



## - Trabajo final de Aplicaciones informáticas -

### ***Alumno***

*Manuel Camacho Romero*

### ***Asignatura***

*Aplicaciones Informáticas*

*4º de arqueología*

### ***Profesor***

*Galo Romero García*

# Índice

1.-Introducción.....3

2.-Objetivos.....3

3.-Metodología.....4

4.-Resultados.....6

5.-Conclusión.....7

6.-Bibliografía.....8

## **1.-Introducción.**

En este trabajo se va a llevar a cabo un estudio y recopilación de asentamientos prehistóricos con diferentes tipologías de asentamiento, es decir, desde yacimientos arqueológicos, hasta asentamientos que son conjuntos arqueológicos y cuevas, las cuales son más abundantes. Además de si son de costa o interior. Todo esto para crear un modelo de regresión logística simple de dos factores.

En cuanto al motivo de este pequeño estudio, constaría de ver la respuesta a la siguiente pregunta: ¿Qué probabilidad hay de que sea costa o interior en un contexto arqueológico primitivo?, es decir en contextos prehistóricos, ¿hay mayor número de lugares de estudios situados en costa o es mayor su situación en zonas de interior?

Para todo ello realizaremos una regresión logística en el programa ``R'', el cual ya ha sido enseñado y monitorizado por nuestro profesor de la asignatura de Aplicaciones informáticas, Galo Romero. La finalidad de utilizar este programa será realizar un modelo logístico de dos variables (costa e interior) en relación con el tipo de asentamiento.

Con este estudio obtendremos una ligera idea de qué tipología es más abundante, y si hay una mayor cantidad de yacimientos situados en costa o en interior.

## **2.-Objetivos.**

Cuando se realiza una regresión logística de dos factores, los principales objetivos son los siguientes:

**-Determinar la relación entre las variables predictivas y la variable de respuesta:** La regresión logística te permite evaluar cómo se relacionan dos variables predictivas con una variable de respuesta binaria. El objetivo es determinar si existe una asociación significativa entre los factores y la probabilidad de que ocurra un evento determinado.

**-Identificar la contribución relativa de cada factor predictivo:** La regresión logística permite estimar los coeficientes de los factores predictivos y evaluar su contribución relativa en la predicción de la variable de respuesta. Esto te ayudará a comprender qué factor tiene una influencia más fuerte en la probabilidad del evento.

**-Evaluar la significancia estadística de los factores:** Mediante la regresión logística, puede realizar pruebas de significancia estadística para determinar si los factores predictivos tienen

un impacto significativo en la variable de respuesta. Esto implica examinar los valores p de los coeficientes y establecer un nivel de significancia adecuado.

**-Realizar predicciones y evaluación del modelo:** Después de ajustar el modelo de regresión logística, puede utilizarlo para realizar predicciones sobre la probabilidad de ocurrencias del evento en nuevos datos. También es importante evaluar la calidad del modelo utilizando métricas de evaluación, como la precisión, el área bajo la curva ROC, la sensibilidad y la especificidad.

**-Interpretar los resultados:** Una vez que haya ajustado el modelo de regresión logística, puede interpretar los coeficientes y las probabilidades configuradas para comprender cómo influyen los factores predictivos en la variable de respuesta. Esto te permitirá extraer conclusiones y tomar decisiones informadas basadas en los resultados del modelo.

### 3.-Metodología.

En cuanto a la metodología, los datos recopilados acerca de los asentamientos están almacenados en una tabla en el programa Excel (figura 1), para la posterior creación del modelo de regresión logística en base a estos datos, deberemos recurrir a ``Rstudio'', del que hablaremos un poco.

Asentamiento/ yacimiento arqueológico	¿Costa o interior?	Localización	Tipología de asentamiento	Cronología del yacimiento
Moreaga	Interior	Uribe Kosta (Bizkaia)	Asentamiento	Paleolítico
Errementariena	Interior	Uribe Kosta (Bizkaia)	Asentamiento	Paleolítico
Cueva de Coimbe	Costa	Asturias	Cueva	Paleolítico
Yacimiento de Bolinkoba	Interior	Bizcaia (Lugo)	Yacimiento arqueológico	Paleolítico
Yacimiento de Landaluze	Costa	Bizcaia (Sopela)	Yacimiento arqueológico	Paleolítico
Cueva de El Mazo	Costa	Cantabria	Cueva	Mesolítico
Las Mesas de Salvavieja	Interior	Estepona (Málaga)	Yacimiento arqueológico	Paleolítico
Dólmenes de Antequera	Interior	Málaga	Conjunto arqueológico	Neolítico
Cerro de la Encina	Interior	Granada	Yacimiento arqueológico	Neolítico
Los Millares	Interior	Almería	Asentamiento	Neolítico
Irikaitz	Interior	Guipuzkoa	Yacimiento arqueológico	Paleolítico inferior
Nerja	Costa	Málaga	Cueva	Pleolítico y Neolítico
Tossal	Costa	Alicante, Valencia	Yacimiento arqueológico	Edad del Bronce
Cueva del Higuérón	Costa	Málaga	Cueva	Neolítico
Cueva de los Murciélagos	Costa	Córdoba	Cueva	Neolítico
La Cueva de El Sidrón	Interior	Asturias	Cueva	Paleolítico
Cueva de Tito Bustillo	Costa	Asturias	Cueva	Paleolítico
cueva de El Pindal	Costa	Asturias	Cueva	Paleolítico
Cueva de Cierro	Interior	Asturias	Cueva	Neolítico
cueva sepulcral de Virués-Martínez	Interior	Granada	Cueva	Neolítico

**Figura 1.** Tabla que contiene los datos.

RStudio es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R. Es una herramienta popular entre los científicos de datos, estadísticos y analistas que utilizan R para realizar análisis de datos y crear visualizaciones.

RStudio proporciona una interfaz amigable y funcionalidades específicas para facilitar el trabajo con R. Tiene características como resaltado de sintaxis, autocompletado de código, herramientas de depuración, administración de paquetes, generación de informes y visualización de resultados.

El IDE está diseñado para mejorar la productividad y la eficiencia al desarrollar proyectos de análisis de datos en R. Proporciona una estructura organizada para trabajar en proyectos, lo que permite a los usuarios mantener archivos, scripts y gráficos en un solo lugar.

RStudio es una herramienta de código abierto y existen dos versiones principales disponibles: RStudio Desktop, que se ejecuta localmente en tu computadora, y RStudio Server, que permite acceder a través de un navegador web y trabajar de forma remota en un servidor.

Tras haber descrito en algunos párrafos lo que se denomina como ``Rstudio`` y sus funcionalidades, empezaremos con la regresión logística. La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de un evento binario, es decir, un evento que puede tener solo dos resultados posibles, como "sí/no", "éxito/fracaso" o "1/0". Se utiliza cuando queremos relacionar una variable de resultado binaria con uno o más factores o variables predictoras.

A continuación, explicaré cómo realizar una regresión logística de dos factores. Supongamos que tenemos una variable de resultado binaria llamada "Y" y dos factores o variables predictoras llamadas "X1" y "X2".

**Paso 1:** Preparación de los datos Recopila los datos que necesita para el análisis. Debes tener una muestra de observaciones en las que tengas los valores de "Y", "X1" y "X2" para cada observación. Asegúrese de tener suficientes datos para obtener resultados confiables.

**Paso 2:** Exploración de los datos antes de realizar la regresión logística, es recomendable realizar una exploración de los datos para familiarizarte con su distribución y características. Puede realizar gráficos y cálculos descriptivos para obtener información preliminar sobre los datos.

**Paso 3:** Codificación de variables predictoras Si tus variables predictoras son categóricas, es necesario codificarlas en variables numéricas antes de realizar la regresión logística. Puede utilizar técnicas como la codificación one-hot, donde se crea una variable binaria para cada

categoría posible. Si tus variables predictoras ya son numéricas, no necesitarás realizar esta codificación.

**Paso 4:** Ajuste del modelo El siguiente paso es ajustar el modelo de regresión logística. En este caso, al tener dos factores, el modelo tendrá la siguiente forma:

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Donde "p" representa la probabilidad de que el evento binario desaparezca, " $\beta_0$ " es el intercepto o término constante, " $\beta_1$ " y " $\beta_2$ " son los coeficientes asociados a las variables predictoras "X1" y "X2" respectivamente. El término "logit(p)" se refiere al logaritmo de la razón de probabilidad (odds ratio) y se utiliza para linealizar la relación entre las variables predictoras y la probabilidad.

**Paso 5:** Estimación de coeficientes Para estimar los coeficientes del modelo, se utiliza una técnica llamada máxima verosimilitud. El objetivo es encontrar los valores de los coeficientes que maximicen la verosimilitud de los datos observados bajo el modelo. Los coeficientes estimados se denotan como " $b_0$ ", " $b_1$ " y " $b_2$ ".

**Paso 6:** Interpretación de los coeficientes Una vez que obtengas los coeficientes estimados, puedes interpretarlos para entender cómo disminuir las variables predictoras a la probabilidad del evento binario. Un coeficiente positivo indica que a medida que aumenta el valor de la variable predictora, la probabilidad del evento también aumenta, mientras que un coeficiente negativo indica una relación inversa.

**Paso 7:** Evaluación del modelo Es importante evaluar la calidad del modelo ajustado. Puede utilizar medidas de ajuste como el coeficiente de determinación ( $R^2$ ),

## 4.-Resultados.

-En primer lugar descargamos con el comando `library()` los paquetes necesarios para la realización del modelo.

```
library ( ggplot2 )
```

```
library ( aod )
```

-Prosiguiendo, se crean los conjuntos de datos que contienen la información. Primero ha de asegurarse de tener un conjunto de datos que contenga una variable dependiente binaria y dos variables independientes.

```
datos <- data.frame(
```

```
nº = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20),
```

```
situacion = c(0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0),
```

```
tipo = c(1, 1, 2, 3, 3, 2, 3, 4, 3, 5, 3, 2, 3, 2, 2, 2, 2, 2, 2, 2))
```

-Ajusta un modelo de regresión logística simple utilizando la función ``glm`` con la familia binomial.

```
modelo_logistico <- glm(nº ~ situacion + tipo, data = datos, family = "binomial")
```

-Representa gráficamente el modelo de regresión logística utilizando ggplot2.

-Instalar ggplot2 si aún no está instalado

-Para la representación gráfica del modelo.

```
ggplot(data = datos, aes(x = situacion, y = nº)) +
```

```
geom_point(aes(color = as.factor(situacion)), shape = 1) +
```

```
stat_function(fun = function(x){predict(modelo_logistico,
```

```
newdata = data.frame(situacion = x),
```

```
type = "response"))}) +
```

```
theme_bw() +
```

```
labs(title = "Regresión logística",
```

```
y = "Probabilidad balance") +
```

```
theme(legend.position = "none")
```

## 5.-Conclusión.

Tras realizar los comandos, como resultado ha dado como conclusión que las cuevas son la tipología más abundante en la costa, mientras que los yacimientos arqueológicos, los conjuntos arqueológicos y los asentamientos son mucho más resaltantes en interior. También cabe a destacar que en las cuevas también encontramos algunas de interior.

Como conclusión comentar que este estudio me ha dado la oportunidad de poder observar más posibilidades, siendo de gran utilidad para este tipo de estudios. De esta forma la implantación de este tipo de herramientas como ``Rstudio`` en la arqueología es realmente útil para la representación y gestión de datos de este tipo, además de el mismo manejo de ellos.

## 6.-Bibliografía.

-Alonso, D. Á., de los Terreros, J. Y. S., de Andrés Herrero, M., Valbuena, Á. A., Pardo, J. F. J., & Hernández, J. A. R. (2013). La secuencia cronoestratigráfica del Paleolítico superior de la cueva de Coímbre (Asturias, España). In *El Cuaternario Ibérico: investigación en el s. XXI: VIII Reunión de Cuaternario Ibérico (Sevilla-La Rinconada 2013)* (pp. 83-86). Asociación Española para el Estudio del Cuaternario.

-Lopez, P., & Cacho, C. (1979). La cueva del Higuerón (Málaga): Estudio de sus materiales. *Trabajos de Prehistoria*, 36, 11.

-Tortosa, J. E. A., García, E. B., Borja, P. G., Puchol, O. G., Benito, J. L. P., Jordá, G. P., ... & PardoB, J. F. J. (2005). Cueva de Nerja (Málaga). Los niveles neolíticos de la Sala del Vestíbulo. In (coord.), *Actas del III congreso del Neolítico en la Península Ibérica (Santander 5 a 8 de octubre de 2003), Santander* (pp. 975-987).

-Alonso, D. Á., & de Andrés Herrero, M. (2012). La transición solutrense-magdalenense en la cueva de El Cierro (Ribadesella, Asturias, España). *Espacio Tiempo y Forma. Serie I, Prehistoria y Arqueología*, (5).

-Fortea Pérez, J., Rasilla Vives, M. D. L., Martínez García, E., Sánchez-Moral, S., Cañaveras Jiménez, J. C., Cuezva Robleño, S.,... & Alonso Peña, J. (2007). La Cueva de El Sidrón (Borines, Piloña, Asturias). Campañas Arqueológicas de 2000 a 2002.

-Moure Romanillo, A. (1990). La cueva de Tito Bustillo (Ribadesella, Asturias): el yacimiento paleolítico. *Excavaciones arqueológicas en Asturias 1983-86*.

## -Webgrafía:

-<https://rpubs.com/>

-<https://github.com/>