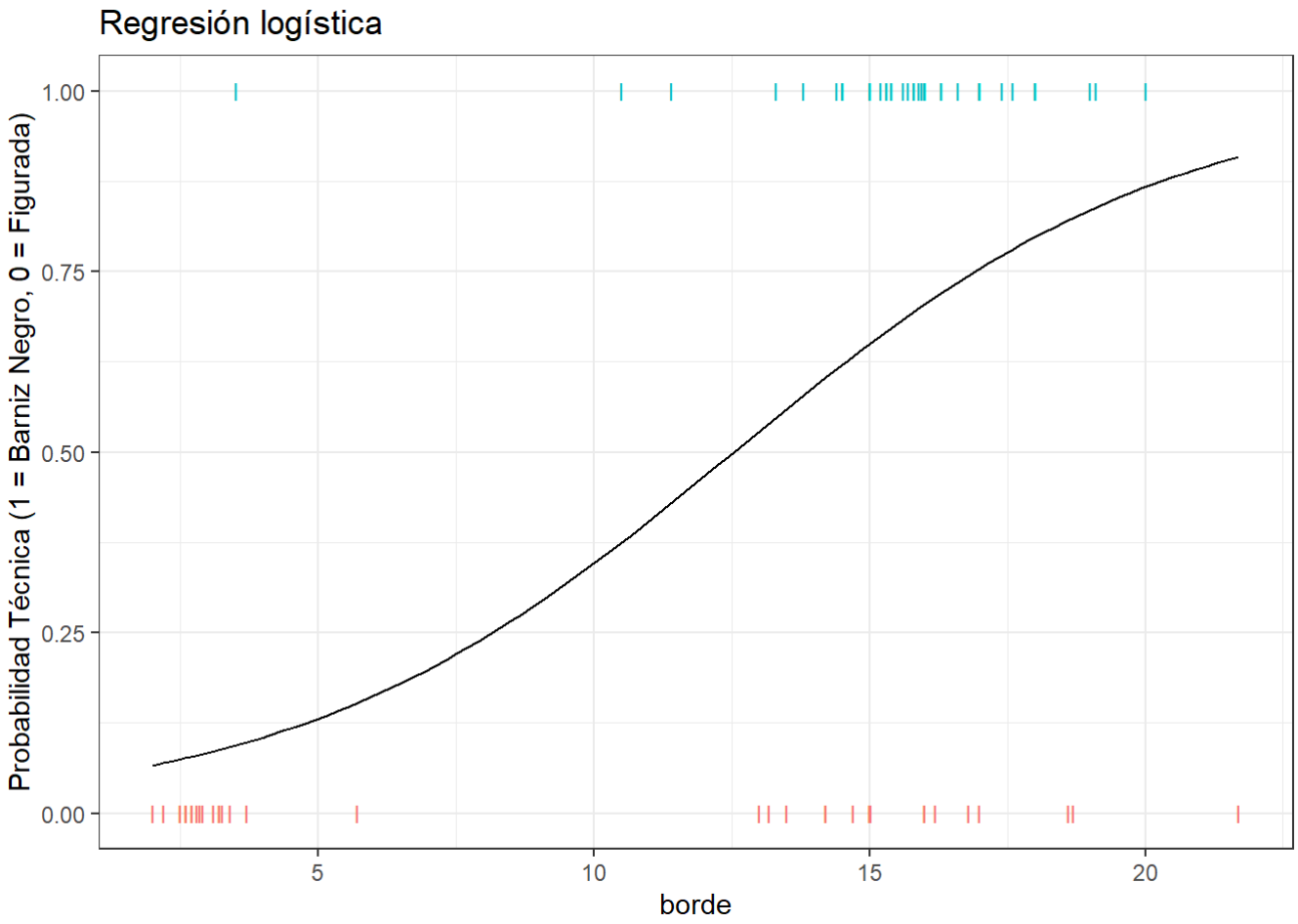


Ahora, podemos hacer lo mismo usando Altura como dependiente, con la variable Tecnica como predictora, y una representación de una regresión logística múltiple, con Forma como predictora y Borde y Altura como dependientes.

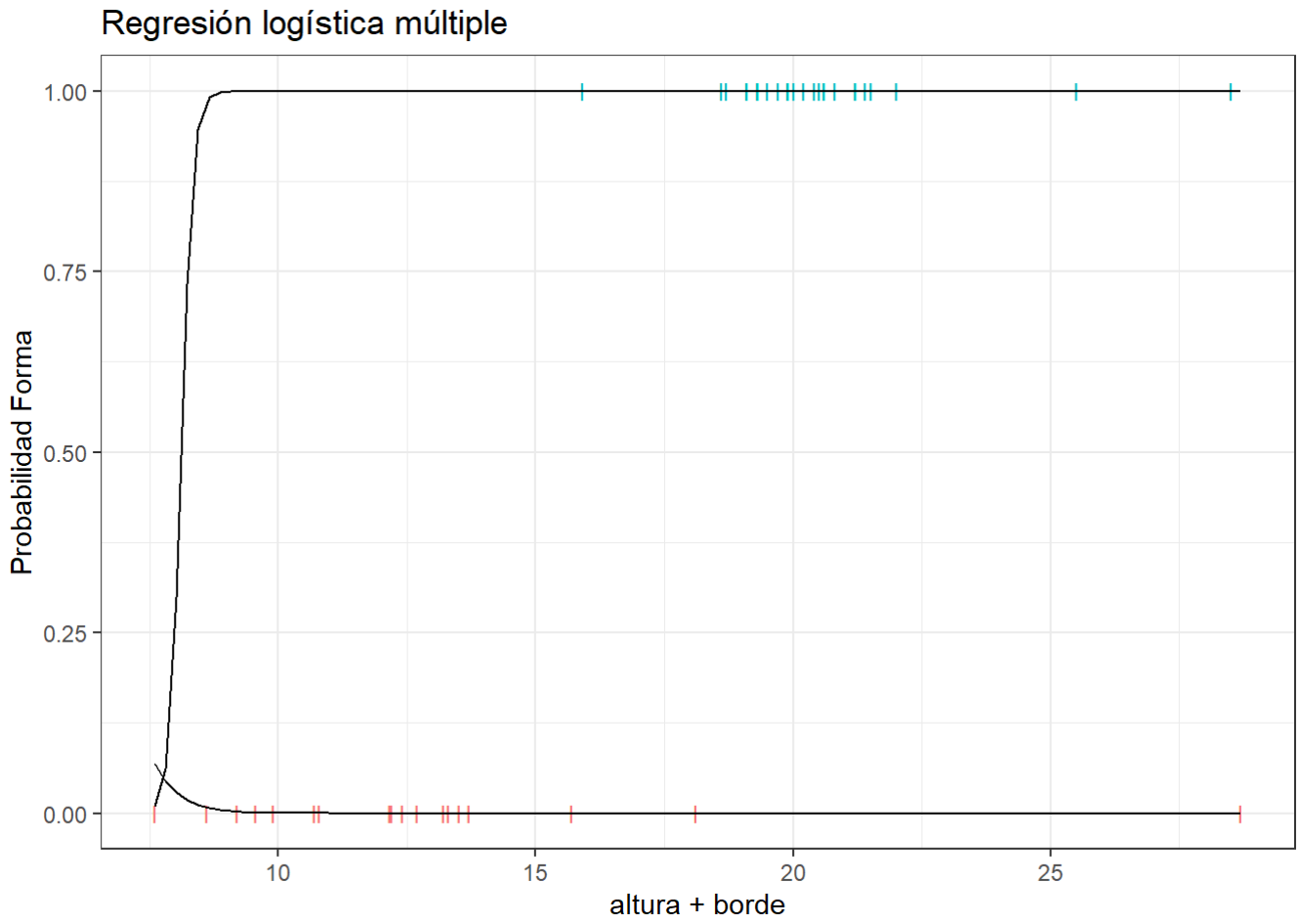
```
#Forma y Altura
ggplot(data = ceramicas, aes(x = altura, y = forma)) +
  geom_point(aes(color = as.factor(forma)), shape = "I", size = 3) +
  stat_function(fun = function(x){predict(modelo1,
                                          newdata = data.frame(altura = x),
                                          type = "response")}) +
  theme_bw() +
  labs(title = "Regresión logística",
       y = "Probabilidad Forma (1 = Kylix, 0 = Lekythos)") +
  theme(legend.position = "none")
```




```
#Técnica y Borde
ggplot(data = ceramicas, aes(x = borde, y = tecnica)) +
  geom_point(aes(color = as.factor(tecnica)), shape = "I", size = 3) +
  stat_function(fun = function(x){predict(modelo_tec,
                                          newdata = data.frame(borde = x),
                                          type = "response")}) +
  theme_bw() +
  labs(title = "Regresión logística",
       y = "Probabilidad Técnica (1 = Barniz Negro, 0 = Figurada)") +
  theme(legend.position = "none")
```




```
#Forma con Borde y Altura


ggplot(data = ceramicas, aes(x = altura + borde, y = forma)) +
  geom_point(aes(color = as.factor(forma)), shape = "I", size = 3) +
    stat_function(fun = function(x){predict(modelo,
                                            newdata = data.frame(borde = x),
                                            type = "response")}) +
  stat_function(fun = function(x){predict(modelo1,
                                            newdata = data.frame(altura = x),
                                            type = "response")}) +
  theme_bw() +
  labs(title = "Regresión logística múltiple",
       y = "Probabilidad Forma") +
  theme(legend.position = "none")
```

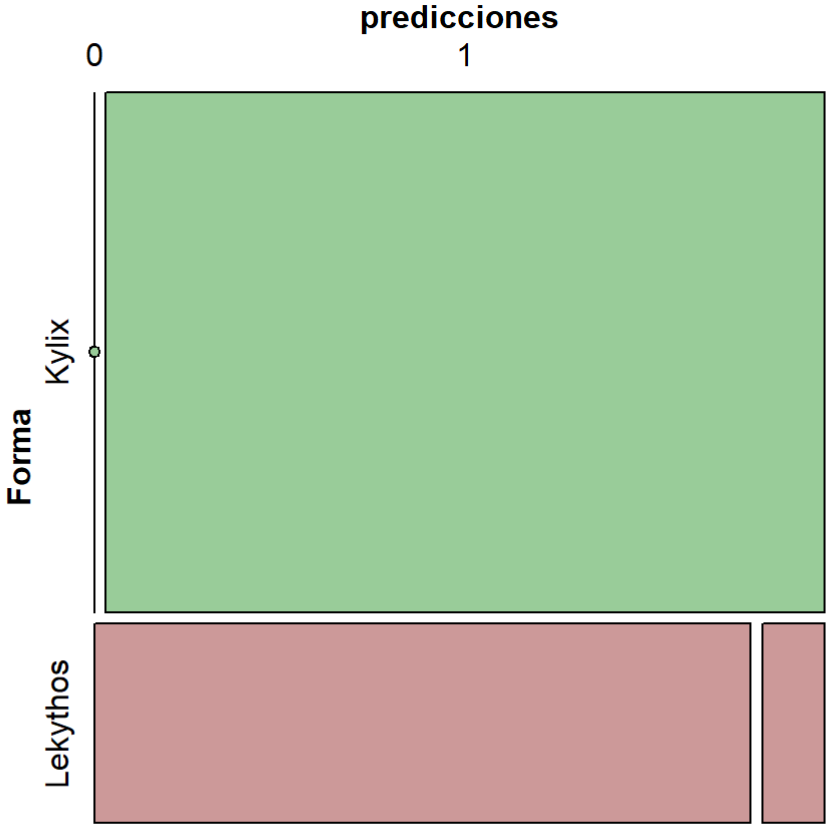


Finalmente, con mosaic () podemos analizar el porcentaje de predicciones correctas, junto al número de falsos positivos y falsos negativos para evaluar el potencial del modelo. Vamos a probar con Forma y Borde, con un threshold de 0,5. Es posible que un modelo prediga mejor una dirección que otra.

```
# Cálculo de la probabilidad del modelo.

newdatafor <- predict(modelo, newdata = data.frame(forma), type = "response")

# Vector de elementos "Lekythos"
pred.modelo <- rep("Lekythos", length(newdatafor))

pred.modelo [newdatafor > 0.5] <- "Kylix"


# Matriz de confusión

matriz <- table(pred.modelo, forma,
dnn = c("Forma", "predicciones"))

matriz
```

```
##           predicciones
## Forma       0  1
##   Kylix     0 60
##   Lekythos 21  2
```

```
library(vcd)
mosaic(matriz, shade = T, colorize = T,
          gp = gpar(fill = matrix(c("#99cc99", "#cc9999"), 2, 2)))
```



Conclusiones

Hemos comprobado (y representado) de forma científica como las medidas morfométricas son relevantes para categorizar cerámicas griegas dentro de categorías, y como, en nuestro caso, el diámetro del borde y la altura nos dicen más que el diámetro de la base, lo cual corrobora lo que se observa en las propias cerámicas a simple vista, pudiendo realizar ahora interpretaciones más objetivas. Pese a que nuestro trabajo se centra en un aspecto muy concreto de unas cerámicas muy concretas (forma y técnica en cerámicas griegas), este método puede (y díria que debe) aplicarse a muchísimos más tipos de cerámicas de cualquier tipo y cronología, así como a otros artefactos y contextos arqueológicos.

Este trabajo plantea un experimento relativamente simple, pero se hacen obvias las inmensas posibilidades que presenta esta metodología. Utilizando bases de datos mucho más amplias y buscando corroborar las hipótesis que abundan en la investigación sobre cerámica griega en la Península Ibérica, se pueden realizar muchos trabajos de gran profundidad que respondan nuestras cuestiones histórico-arqueológicas de forma objetiva y matemática.

Está claro que R debe formar parte del arsenal de herramientas del arqueólogo del s. XXI, para así dejar atrás esa arqueología poco científica, tan ligada al tradicionalismo de las humanidades, que sigue a la orden del día.

Bibliografía consultada para la obtención de datos

Para Ibiza: Base de datos del Museo de Prehistoria de Valencia y base de datos CERES del Museo Arqueológico Nacional.

Para Castellones de Ceal y El Pajarillo: CERES.

Para Cancho Roano: Gracia 2003: Gracia Alonso, F. (2003): “Las cerámicas áticas del palacio-santuario de Cancho Roano”, Cancho Roano VIII. Los materiales arqueológicos I. Celestino Pérez, S. (Ed.). Mérida: IAM-CSIC: 23-194.

Para Mértola: Arruda et al. 1998: Arruda, A. M.; Barros, P. y Lopes, V. (1998): “Cerámicas áticas de Mértola”. Conimbriga. Coimbra 37: 122-149.

Para Castro Marim: Arruda et al. 2020: Arruda, A. M.; Ferreira, D.; Sousa, E. D. (2020). A Cerâmica Grega do Castelo de Castro Marim. Lisboa: UNIARQ. Centro de Arqueologia da Universidade de Lisboa.

Para Cástulo: CERES.

Para Cerro del Castillo: Martín et al. 1995: Martín Ruiz, J. A.; Martín Ruiz, J. M.; García Carretero, J. R. (1995): “Las copas tipo Cástulo del Cerro del Castillo (Fuengirola, Málaga): Una aportación al estudio de su distribución en el área del Estrecho”. En Ripoll Perelló, E. y Ladero Quesada, M. F. (Eds.): Actas del II Congreso Internacional el Estrecho de Gibraltar. UNED, Ayuntamiento de Ceuta: 273-286.

Para Ampurias: Base de datos del Museo de Prehistoria de Valencia. Para El Sec: Arribas et al. 1987: Arribas, A.; Trías, G.; Cerdá, D.; De Hoz, J. (Eds.) (1987): El Barco de El Sec (Costa de Calviá, Mallorca). Estudio de materiales. Ajuntament de Calviá, Universitat de les Illes Balears. Mallorca.

Para Mesas de Asta: obtención propia en el Museo Arqueológico de Jerez de la Frontera.