Regresión lineal simple

Manuel Martín

2024-03-12

### Ejercicio 1

Si, en ello se basa la disciplina arqueológica para conocer y explicar los fenómenos y/o sucesos historicos. Sin embargo, una explicación completa de estos fenómenos y/o sucesos requiere en muchos casos en aproximaciones pluridisciplinares, que no solo analizan los restos materiales, sino también los comportamientos, la información textual o los condicionantes climáticos, por mencionar algunos.

### Ejercicio 2

El análisis de Pearson no establece relaciones causa-efecto entre las variables analizadas sino que indica la fuerza y dirección de la relación lineal entre dos variables. Por lo que este análisis no implica causalidad entre las variables analizadas.

### Ejercicio 3

La causalidad es una relacion causa-efecto en la que un evento/suceso provoca como resultado un evento/suceso. Un ejemplo de este tipo de relaciones de causalidad en historia y arqueología podría ser: La conquista del 711 (causa) supuso la islamización y arabización de la población (efecto).

### Ejercicio 4

Los parámetos que determinan una regresión lineal son la pendiente y la ordenada en el origen, también denominada intersección.

### Ejercicio 5

No, porque en un plano cartesiano el eje “x” es el de abscisas y el eje “y” es el de ordenadas.

### Ejercicio 6

Por un lado, la recta de regresión ilustra la relación entre una variable independiente y otra dependiente en un modelo de regresión lineal. Por su parte el plano de regresión es una superficie de tres dimensiones en la que lo que se representa es la relacion entre dos o más variables independientes y una dependiente, en un modelo en este caso de regresión múltiple.

### Ejercicio 7

Estos supuestos son cuarto: linealidad, homocedasticidad, normalidad e independencia.

### Ejercicio 8

#Datos  
distancia <- c( 1.1,100.2,90.3,5.4,57.5,6.6,34.7,65.8,57.9,86.1)  
cuentas <- c(110,2,6,98,40,94,31,5,8,10)  
#Creamos un data frame  
datos <- data.frame(distancia, cuentas)  
#Ajustamos para que sea una regresión lineal  
recta <- lm(cuentas ~ distancia, data = datos)  
summary(recta)

##   
## Call:  
## lm(formula = cuentas ~ distancia, data = datos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.644 -12.672 7.693 8.730 15.825   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 95.3710 9.7188 9.813 9.77e-06 \*\*\*  
## distancia -1.0872 0.1579 -6.885 0.000126 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.52 on 8 degrees of freedom  
## Multiple R-squared: 0.8556, Adjusted R-squared: 0.8375   
## F-statistic: 47.4 on 1 and 8 DF, p-value: 0.0001265

En el resumen del modelo, podemos conocer cual es la pendiente y la intersección de nuestra recta. La pendiente es el coeficiente para nuestra variable independiente (distancia) en este caso -1.0872, mientras que la intersección es el coeficiente independiente (cuentas), en este caso 95.3710

### Ejercicio 9

Estos parámetros de la ecuación de regresión representan su pendiente (-1.0872) y su intersección (95.3710) y describen como varia la variable dependiente según lo hace la independiente.

### Ejercicio 10

Un intercepto o intersección = 0 implica que en el momento que la variable independiente es 0 la dependiente también es 0

### Ejercicio 11

Para los análisis de regresión lineal se utiliza una minimización de los errores cuadráticos para así calcular los valores de la recta de regresión.

### Ejercicio 12

Para establecer el error asociado a nuestro cálculo de cuentas para el yacimiento situado a 1.1 km necesitaríamos comparar nuestra predicción (94.17507) con el valor real. Pero como no lo tenemos, no podemos calcular el error de manera directa. Por eso calculamos el error estándar de nuestra predicció.

#Calculamos el número de cuentas para este yacimiento   
distancia\_nueva <- 1.1  
prediccion <- predict(recta, newdata = data.frame(distancia = distancia\_nueva))  
print(prediccion)

## 1   
## 94.17507

#calculamos el error estándar de esta predicción  
prediccion\_error <- predict(recta, newdata = data.frame(distancia = distancia\_nueva), se.fit = TRUE)$se.fit  
print(prediccion\_error)

## [1] 9.576585

Al ser nuestro error estándar 9.576585, significa que en promedio podemos esperar que las predicciones de el modelo varíen en torno a +- 9.576585 unidades del valor real de la variable dependiente.

### Ejercicio 13

Para calcular los residuos del modelo, restamos las predicciones del modelo al valor real de las cuentas.

cuentas\_prediccion <- c(6, 98, 40, 94, 31, 5, 8, 10)  
predicciones\_cuentas <- c(-6.682842, 85.520196, 28.938591, 84.216973, 53.69983, 19.924631, 28.504183, -2.121561)  
residuos <- cuentas\_prediccion - predicciones\_cuentas  
print(residuos)

## [1] 12.682842 12.479804 11.061409 9.783027 -22.699830 -14.924631 -20.504183  
## [8] 12.121561

### Ejercicio 14

Para comprobar si se cumple el supuesto de normalidad empleamos la prueba de normalidad de Shapiro-Wilk.

shapiro.test(residuos)

##   
## Shapiro-Wilk normality test  
##   
## data: residuos  
## W = 0.72999, p-value = 0.004895

El estadistico de la prueba da W = 0.72999 y el p-value = 0.004895. Por lo que al ser inferior a 0.05, podemos establecer que los datos residuales no están distribuidos de manera normal.

### Ejercicio 15

Los dos conjuntos de datos empleados para la modelización lineal son los de entrenamiento y los de prueba. Siendo imprescindible en su preparación establecer una semilla que permita generar los datos de manera semialeatoria.

set.seed(123)   
indices\_entrenamiento <- sample(1:length(cuentas), 0.7 \* length(cuentas))   
datos\_entrenamiento <- datos[indices\_entrenamiento, ]  
datos\_prueba <- datos[-indices\_entrenamiento, ]

### Ejercicio 16

library(caret)

## Warning: package 'caret' was built under R version 4.3.3

## Loading required package: ggplot2

## Loading required package: lattice

control <- trainControl(method = "cv", number = 5)  
modelo\_cv <- train(cuentas ~ ., data = datos, method = "lm", trControl = control)  
print(modelo\_cv)

## Linear Regression   
##   
## 10 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 8, 8, 8, 8, 8   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 17.32892 1 15.87944  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Ilustramos de esta manera cual es el rendimiento de nuestro modelo de regresión lineal cuando empleamos validación cruzada con 5 pliegues.

Desgranando los resultados el RMSE se refiere a una medida del precisión del modelo, cuanto más bajo sea mejor capacidad de predicción tendrá el modelo.

Rsquared mide como las variables predictoras son capaces de explicar la variabilidad de la variable respuesta (dependiente). En este caso el valor es aproximadamente 1, lo que implica que el modelo es capaz de explicar la toda la variabildiad presente en los datos.

MAE es otra medida de precisión del modelo, de nuevo cuanto menor sea, mayor será la capacidad predictiva del modelo.

### Ejercicio 17

Si los coeficientes de regresión se han calculado con un intervalo de confianza del 95% significa que la probabilidad de que la correlación lineal entre los coeficientes de regresión y la variable de respuesta se deba al azar es de un 5% o lo que es lo mismo un nivel de significación de 0.05. Si el nivel de significación es 0.01 los coeficientes de regresión se habrán calculado con un intervalo de confianza del 99%

### Ejercicio 18

Si las estimaciones en un modelo lineal son menos precisas para un rango de valores en comparación a otro, hablamos de que hay indicios de heterocedasticidad.

### Ejercicio 19

El coeficiente de determinación (R-cuadrado), que es una medida que examina hasta que punto las diferencias en una variable se pueden explicar por la diferencia en una segunda variable.

### Ejercicio 20

Al hablar de una observación atípica hablamos de un valor que es notablemente diferente del resto de datos y que puede provocar distorsiones del modelo. Mientras que hablamos de apalancamiento cuando se trata de valores muy altos o bajos en las variables predictoras y que pueden tener como resultado un impacto muy marcado en la estimación de coeficientes de nuestro modelo.