UNIDAD 4: Análisis de Regresión y Correlación
TRABAJO PRÁCTICO IV

1. Se dan los siguientes pares ordenados, correspondientes a observaciones en que X es una variable no aleatoria e Y es una variable aleatoria. Asuma que la relación es lineal, es decir, $E(Y/x) = \beta_0 + \beta_1 x$ y que todos los supuestos pueden ser sostenidos

x	0	2	4	6	8
y	9	11	11	14	15

Práctica con cálculos básicos

a) Representar los datos en un diagrama de dispersión.

b) Calcular la suma, la media, suma de cuadrados de X corregidos por la media, varianza muestral y desvío estándar muestral para X.

c) Idem para Y.

d) Encontrar estimación por mínimos cuadrados de β_0 y β_1 y graficar la recta de ajuste en el diagrama anterior.

e) Indicar cuales son los residuales correspondientes a cada valor de Y estimado o predicho. Verificar que la suma de los residuales es cero.

f) Calcular:

f1) la suma de cuadrados total o suma de cuadrados de la variable Y (SCtotal ó SCy)

f2) la suma de cuadrados "explicada por la regresión" (SCexplicada = b^2 SCx)

f3) la suma de cuadrados "no explicada por la regresión" ó "suma de cuadrados del error" (SCno explicada ó SCerror = SCy - SCexplicada).

f4) Verificar que la suma de cuadrados del error (RSS) obtenida de este modo coincide con la que obtendría utilizando la suma de los residuales al cuadrados del inciso c.

g) Calcular el coeficiente de determinación e interpretar el resultado obtenido.

h) Calcular el desvío estándar residual o error medio cuadrático usando de la suma de cuadrados no explicada obtenida anteriormente, e interpretar el resultado obtenido.

i) Verificar la hipótesis Ho: $\beta_1 = 0$ usando:

i) una prueba t ; ii) una prueba F ; por qué se dice que ambas pruebas son equivalentes?

j) Usar el modelo para obtener una estimación puntual del valor de Y correspondiente a X= 5 y calcular el desvío estándar de esta estimación para el caso:

j1) el valor de Y estimado representa a la media de la subpoblación de valores de Y asociada a X=5 (estimación de μ i dado que X= x_i)

j2) el valor de Y estimado representa a un individuo cualquiera de la subpoblación de Y asociada a X= 5 (estimación de y_{ij} dado que X= x_i).

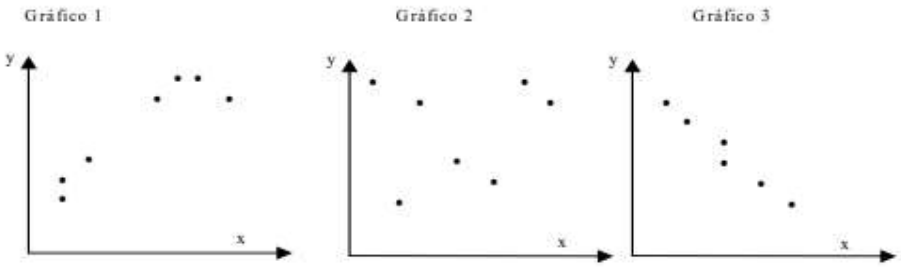
k) Usar los resultados del inciso anterior para calcular intervalos del 95% de confianza para las dos estimaciones allí obtenidas (para la estimación de μ i y para yij). **Interprete** ambos intervalos.

2. A continuación se presentan tres afirmaciones referidas a las conclusiones de un estudio realizado acerca de las tasas de nacimiento, suicidio, crecimiento y productividad, junto con tres gráficos de dispersión.

Afirmación 1: En países con un desarrollo tecnológico alto tales como Japón, Estados Unidos, Alemania, Inglaterra, Francia, Italia y Canadá, se tienen bajas tasas de nacimiento (TN) asociadas con altas tasas de suicidio (TS)

Afirmación 2: Algunos economistas afirman que independientemente de los países que se estudian, a altas tasas de crecimiento (TC) se asocian altas tasas de productividad (TP)

Afirmación 3: Tanto economistas como demógrafos, afirman que las tasas de suicidio (TS) no parecen estar correlacionadas con las tasas de productividad (TP)



Indique a qué grafico corresponde cada una de las afirmaciones dadas.

3. Los datos del archivo *EdadPesoGrasas.xlsx* corresponden a tres variables medidas en 25 individuos: edad, peso (en kg) y cantidad de grasas en sangre (mg./dl.)

a) Realice una matriz de diagrama de dispersión. ¿sugiere la pertinencia del modelo de regresión lineal simple entre algún par de variables?
b) Halle la estimación de la recta de ajuste mínimo cuadrático que le permita expresar la cantidad de grasa en la sangre en función de la edad de los individuos. Interprete si corresponde los coeficientes.
c) Represente gráficamente la recta estimada en el diagrama de dispersión correspondiente.
d) Interprete R^2
e) Realice un estudio de los residuales para verificar el cumplimiento de los supuestos. Efectúe un estudio de balanceo, influencia y valores atípicos.
f) ¿Es posible afirmar que existe relación lineal entre la cantidad de grasa en la sangre y la edad de los individuos? Utilice la tabla ANOVA.
g) Calcule e interprete un intervalo de confianza del 95% de la cantidad de grasa media para un individuo de 32 años.
h) Calcule e interprete un intervalo de predicción del 95% de confianza de la cantidad de grasa para un individuo de 50 años.

4. Las calificaciones obtenidas por un grupo de 10 alumnos al aplicarles dos test son las siguientes (la calificación varía de 0 a 20)

Test 1	2	7	8	9	10	12	14	10	16	12
Test 2	5	8	10	12	12	14	15	16	18	20

Realice el test apropiado que le permita evaluar la correlación lineal entre las dos variables. Comente los resultados del mismo.

5. Se cree que la pureza del Oxígeno producido en un proceso está relacionada con el porcentaje de hidrocarburos en el condensador principal. El archivo Pureza.csv contiene los datos obtenidos en un estudio del proceso de purificación.

a. Represente las variables en un diagrama de dispersión y efectúe una descripción del mismo.

b. Halle la recta de ajuste mínimos cuadrados que le permita expresar la pureza del Oxígeno en función del porcentaje de hidrocarburos en el condensador principal. **Represente** la recta estimada en el diagrama de dispersión. **Interprete** los coeficientes si correspondiera.

c. Interprete R^2.

d. Realice un estudio de los residuales para verificar el cumplimiento de los supuestos. Efectúe un estudio de balanceo, influencia y valores atípicos.

e. Plantee las hipótesis correspondientes, obtenga la tabla ANOVA y concluya.

f. Calcule e **interprete** un intervalo del 95% de confianza para la pendiente del modelo.

g. Para un porcentaje de hidrocarburos del 1%: Calcule e interprete el intervalo de confianza y el de predicción. Utilice un nivel de confianza del 95%.

h. Suponga que la pureza del Oxígeno y el porcentaje de hidrocarburos son variables aleatorias con distribución de probabilidad conjunta normal bivariada. Formule y efectúe las hipótesis que correspondan a este problema sobre el coeficiente de correlación. **Concluya**.

6. Se ajustó un modelo de regresión lineal simple a un conjunto de datos y se obtuvo:

- $\hat{y}_i = 11{,}5 - 1{,}5x_i$
- La hipótesis bilateral H_0: $\beta_1 = 0$ no resultó significativa al 5%. Se comparó un $t_{ob} = -4{,}087$ con $t_{2,0.05}$
- La estimación de σ^2 resultó $s_e^2 = 1{,}75$

a. Complete la tabla ANOVA correspondiente a este ajuste utilizando los resultados dados.

b. Calcule e interprete el coeficiente de determinación.

7. Cargue en Rcmdr el conjunto de datos *trees* del paquete *datasets* y obtenga ayuda sobre el significado de las variables que conforman este conjunto de datos. Se desea modelar la relación entre la variable `Volume` (Respuesta) y la variable `Girth`.

a. Represente un diagrama de dispersión y efectúe una descripción del mismo.

b. Exprese la recta de ajuste mínimos cuadrados. **Represente** la recta en el diagrama de dispersión. **Interprete** los coeficientes cuando sea pertinente.

c. Interprete R^2.

d. Realice un estudio de residuales para verificar el cumplimiento de supuestos. Incluya un estudio de balanceo, influencia y valores atípicos.

e. Plantee las hipótesis correspondientes, obtenga la tabla ANOVA y concluya.

f. Calcule e **interprete** un intervalo del 95% de confianza para la pendiente del modelo.

g. Para `Girth = 17`: Calcule e interprete el intervalo de confianza y el de predicción. Utilice un nivel de confianza del 99%.