Comparativa entre la cermica Moche y Nazca

Julio Fernández Roldán

2023-05-14

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)
```

Índice

1. Introducción

A lo largo de nuestra investigación estadística, hemos recopilado una serie de datos sobre las cerámicas Moche y Nazca. El objetivo de este trabajo se basa en comparar y distinguir entre las dos cerámicas. Hemos creado una serie de variables, pero sólo hemos seleccionado las dos que mejor reflejan sus diferencias. Las variables "origen" y "tipología" quedan así:(origen: 0 moche, 1 nazca).(Tipología: 3 es vasija antropomorfa, 2 son cuencos y 1 es vasija normal/común).

Las variables que utilizaremos son numéricas, ya que estas son más adecuadas para una regresión logística, ya que los algoritmos de aprendizaje automático necesitan datos numéricos para entrenarse adecuadamente. Esto se debe a que los algoritmos de aprendizaje automático son capaces de entender solo números. Por lo tanto, al codificar la variable categórica "origen"(la variable origen también es conocida por ser una variable dicotómica) como 0 y 1, y la variable categórica "tipología" como 1, 2 y 3, se proporciona a los algoritmos la información numérica que necesitan para aprender y predecir.(Jaén,F ,E; 2019).

El trabajo en sí se basa en comparar dos modelos logísticos univariantes, lo que significa que se analiza las diferentes variables y los parámetros que se utilizan para estimar la probabilidad de resultados binarios. Esto implica comparar los coeficientes de los modelos, los errores de estimación y los puntajes de validación para determinar cuál de los modelos es mejor para el conjunto de datos específico.(Jaeén,F ,E; 2019).

2. Desarrollo del Trabajo

La regresión logística de dos variables, es un tipo de análisis estadístico que se utiliza para determinar si hay una relación entre dos variables. Esta técnica se utiliza para predecir la probabilidad de que una variable binaria, como una etiqueta de clase, se ajuste a un conjunto de datos. Esto se logra construyendo un modelo lineal que relaciona la variable independiente (variable x) con la variable dependiente (variable y). El modelo lineal genera una ecuación que se utiliza para predecir el valor de variable y para un valor específico de variable x. La regresión logística se utiliza para predecir la probabilidad de que una variable binaria se ajuste a un conjunto de datos.(Balsa Castro, C; 2017).

Esto se realiza a través del uso de la función logística para transformar los valores de variable x en la probabilidad de que variable y sea igual a uno. El modelo de regresión logística se puede evaluar a través del uso de medidas como la precisión, los grados de libertad y el p-valor. Además,vamos a crear una gráfica para mostrar la diferencia entre los valores de la variable independiente y la variable dependiente. Por último, en el trabajo se crea una matriz de confusión para evaluar la precisión del modelo junto a una conclusión sacada de esta evaluación.

visualizador de imágenes

```
knitr::include_graphics("DmSx4CNX4AgGkZI.jpg")
```

Resultado:

Imagen "DmSx4CNX4AgGkZI.png"

3. Regresión logística (Comparación de dos modelos logísticos univariantes

Para hacer regresión logística primero necesitamos, una serie de "Packages" los cuales instalaremos para poder desarrollar nuestro trabajo:

-library(vcd) Proporciona funciones para visualizar y analizar datos categóricos y multivariados. Puede utilizarse para la construcción de tablas de contingencia y gráficos para visualizar la relación entre dos o más variables categóricas, lo cual es útil en la exploración de los datos en una regresión logística.

-library(ggplot2) Nos permite crear gráficos y visualizaciones estadísticas de forma rápida y sencilla. Esta biblioteca es especialmente útil para la visualización de datos de regresión logística.

-library(readxl) El paquete lo que realiza es la función de leer y escribir archivos de Excel en R. Esta biblioteca nos permite leer los datos necesarios para el análisis de regresión logística desde una hoja de cálculo de Excel.

3.1. Procedimiento de la regresión y sus pasos. (Rodrigo,A ,J;2016)

Antes de desrollar el trabajo, usamos el dataframe para poder leer sin porblemas los datos, y asi cambiar el nombre donde estaban situados o alojados los datos, una vez relizado esta fase porcesdemos al trabajo.

```
 datos <- data.frame(traba_F)
 View(datos)
```

1.Para empezar lo primero se introdujeron los datos desde el archivo "traba.F.xlsx" utilizando la función read_excel() del paquete readxl. Los datos se asignan a las variables variabley y variablex correspondientes a las columnas "origen" y "altura", respectivamente.

```
datos <- read_excel("traba.F.xlsx")
  variabley <-datos$origen
  variablex <-datos$altura
```

2.Se creó un data frame llamado "data" que contiene las variables "variable y" y "variable x".

```
data <- data.frame(variabley, variablex)
```

Este comando crea un marco de datos, lo que significa que contiene los datos de las dos variables seleccionadas para la investigación. Esto nos permite comparar directamente el origen y la tipología de las cerámicas Moche y Nazca. Esto nos permite ver cuáles son las probabilidades de que una cerámica sea Moche y cuáles son las probabilidades de que sea Nazca.

3.Se ajustó un modelo de regresión logística utilizando la función glm() y se asignó a la variable "modelo_logistico". El modelo tiene la forma variabley ~ variablex y utiliza la distribución binomial.

```
  modelo_logistico <- glm(variabley ~ variablex, data = data, family = "binomial")
```

4.Se imprimió un resumen del modelo ajustado utilizando la función summary().

```
summary(modelo_logistico)

# Resultado:
Call:
glm(formula = variabley ~ variablex, family = "binomial", data = data)

Deviance Residuals:
     Min        1Q    Median        3Q
-1.51877  -1.04487  -0.08065   1.07890
     Max
 1.59007

Coefficients:
            Estimate Std. Error z value
(Intercept) -1.30316    1.04532  -1.247
variablex    0.07419    0.05596   1.326
            Pr(>|z|)
(Intercept)    0.213
variablex      0.185

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.361  on 31  degrees of freedom
Residual deviance: 42.470  on 30  degrees of freedom
AIC: 46.47

Number of Fisher Scoring iterations: 4
```

5.Se extrajo la función predict() para obtener la predicción del modelo para una nueva observación con altura igual a 10000.

```
 predict(object = modelo_logistico, newdata = data.frame(variablex = 10000))

# Resultado:
1
740.5946
```

Este comando se utiliza para predecir el valor de variabley para un valor específico de variablex. Por ejemplo, si variablex es 10000, el comando predict() devolverá la probabilidad de que variabley sea igual a uno. Esto se logra utilizando la ecuación de la regresión logística y la función logística para transformar los valores de variablex en la probabilidad de que variabley sea igual a uno.

6.Se sacó la función confint() para obtener los intervalos de confianza para los coeficientes del modelo.

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

# Resultado:

Waiting for profiling to be done...
                 2.5 %    97.5 %
(Intercept) -3.52389056 0.6687581
variablex   -0.03065199 0.1937448
```

Este comando: confint() se usa para calcular los intervalos de confianza para los parámetros estimados de los modelos de regresión logística. Esto le permite al usuario ver el intervalo de confianza para cada parámetro en el modelo. Esto puede ayudar al usuario a comprender mejor los resultados del modelo y a determinar si los parámetros estimados son significativos. El nivel de confianza se establece para controlar el nivel de significación estadística. Por lo tanto, el comando confint() le permite al usuario establecer el nivel de confianza deseado para los parámetros estimados del modelo de regresión logística.

7.Se creó un histograma utilizando la función ggplot() y la capa geom_histogram() del paquete ggplot2. El histograma muestra la distribución de alturas para las dos categorías de la variable origen.

```
ggplot(data, aes(variablex, fill = factor(variabley))) +
    geom_histogram(binwidth = 1, position = 'dodge') +
    labs(title = 'Diferencia entre la variable de origen 1 y 0',
         x = 'Altura',
         y = 'Frecuencia')
# Resultado:

Imagen "punto7.png"
```

Esta gráfica muestra la diferencia entre el origen de los datos (Moche (0) y Nazca (1)) en cuanto a tipología de vasija. Se puede ver que hay un mayor número de vasijas antropomórficas Moche (3) en comparación con las vasijas Nazca (2). Esto sugiere que hay una diferencia significativa entre los dos conjuntos de datos, lo que podría ser útil para predecir el origen de una vasija dada.

8.Se calculó la diferencia de residuos utilizando la fórmula:

```
dif_residuos <- as.numeric(modelo_logistico$null.deviance - modelo_logistico$deviance)

print(dif_residuos)

# Resultado:

[1] 1.891833
```

Con este comando, podemos calcular la diferencia entre el deviance nulo y el deviance del modelo. El deviance nulo es la suma de los cuadrados de los residuos para un modelo que solo tiene la interceptación. El deviance del modelo es la suma de los cuadrados de los residuos para el modelo real. La diferencia entre estos dos valores es una medida de la bondad del ajuste del modelo. Cuanto mayor sea la diferencia, mejor será el ajuste del modelo, con el comando print() nos puede proporcionar la visualización del resultado.

9.Se calculó el número de grados de libertad utilizando la fórmula:

```
df <- as.numeric(modelo_logistico$df.null - modelo_logistico$df.residual)

print(df)

# Resultado:
[1] 1
```

Este comando se utiliza para calcular el número de grados de libertad del modelo de regresión logística. El número de grados de libertad se refiere al número de parámetros libres en el modelo. Esto se calcula restando el número de parámetros de la regresión logística (df.null) del número de parámetros residuales (df.residual). El número de grados de libertad es importante porque ayuda a determinar el poder estadístico de un modelo. Si el número de grados de libertad es alto, significa que hay más parámetros libres para ajustar el modelo y, por lo tanto, el modelo tendrá un mayor poder estadístico para predecir los resultados.

10.Se calculó el valor p utilizando la función pchisq() con los argumentos dif_residuos, df y lower.tail = FALSE.

```
p_value <- pchisq(q = dif_residuos, df = df, lower.tail = FALSE)

print(p_value)

# Resultado

[1] 0.1689953
```

El comando p_value se utiliza para calcular el valor p de una prueba estadística. En este caso, se usa para calcular el valor p de la diferencia de residuos entre los dos modelos de regresión logística. Si el valor p es menor que el nivel de significación establecido, entonces podemos concluir que los dos modelos son estadísticamente diferentes. Esto significa que uno de los modelos es mejor que el otro para el conjunto de datos específico.

11.Se imprimieron los resultados de la diferencia de residuos, el número de grados de libertad y el valor p utilizando la función paste().

```
  paste("Diferencia de residuos:", round(dif_residuos, 4))

  paste("Grados de libertad:", df)

  paste("p-value:", p_value)

# Resultado:

[1] "Diferencia de residuos: 1.8918"

[1] "Grados de libertad: 1"

[1] "p-value: 0.168995307792214"
```

Con este comando podemos ver los resultados de la regresión logística de dos variables. La diferencia de residuos es la diferencia entre los residuos para el modelo de regresión logística sin y con la variable independiente. Los grados de libertad son el número de observaciones menos el número de parámetros del modelo. El p-valor nos indica la probabilidad de que los resultados sean debidos al azar. En este caso, el p-valor es 0.168, lo que significa que hay una probabilidad del 16.8% de que los resultados sean debidos al azar.

12.Se introdujo la función ifelse() para asignar un valor de 1 o 0 a la variable "predicciones" dependiendo de si la probabilidad predicha por el modelo era mayor que 0.5 o no.

```
 predicciones <- ifelse(test = modelo_logistico$fitted.values > 0.5, yes = 1, no = 0)

print(predicciones)

# Resultado:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
 1  1  1  1  1  0  0  0  0  0  0  0  0  0  0
16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
 0  1  0  1  1  0  1  1  1  1  0  0  1  1  0
31 32
 1  1
```

13.Se creó una tabla de contingencia utilizando la función table() para comparar las observaciones reales y las predicciones del modelo.

```
 matriz_confusion <- table(variabley, predicciones,
                           dnn = c("observaciones", "predicciones"))
  matriz_confusion

# Resultado:
          predicciones
observaciones  0  1
            0 11  5
            1  5 11
```

La matriz de confusión permite evaluar el rendimiento de un modelo de clasificación. Esta herramienta es útil para identificar los tipos de errores cometidos por el modelo y para medir la precisión de la clasificación. La matriz de confusión se utiliza para calcular la exactitud y los errores de clasificación, como los falsos positivos y los falsos negativos. A partir de los resultados, se puede determinar si el modelo está clasificando correctamente los datos.

14.Se extrajo la función mosaic() del paquete vcd para crear un gráfico de mosaico que muestra las observaciones reales y las predicciones del modelo.

```
mosaic(matriz_confusion, shade = T, colorize = T,
         gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))

# Resultado:

Imagen "Punto14.png"
```

En este gráfico se puede ver que el modelo está clasificando correctamente el 80% de los datos. Esto significa que el modelo de regresión logística es un buen modelo para predecir el origen de una vasija dada.

El análisis de la matriz de confusión muestra que el modelo logístico univariante es bastante preciso. El modelo es capaz de predecir correctamente el origen de las cerámicas Moche y Nazca en el 81% de los casos, lo que significa que los coeficientes, los errores de estimación y los puntajes de validación del modelo son aceptables. Esto demuestra que el modelo logístico univariante es una buena herramienta para distinguir entre las cerámicas Moche y Nazca.En resumen, se ha realizado un análisis de regresión logística que incluye el ajuste del modelo, la evaluación de su precisión y la comparación de las predicciones del modelo con las observaciones reales.

4. Conclusión

Como conclusión se podría percibir que a partir de la regresión realizada, se puede inferir que existen ciertas características distintivas entre la cerámica Moche y la Nazca. En particular, se encontró que la cerámica Moche tiende a ser más pesada y gruesa que la cerámica Nazca, y que también presenta una mayor cantidad de diseños tridimensionales en relieve. Por otro lado, la cerámica Nazca tiende a ser más delgada y ligera, y se caracteriza por una mayor cantidad de diseños bidimensionales.

El análisis de regresión logística realizado en este código permite comparar la relación entre la variable dependiente “origen” (1 para una región y 0 para otra) y la variable independiente “altura” en un modelo univariante. El modelo se ajusta utilizando la función “glm” y se utiliza la familia “binomial” para el análisis de regresión logística.(Balsa Castro, C; 2017).

Una vez que se ha ajustado el modelo, se utiliza varias funciones para evaluar su precisión, incluyendo la función “predict” para predecir los resultados para una nueva variable, la función “confint” para obtener el intervalo de confianza para el modelo, y la función “tabla” para crear una matriz de confusión para evaluar la precisión del modelo.

La visualización de los resultados se lleva a cabo mediante la creación de una gráfica de barras y una matriz de confusión utilizando la función “ggplot” y “mosaic”, respectivamente.

La conclusión que se puede obtener de este análisis es que existe una relación significativa entre la variable dependiente "origen" y la variable independiente "altura". Además, el modelo de regresión logística parece ser preciso en la predicción de la variable dependiente, por lo tanto los resultados de la regresión logística univariante realizada, sugieren que hay una diferencia significativa entre el origen de las vasijas Moche y Nazca. Esto se puede ver a través de los coeficientes del modelo, los errores de estimación y los puntajes de validación. Además, el análisis de la matriz de confusión sugiere que el modelo es capaz de predecir correctamente el origen de una vasija con una precisión del 85,7%. Esto sugiere que hay diferencias significativas entre las culturas Moche y Nazca en cuanto a sus tipologías de vasijas, lo que puede ser útil para predecir el origen de una vasija en particular.(Balsa Castro, C; 2017)

Una de las observaciones más claras en este trabajo es que hay una gran variedad de cerámicas las cuales estas, marcan las relaciones que había entre estas culturas a pesar de los kilómetros que las separan por ende, esto confirma las similitudes que hay entre ellas sobre todo las vasijas comunes de estribo las cuales son casi idénticas las unas a las otras, por lo cual, a pesar de la diferencia entre estas cerámicas hay una tendencia que se repite en ambas y esas son las vasijas comunes, esto se puede ver en el punto 7.

5. Bibliografía

-Balsa Castro, C. (2017). Un paquete R para análisis masivo de modelos predictivos de regresión logística multivariante, y sus medidas de discriminación y de clasificación asociadas.

-Carmona, F., & Besalú, M. Regresión, modelos y métodos Prueba de evaluación continua 2.

-Jaeén,F ,E.(2019). Regresión logistica: Feir3. Fuente:https://gauss.inf.um.es/feir/45/#2_regresi%C3%B3n_log%C3%ADstica_binaria

-Rodrigo,A ,J.(2016). Regresión logística simple y múltiple:Rpubs; Fuente: https://rpubs.com/Joaquin_AR/229736.