Alumna: Elena Carretero Moreno

ACTIVIDAD. CORRELACION LINEAL

1) La correlación lineal es un método estadístico para estudiar la relación entre dos variables. Existen dos variables continuas, que tiene que tener parámetros para poderlas cuantificar , con parámetros como la covarianza (grado de variación de estas dos variables aleatorias, que se estandarizan generando los coeficientes de correlación).
2) La diferencia entre las pruebas paramétricas y no paramétricas reside en qiue las paramétricas deben ajustarse a una distribución mientras que las otras no, pues las paramétricas necesitan condiciones de validez, asumiendo distribuciones estadísticas subyacentes. Las pruebas paramétricas tienen mas potencia estadística, mientras que las no paramétricas son válidas en un rango más amplio de situaciones.

9) La diferencia entre una relación lineal y una monótona, sabemos que la relación lineal entre dos variables es cuando el cambio en una es proporcional al de la otra (representado con una línea recta). Mientras que la monótona cuando una variable aumenta, la otra esta en la misma dirección(aumentando o disminuyendo) aunque no tiene que ser una tasa constante. Ponemos el ejemplo en R en el directorio.

ACTIVIDAD. REGRESIÓN LINEAL SIMPLE.

1) Si pretendiésemos explicar un fenómeno del pasado, en mi opinión si se puede inferir la respuesta a dichos elementos usando los restos materiales presentes, puesto que estos son los restos de esa actividad que se ha llevado a cabo en el pasado, obviamente sin olvidar que la mayoría de restos son orgánicos y ya se han degradado, por lo que su historia no puede ser reconstruida al 100%. Pero hay objetos que excepcionalmente se han conservado y que nos ayudan a tener una retrospectiva de este pasado, por lo que estos restos pueden ser cuantificados y examinados para la asociación de estos a un modelo común o a una posible. En estadística los datos históricos pueden ser usados con modelos de regresión lineal, y para cálculos más detallados tendríamos que ser más específicos en los datos que utilizaremos.
2) La correlación lineal de Pearson sirve para medir el grado de relación lineal de dos variables, aunque no establece una relación de causa efecto. Aunque sea una correlación alta no dice si una variable causa cambios en otra. Para identificar la relación causa efecto, se necesitan estudios específicos.
3) La causalidad es la relación entre dos eventos, y asi uno es el resultado del otro, por una conexión causa efecto (siendo el efecto la consecuencia). Un ejemplo de causa efecto es el tiempo invertido en un trabajo y los resultados mejores o peores de este, y para establecer la causalidad, se buscan cumplir criterios, como la temporalidad, el enfoque, la practica y el resultado de experimentos. En estadística consiste en hacer un diseño experimental donde se manipula activamente la variable(causa) y se ven los resultados (consecuencia)
4) En una ecuación de regresión lineal se involucran los siguientes parámetros:

a. Intercepto, que es el valor de la variable dependiente (todas las independientes son 0), este representa el punto donde la línea de regresión pone el eje Y.
b. Coeficiente o pendiente, la cantidad que la variable cambiante por el cambio en la independiente. En una regresión lineal múltiple, con mas de una variable independiente, habría un coeficiente para cada una.

En R, la función lm es la que funciona para calcular este tipo de datos, pudiendo ser tanto para regresiones lineales simples como múltiples. El código producirá resúmenes para los coeficientes del modelo, estadísticas (r-cuadrado o p-valor). Y la estimación del intercepto es para ver como cambia el valor de "y" y "x".

5) No, en un plano cartesiano, el eje "x" y el "y" forman el sistema de coordenadas para visualizar las posiciones en un plano, por lo que el eje de abscisas suele representar la variable independiente de la grafica y el de ordenadas la dependiente.
6) La recta en un plano de regresión es una línea recta que va sobre los datos en una gráfica de dispersión mientras que el plano es utilizado para una regresión lineal múltiple, con unas cuantas variables independientes además de la variable dependiente, por lo que se genera un plano, es una superficie que se extiende en 3 dimensiones para representar la relación entre todas las variables independientes con la dependiente.
7) Los supuestos son:
    a. Linealidad: relación lineal entre variable dependiente e independiente (se puede comprobar por un diagrama de dispersión.
    b. Independencia: Observaciones independientes, los residuos no deben influir entre sí
    c. Homocedasticidad: Los residuos que tienen variabilidades similares en la línea de regresión.
    d. Normalidad de los errores: Los residuos tienen que seguir una distribución normal.
    e. No multicolinealidad: las variables independientes no se relacionan entre sí, porque puede entonces ser difícil saber el efecto independiente de la variable dependiente.
8) **Ejercicio realizado en R**
9) La interpretación del ejercicio realizado en R seria los parámetros intercepto, que te dice el valor esperado de y, pero solo cuando las variables que son independientes son 0, siendo importante para la línea de regresión en el espacio. También mide el coeficiente de las variables independientes, que muestra el efecto promedio en la variable dependiente cuando incrementa una unidad en la independiente (las demás independientes son constantes).
10) Tener un intercepto de valor 0, puede significar que la línea de regresión está pasando por su origen en un plano cartesiano (cuando la variable independiente es 0, la dependiente también es 0). Lo que implicaría que en la interpretación práctica, el intercepto de cero puede o no tener sentido práctico dependiendo del fenómeno. En cuanto al ajuste del modelo a los datos, este intercepto puede indicar que los datos se ajustan de manera que la relación de dependencia empieza al principio, para finalmente un posible sobre juste o limitación en el procedimiento, pues forzarlo a cero es peligroso si no se basa en una justificación teórica y puede limitar la capacidad del modelo para capturar la variabilidad de los datos y no reflejar bien el fenómeno que queremos comprender.
11) La forma de ponderación refleja que (x_i) y (y_i) son las variables independientes y dependientes. **Ejercicio hecho en R.** En este, y es la variable dependiente y x es la

independiente, mientras que b_O es el intercepto con el eje y, y b_1 es la pendiente de la recta.

12) **Hecho en R.**
13) Estos son los datos de residuos de la tabla:

```
Residuals:
   Min     1Q Median     3Q    Max
-22.70 -16.32  10.42  12.21  12.68
```

14) El supuesto de normalidad no se cumple en este caso, pues con el shapiro.test el valor p es 0.004895, lo que indica que los valores no se distribuyen normalmente, pues unos valores que se distribuyen normalmente tienen un valor p más alto de 0.05.
15) En la modelización lineal hay que emplear 2 datos, un conjunto de entrenamiento (ajusta la regresión lineal, con la que el algoritmo aprende la relación entre variables) y un conjunto de prueba (para evaluar como predice datos el modelo). Estos datos se preparan limpiando los datos, haciendo una división en conjuntos (con porcentajes que varían dependiendo del conjunto) y la estandarización o normalización (transformar datos para tener media de 0 y desviación estándar de 1, o que estén en escala entre 0 y 1). La preparación sería en un modelo similar a este ejemplo:
    a. set.seed(123)
    b. indices_entrenamiento <- sample(1:length(cuentas_), 0.7 * length(cuentas_))
    c. datos_entrenamiento <- datos[indices_entrenamiento,]
    d. datos_prueba <- datos[-indices_entrenamiento,]
16) **Hecho en R.**
17) Con el intervalo de confianza en un 95%, tenemos un 5% de probabilidad de que la correlación sea debido a que es al azar, por que el nivel asociado a es del 5% (100%-95%). Y con un nivel de significación del 0.01 tendríamos el intervalo de confianza al 99% en los coeficientes propios de regresión.
18) Sería entonces que hay indicios de heterocedasticidad, porque cuando las estimaciones varían a lo largo de rangos de valores (el tamaño de error de las predicciones cambia dependiendo del valor). En esta situación, tenemos que los errores tienen varianzas que no tienen constancia en la variable independiente. Aparece esta indicando que el error varia en los rangos de la variable.
19) Es el coeficiente de determinación, que se conoce también como (R^2)(R cuadrado), que varia entre 0 y 1, donde el valor 0 sería que el modelo no explica nada de la variabilidad y 1 que si explica toda la variabilidad de los datos de la respuesta de la media. En nuestro caso práctico, tenemos un R^2 de O.8258854, lo que sugiere que el 82% de la variabilidad del modelo en la variable dependiente podría ser perfectamente explicada por el modelo (aunque este índice alto no implica que sea totalmente correcto).
20) Una observación atípica es un punto de datos que difiere de otras observaciones en el conjunto, pudiendo estar muy por encima o debajo de los otros puntos de daros, lo que a su vez puede dar variabilidad en la medición y errores. Es un punto con valor en la variable dependiente inusual por su valor en la independiente (pueden llevar a estimaciones sesgadas, por lo que puede ser recomendable excluirlos).
    También una observación con alto apalancamiento seria aquella que tiene valores extremos para varias variables independientes en comparación con el resto, por lo que pueden ejercer una influencia en la estimación de los parámetros. Aunque no todas son problemáticas, solo las que tienen residuos grandes. Ambas tienen diferencia en que mientras un outlier es de un valor inusual de la dependiente (y), la de alto apalancamiento es de los valores extremos de las variables independientes(x).