Econometria

Agustin Riquelme y Heriberto Espino

2023-08-30

# Regresion lineal simple

Se muestran las calificaciones del examen de matemáticas de 10 estudiantes de primer año de universidad, junto con sus calificaciones finales en cálculo. Se busca encontrar la recta de predicción de mínimos cuadrados para estos datos, creyendo que la calificación de matemáticas puede ser un determinante para la calificación de cálculo

Tabla de calificaciones

Matemáticas

Cálculo

39

65

43

78

21

52

64

82

57

92

47

89

28

73

75

98

34

56

52

75

Sabemos que nuestra variable dependiente o de respuesta viene siendo la calificación de cálculo, pues buscamos ver si a través de las calificaciones de matemáticas se puede predecir su valor. Mientras que nuestra variable predictora o independiente viene siendo la calificación de matemáticas.

Sabemos que de nuestra muestra de datos, el promedio de ambos conjuntos y su tamaño son los siguientes:

n <- length(x)  
xprom <- mean(x)   
yprom <- mean(y)

## La longitud de los datos es: 10

## El promedio de x es: 46

## El promedio de y es: 76

Ahora, para poder realizar el cálculo de los estimadores de mínimos cuadrados para y , es necesario conocer la suma de cuadrados de los vectores por separado al igual que el producto cruzado, por lo que se realizará una tabla con las estimaciones

SCxy <- sum(x \* y) - n \* xprom \* yprom  
SCx <- sum(x^2) - n \* xprom^2  
SCy <- sum(y^2) - n \* yprom^2

Tabla de cuadrados

x

y

x.2

xy

y.2

39

65

2474

1894

2056

43

78

2474

1894

2056

21

52

2474

1894

2056

64

82

2474

1894

2056

57

92

2474

1894

2056

47

89

2474

1894

2056

28

73

2474

1894

2056

75

98

2474

1894

2056

34

56

2474

1894

2056

52

75

2474

1894

2056

Suma de cuadrados

x

y

x.2

xy

y.2

460

760

2474

1894

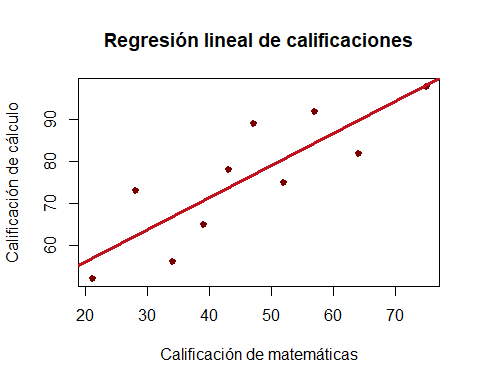
2056

Ahora, con estos datos ya se puede realizar el cálculo de las betas y a su vez crear la recta de regresión, la cual sigue la ecuación de una recta convencional. Donde representa el coeficiente de la variable predictora en el modelo y su intersección.

β1 <- SCxy / SCx  
β0 <- yprom - β1\*xprom  
recta <- β0 + β1\*x

## La estimación para β1 es: 0.7655618

## La estimación para β0 es: 40.78416



Para la estimación de la varianza será necesario conocer la suma cuadrada de la estimación de los errores al igual que los grados de libertad, pero como estamos estimando dos parámetros, los grados de libertad van a corresponder a

SSE <- SCy - β1 \* SCxy  
df <- n - 2  
Scuad <- SSE / df  
S <- sqrt(Scuad)

## La estimación para SSE es: 606.0259

## La estimación para la varianza es: 75.75323

## La estimación de la desviación estándar es: 8.703633

A su vez podemos realizar la estimación del coeficiente de variación el cual se estima sea menor a 0.1 o bien 10%

CV <- 100 \* S / yprom

## El coeficiente de variación es: 11.45215

Ahora para realizar la inferencia sobre siempre se va a mantener la hipótesis que y como alternativa que es diferente de cero. Es decir:

Para ello se usará un nivel de significancia de , que suele ser de las más comunes. Con ello se realizará la estimación de la región de rechazo a través de los cuantiles t student de sus colas (prueba de dos colas), y con el módelo que ya creamos, calcularemos el valor p para así determinar si existe evidencia suficiente para rechazar

α <- 0.05  
qti <- qt(α/2, df, lower.tai = TRUE)  
qtd <- qt(α/2, df, lower.tai = FALSE)  
tc <- β1 / ( S / sqrt(SCx) )  
pvalor <- 2 \* pt( -abs(tc), df)

## El valor estadístico de t es: 4.375015

## El valor p del módelo es: 0.002364532

Como el valor absoluto del estadístico es mayor que el de la cola derecha, se rechaza la hipótesis nula y a su vez con el p valor, al ser menor que el nivel de significancia, de igual manera la hipótesis nula se rechaza

## Rechazamos H0, β1 ≠ 0  
## 0.7655618 ≠ 0

## Rechazamos H0, β1 ≠ 0  
## 0.7655618 ≠ 0

Ahora para calcular el intervalo de confianza de se calculará el límite superior para

limsup <- β1 + qtd \* S / sqrt(SCx)  
liminf <- β1 - qtd \* S / sqrt(SCx)

## El límite superior de β1 es: 1.169078

## El límite inferior de β1 es: 0.3620458

Además para saber si los datos están o no fuertemente relacionadas o no, es decir, cambian simultaneamente, con reacciones opuestas o no tienen ninguna relación alguna nos va a servir observar el coeficiente de correlación que se determina de la siguiente manera:

r <- SCxy / sqrt( SCx \* SCy )

## El coeficiente de correlación es: 0.8397859

Un coeficiente alto y cercano a 1 nos indica una relación significativa entre las variables y que si una aumenta, la otra tiene en gran parte el mismo cambio. En caso de ser negativo alto, nos menciona que cuando una aumente, la otra llevará el camino contrario, es decir, disminuirá, en este caso las variables están fuertemente correlacionadas.

Otro dato que nos ayudará a determinar la precisión del modelo, va a ser el coeficiente de determinación, cuyo dominio se encuentra entre 0 y 1, representando el 1 una precisión casi perfecta en el modelo mientras que el 0 no logrará modelar los valores en absoluto. En este caso tenemos un valor alto que dice que nuestro modelo tiene una precisión alta.

r2 = 1 - SSE / SCy

## El coeficiente de determinación es: 0.7052403

A su vez, podemos realizar el cálculo de los errores,tanto del promedio de la variable de respuesta y el de la predicción, a continuación se realizará el cálculo de cada uno de ellos. Entre menores sean dichos errores en promedio, se dice que el modelado será de igual manera más preciso, pues implicará que se suela acercar a los valores reales de la población.

erroryprom <- S \* sqrt( (1/n) + (x - xprom)^2 / SCx )   
errorpredic <- S \* sqrt( 1 + (1/n) + (x - xprom)^2 / SCx )

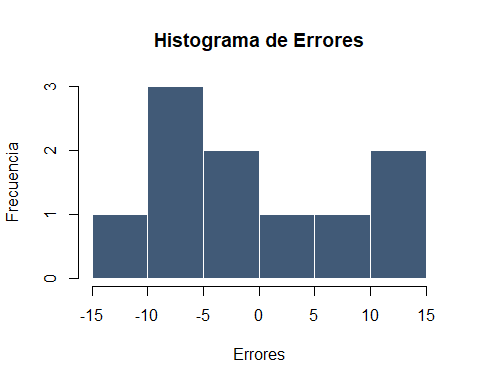
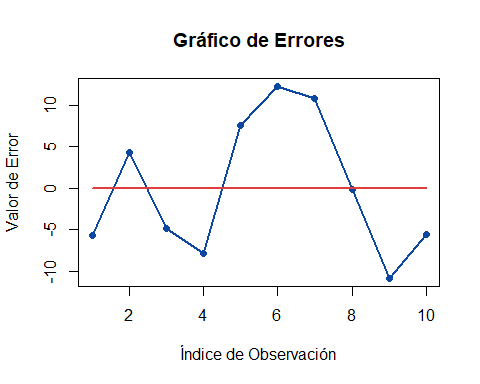
cat("El error promedio de la variable de respuesta es: ", erroryprom)

## El error promedio de la variable de respuesta es: 3.012589 2.801946 5.168429 4.182836 3.358618 2.757887 4.182836 5.772913 3.461873 2.945782

cat("El error promedio de la predicción es: ", errorpredic)

## El error promedio de la predicción es: 9.210262 9.14353 10.12254 9.65657 9.329177 9.130125 9.65657 10.44413 9.366846 9.188627

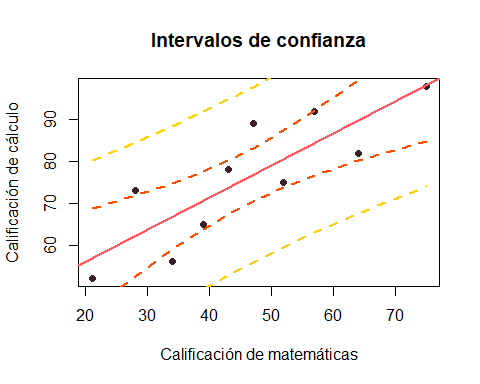
Analizando los errores igual podremos obtener unas gráficas que nos ayudarán a obtener más datos acerca de estos y de una manera más visual. Primeramente tendremos una gráficad e puntos tradicional donde podremos observar qué tanto se alejan de la recta de regresión y en el historgrama podremos observar qué intervalos de errores se presentan com mayor frecuencia, además se podrían realizar boxplots o gráficos qq pero en esta ocasión creemos que estos dos son los que dan resultados más visuales.



## El error promedio es: -1.421216e-15

De igual manera tendremos el cálculo de los intervalos de confianza que nos ayudarán a marcar límites los cuales nuestra serie de datos y predicciones no rebasarán, en el segundo caso, los límites abarcarán un espacio más extenso al tratarse de una predicción y ser mucho menos precisos a la hora de realizar las estimaciones. Dichos límites se marcarán en una gráfica a continuación para una mejor apreciación

lieeyprom <- recta - qtd \* erroryprom  
lseeyprom <- recta + qtd \* erroryprom  
lieepredic <- recta - qtd \* errorpredic  
lseepredic <- recta + qtd \* errorpredic



Como últimos pasos para nuestro análisis de regresión, realizaremos dos pruebas de normalidad para determinar si nuestros datos se aproximan a una normal, esto lo realizaremos mediante las dos pruebas clásicas de normalidad que son Shapiro y Anderson-Darling

normalidad\_shapiro <- shapiro.test(error)

## Como el pvalor ≥ α, no rechazamos H0(los datos son normales) y no hay evidencia que garantice que NO son normales 0.3262792 ≥ 0.05

library(nortest)  
normalidad\_anderson <- ad.test(error)

## Como el pvalor ≥ α, no rechazamos H0(los datos son normales) y no hay evidencia que garantice que NO son normales 0.3115597 ≥ 0.05

Generalmente nos quedaremos con el p valor más alto, en este caso el de Shapiro, el cuál dice que los datos se comportan de manera normal

Además sobre los datos podemos realizar pruebas de correlación e independencia, de la primera ya cálculamos una. Pero en este caso se dará el nombre y los valores que arroja, para el caso de independencia tenemos la de y Fisher

tabla\_contingencia <- table(x, y)  
independencia\_chi\_cuadrada <- chisq.test(tabla\_contingencia)

## Warning in chisq.test(tabla\_contingencia): Chi-squared approximation may be  
## incorrect

## pvalor > α, No hay evidencia para decir que no son independientes 0.2313417 > 0.05

independencia\_fisher <- fisher.test(tabla\_contingencia)

## pvalor > α, No hay evidencia para decir que no son independientes 1 > 0.05

A su vez, las pruebas de correalción son las siguientes

correlacion\_pearson <- cor(x, y)

## Hay una correlación fuerte, coeficiente de correlación de Pearson: 0.8397859

correlacion\_spearman <- cor(x, y, method = "spearman")

## Hay una correlación fuerte, coeficiente de correlación de Spearman: 0.8787879