Actividad. Regresión Lineal Simple

Daniel Orellana Pérez

1.Si pretendiésemos explicar un suceso y/o fenómeno acontecido en el pasado ¿Podemos inferir la respuesta asociada a dichos eventos en base a los restos materiales presentes?

Como arqueólogos, nuestro trabajo precisamente es ese, dar respuesta y explicar los fenómenos acontecidos en el pasado. Este trabajo lo realizamos a través de los restos materiales procedentes del registro arqueológico. Si bien, hay que tener en cuenta que nunca sabremos del todo cómo aconteció un suceso o fenómeno por bien documentado que esté puesto que no lo hemos podido experimentar y ver con nuestros propios ojos. Los arqueólogos trabajamos a través de hipótesis y estudios intentando dar respuesta a ello, pero con la convicción que con el tiempo se demuestre que nuestra teoría era errada y pueda ser sustituida por otra.

2. Haciendo referencia al análisis de correlación lineal de Pearson, ¿establece este algún tipo de relación causa-efecto de una variable sobre otra(s)?

Haciendo referencia al análisis de correlación lineal de Pearson no que se establece una relación causa-efecto de una variable sobre otra puesto que, este tipo de correlación lineal se encarga del análisis de la fuera y la relación entre dos variables.

3. Define causalidad. Exponga algún ejemplo.

La causalidad es la relación causa-efecto entre las variables independientes (predictoras) y la variable dependiente (la respuesta). Un ejemplo de ello sería que por un supuesto estudiáramos la relación existente entre la cantidad de yacimientos de época Flavia (variable independiente)y la utilización del mármol procedente de canteras italianas (variable dependiente).

4. ¿Podrías mencionar los parámetros involucrados en la ecuación de regresión lineal?

Los parámetros involucrados en la regresión son la pendiente y el intercepto.

5. En un plano cartesiano, si afirmo que el eje ‘x’ también se denomina eje de ordenadas, ¿estoy en lo cierto?

No, no estaría en lo correcto puesto que, en un plano cartesiano, el eje "x" se denominaría eje de abscisas.

6. ¿Sabrías diferenciar entre recta de regresión y plano de regresión?

Una recta de regresión corresponde, como su propio nombre indica a una recta ajustada a los datos correspondientes dentro de un plano cartesiano en una regresión lineal simple (dos dimensiones). El plano de regresión se corresponde a diferentes variables independientes dentro de una regresión lineal múltiple (tres dimensiones).

7. ¿Cuáles son los supuestos (o hipótesis) del análisis de regresión lineal?

Dentro de los supuestos (o hipótesis) del análisis de regresión lineal encontramos:

1.-Homocedasridad: Los residuos tienen varianza constante en cada nivel de x. Esto provoca que dichos residuos estén alrededor de la línea de regresión.

2.- Independencia: Los residuos como el propio nombre indica, son independientes, por tanto, no presentan correlaciones entre sí. En particular, no existe correlación entre residuos consecutivos en datos de series de tiempo.

3.- Normalidad: Los residuos del modelo se distribuyen como su nombre indica de manera normal.

4.- Relación lineal o linealidad: existe una relación lineal entre la variable independiente (x), y la variable dependiente (y). Por tanto, cualquier cambio en una, es proporcional a la otra.

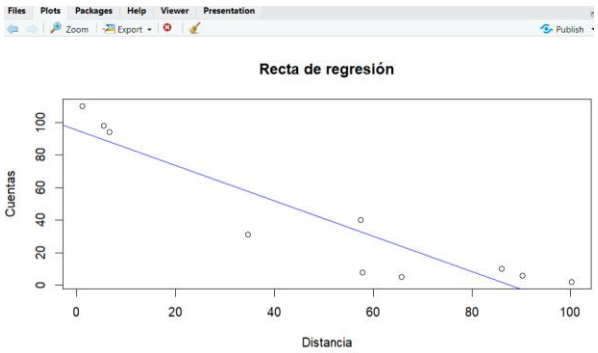
8. Dados los siguientes datos, calcula la recta de regresión que mejor se adapte a nuestra nube de puntos siendo "cuentas" la variable dependiente o de respuesta y "distancia" la variable independiente o explicativa.

Figura 1: Tabla de datos referidos a número de cuentas por yacimiento y distancia (km) del yacimiento a la mina. Fuente: Elaboración propia.

cuentas	distancia
110	1.1
2	100.2
6	90.3
98	5.4
40	57.5
94	6.6
31	34.7
5	65.8
8	57.9
10	86.1

```
#Ejercicio 8
cuentas <- c(110, 2, 6, 98, 40, 94, 31, 5, 8, 10)
distancia <- c(1.1, 100.2, 90.3, 5.4, 57.5, 6.6, 34.7, 65.8, 57.9, 86.1)
datos <- data.frame(cuentas, distancia)
modelo <- lm(cuentas ~ distancia, data = datos)

summary(modelo)
plot(distancia, cuentas, main = "Recta de regresión", xlab = "Distancia", ylab = "Cuentas")
abline(modelo, col = "blue")
```



9. ¿Serías capaz de interpretar el significado de los parámetros de la ecuación de regresión?

10. ¿Qué implicaciones conlleva obtener un intercepto con valor '0'?

11. ¿Qué ponderación lleva a cabo el análisis de regresión lineal para calcular los valores de los parámetros que configuran la recta de regresión?

La ponderación que lleva a cabo el análisis de regresión lineal para calcular los parámetros que configuran la recta de regresión es el MCO, es decir, el método de mínimos metros cuadrados ordinarios.

12. ¿Cuál sería el error asociado a mi modelo en la estimación del número de cuentas para un yacimiento que se encuentra a 1.1 km de la mina?

13. ¿Cómo calcularías los residuos del modelo dado los siguientes datos?

cuentas	predicciones
6	-6.682842
98	85.520196
40	28.938591
94	84.216973
31	53.69983
5	19.924631
8	28.504183
10	-2.121561

Figura 2: Tabla de datos referidos a número de cuentas por yacimiento y su estimación atendiendo al modelo lineal. Fuente: Elaboración propia.

#Ejercicio 13

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

```
> cuentas_obs <- c(6, 98, 40, 94, 31, 5, 8, 10)
> predicciones <- c(-6.682842, 85.520196, 28.938591, 84.216973, 53.69983, 19.924631, 28.504183, -2.121561)
> residuos <- cuentas_obs - predicciones
> print(residuos)
[1]  12.682842  12.479804  11.061409   9.783027 -22.699830 -14.924631 -20.504183
[8]  12.121561
```

14. Con los datos residuales, verifica si se cumple (o no) el supuesto de normalidad.

#Ejercicio 14

```
qqnorm(residuos)
qqline(residuos)
```

15. ¿Que 2 de conjuntos (de datos) se han de emplear en la modelización lineal? ¿Cómo llevarías a cabo la preparación de estos?

16. Evalúa la capacidad predictiva del modelo implementando una validación cruzada simple.

```
cuentas <- c(6, 98, 40, 94, 31, 5, 8, 10)
distancia <- c(1.1, 100.2, 90.3, 5.4, 57.5, 6.6, 34.7, 65.8)

set.seed(123)
indices_entrenamiento <- sample(1:length(cuentas), 0.7 * length(cuentas))
indices_prueba <- setdiff(1:length(cuentas), indices_entrenamiento)
```

```
cuentas_entrenamiento <- cuentas[indices_entrenamiento]
distancia_entrenamiento <- distancia[indices_entrenamiento]

cuentas_prueba <- cuentas[indices_prueba]
distancia_prueba <- distancia[indices_prueba]

modelo <- lm(cuentas_entrenamiento ~ distancia_entrenamiento)

predicciones <- predict(modelo, data.frame(distancia_entrenamiento = distancia_prueba))

error_cuadratico_medio <- sqrt(mean((cuentas_prueba - predicciones)^2))
r_cuadrado <- cor(predicciones, cuentas_prueba)^2

print(paste("Error cuadrático medio:", error_cuadratico_medio))
print(paste("Coeficiente de determinación (R^2):", r_cuadrado))
```

```
> print(paste("Error cuadrático medio:", error_cuadratico_medio))
[1] "Error cuadrático medio: 61.6467456802843"
> print(paste("Coeficiente de determinación (R^2):", r_cuadrado))
[1] "Coeficiente de determinación (R^2): 0.0305863215184227"
```

17. Si mis coeficientes de regresión se han calculado con un intervalo de confianza del 95% ¿cuál será la probabilidad de que la correlación lineal entre los coeficientes de regresión y la variable de respuesta o explicada se deba al azar? ¿Y si tengo un nivel de significación (Alpha) de 0?01, con que Intervalo de Confianza he obtenido mis coeficientes de regresión?

18. Si las estimaciones arrojadas por mi modelo lineal resultan menos precisas (mayor error) en un determinado rango de valores con respecto a otro, decimos ¿qué hay indicios de homocedasticidad o heterocedasticidad?

Si las estimaciones arrojadas por mi modelo lineal fueran menos preciosas debemos indicar que hay indicios de heterocedasticidad en los residuos del modelo puesto que esto significa que nuestros residuos estarían alrededor de la línea de regresión.

19. ¿Qué medida de precisión estadística nos indica el % de variabilidad explicada de la variable dependiente por nuestro modelo lineal?

La medida de precisión estadística nos indica el % de variabilidad explicada de la variable dependiente por nuestro modelo lineal es el coeciente de determinación (R al cuadrado).

20. Explica la diferencia entre una observación atípica y una observación que produzca lo que se conoce en estadística como “apalancamiento” del modelo