GRADO EN INTELIGENCIA ARTIFICIAL (MOSTOLES)

2361 - PROBABILIDAD Y ESTADISTICA - MAÑANA A - 2Q

Exámenes > Parcial1 - Marzo 2024

Comenzado el	miércoles, 20 de marzo de 2024, 11:10
Estado	Finalizado
Finalizado en	miércoles, 20 de marzo de 2024, 11:56
Tiempo empleado	46 minutos 3 segundos
Calificación	**8,67** de 10,00 (**86,72**%)

Pregunta 1

Correcta

Se puntúa 0,75 sobre 0,75

Para evaluar cómo diversos factores influyen en el tiempo de respuesta de una aplicación web, se han registrado ciertas métricas relacionadas en múltiples escenarios de uso. Los datos se recopilaron durante un período de tiempo considerable, y se realizó un análisis de regresión para comprender la relación entre el tiempo de respuesta de la aplicación (en milisegundos) y el número de usuarios conectados de forma concurrente. Los resultados de este análisis se presentan a continuación:

```
> summary(regresion1, digits = 2)
```

```
Call:
lm(formula = TiempoRespuesta ~ UsuariosConcurrentes, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-51.581 -21.061  -2.468  16.813  64.344

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.77809    9.11245   0.085    0.932
UsuariosConcurrentes  0.15002    0.01223  12.271   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.23 on 73 degrees of freedom
Multiple R-squared:  0.6735,    Adjusted R-squared:  0.669
F-statistic: 150.6 on 1 and 73 DF,  p-value: < 2.2e-16
```

```
> summary(regresion2, digits = 2)
```

```
Call:
lm(formula = UsuariosConcurrentes ~ TiempoRespuesta, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-265.133 -109.843   -5.932  122.709  233.080

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     226.0331    42.2507    5.35  9.7e-07 ***
TiempoRespuesta   4.4892     0.3659   12.27  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 143.5 on 73 degrees of freedom
Multiple R-squared:  0.6735,    Adjusted R-squared:  0.669
F-statistic: 150.6 on 1 and 73 DF,  p-value: < 2.2e-16
```

Por termino medio, ¿cómo se ve afectado el tiempo de respuesta cuando disminuye el número de usuario en 1?

Seleccione una:

- a. Ninguna de las otras opciones es correcta.
- b. Se reduce en 0.7781 milisegundos.
- c. Se reduce en 4.4892 milisegundos.
- d. Se reduce en 226.0331 milisegundos.
- e. Se reduce en 0.15 milisegundos. ✔ Es correcto, la pendiente de la recta de regresión.

a. Incorrecta.
b. No es correcto, hay que observar la pendiente.
c. No es correcto, si, hay que observar la pendiente pero, en este caso, habría que tener en cuenta la otra regresión.
d. No es correcto, hay que observar la pendiente, y, en este caso, habría que tener en cuenta la otra regresión.
e. Es correcto, la pendiente de la recta de regresión.

La respuesta correcta es: Se reduce en 0.15 milisegundos.

Pregunta **2**

Correcta

Se puntúa 1,50 sobre 1,50

Nota: la tolerancia es 0.001.
Una empresa de software quiere cambiar su interfaz de usuario. Por encuestas realizadas en el pasado, clasifica a sus usuarios en tres grupos: "abiertos", "indiferentes" y "reacios" a adaptarse. Estudios de la empresa indican que la probabilidad de que un usuario acepte el nuevo diseño de la interfaz es: 0.42, 0.36, 0.14, respectivamente, según pertenezca a uno de los tres grupos anteriores. Además se sabe que el 23% de los usuarios pertenecen al grupo de "abiertos", el 23% de los usuarios a "indiferentes" y el 54% de los usuarios a "reacios" a adaptarse.

- Si se entrevista a un usuario al azar, ¿cuál es la probabilidad de que pertenezca al grupo de abiertos y que acepte la interfaz? 0.0966 ✔
- Si se entrevista a un usuario y dice que no le ha gustado la interfaz (no la acepta), ¿cuál es la probabilidad de que pertenezca al grupo de indiferentes? 0.1975839 ✔

- Sean los sucesos,

  A={ Es usuario abierto a adaptarse}
  I={ Es usuario indiferente al cambio}
  R={ Es usuario reacio a adaptarse}
  X={ Aceptar interfaz}
  El enunciado nos indica que:
  $P(A) = 0.23; \quad P(I) = 0.23; P(R) = 0.54$
  $P(X|A) = 0.42; \quad P(X|I) = 0.36; \quad P(X|R) = 0.14$
- Si se entrevista a un usuario al azar, la probabilidad de que pertenezca al grupo de abiertos y que acepte la interfaz es:

  $P(A \cap X) = P(X|A) \cdot P(A) = 0.42 \cdot 0.23 = 0.097$
- Si se entrevista a un usuario y dice que no le ha gustado la interfaz, la probabilidad de que pertenezca al grupo de indiferentes será, aplicando el Teorema de Bayes:

  $P(I|X^c) = \frac{P(X^c|I) \cdot P(I)}{P(X^c)} = \frac{(1-0.36) \cdot 0.23}{1-0.255} = 0.198$

  donde, usando el Teorema de la Probabilidad Total, la probabilidad de que la interfaz sea aceptada es:

  $P(X) = P(X|A) \cdot P(A) + P(X|I) \cdot P(I) + P(X|R) \cdot P(R) = 0.23 \cdot 0.42 + 0.23 \cdot 0.36 + 0.54 \cdot 0.14 = 0.255$

Pregunta **3**

Correcta

Se puntúa 0,75 sobre 0,75

Nota: la tolerancia es 0.1.

Del análisis bidimensional de las variables X e Y observadas en una muestra de tamaño 34, se han obtenido los siguientes resultados:

$$\sum_{i=1}^{34} x_i = 2055 \quad \sum_{i=1}^{34} y_i = -2081 \quad \sum_{i=1}^{34} x_i^2 = 128025 \quad \sum_{i=1}^{34} y_i^2 = 133899 \quad \sum_{i=1}^{34} x_i y_i = -129052$$

Se pide el coeficiente de correlación lineal entre X e Y:

Seleccione una:

- a. -0.656 ✔
- b. Ninguna de las otras opciones es correcta
- c. -0.43
- d. 0.656
- e. 0.43

A partir de los sumatorios de los que disponemos se pueden calcular las medias y varianzas de las variables X e Y y la covarianza entre ellas:

$$\bar{x} = \frac{2055}{34} = 60.441$$

$$\frac{\sum_{i=1}^{34} x_i^2}{34} \frac{128025}{34} = 3765.441$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{34} x_i^2 - \bar{x}^2 = 3765.441 - 60.441^2 = 112.327$$

$$\bar{y} = \frac{-2081}{34} = -61.206$$

$$\frac{\sum_{i=1}^{34} y_i^2}{34} \frac{133899}{34} = 3938.206$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{34} y_i^2 - \bar{y}^2 = 3938.206 - -61.206^2 = 192.032$$

$$\frac{\sum_{i=1}^{34} x_i y_i}{34} \frac{-129052}{34} = -3795.647$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{34} x_i y_i - \bar{x}\bar{y} = -3795.647 - 60.441 \cdot -61.206 = -96.295$$

Podemos resumir estos valores indicando el vector de medias y la matriz de varianzas y covarianzas de estos datos:

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} 60.441 \\ -61.206 \end{pmatrix}; \quad S = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix} = \begin{pmatrix} 112.327 & -96.295 \\ -96.295 & 192.032 \end{pmatrix}$$

- El valor de la pendiente de la recta de regresión es:
  $b = \frac{s_{xy}}{s_x^2} = \frac{-96.295}{112.327} = -0.857$ En cuanto al termino independiente, su valor es:
  $a = \bar{y} - b \cdot \bar{x} = -61.206 - -0.857 \cdot 60.441 = -9.408$ En consecuencia, la recta de regresión es:
  $\hat{y} = -9.408 + -0.857 \cdot x$
- El coeficiente de correlación entre X e Y es :
  $r_{xy} = \frac{s_{xy}^2}{s_x * s_y} = \frac{-96.295}{\sqrt{112.327 \cdot 192.032}} = -0.656$
- #Con R:

```
n=34
sum_x=2055
sum_y=-2081
```

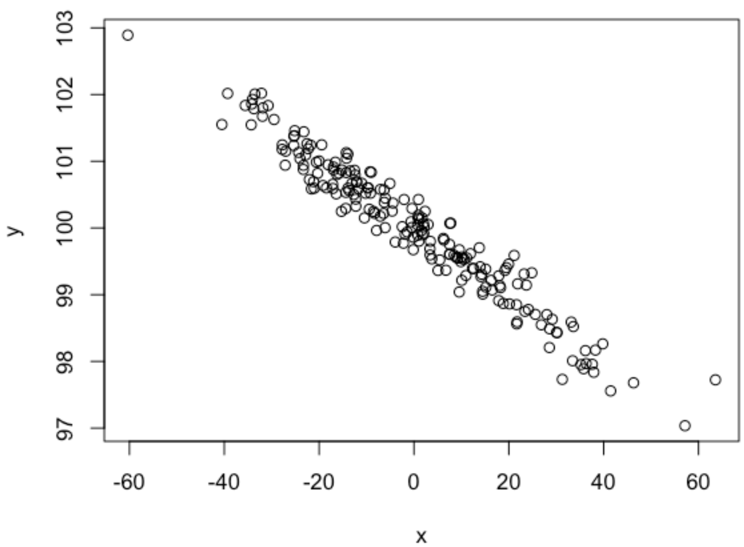
```
sum_x2=128025
sum_y2=133899
sum_xy=-129052

media_x=sum_x/n
media_y=sum_y/n

media_x2=sum_x2/n
media_y2<-sum_y2/n
media_xy<-sum_xy/n

var_x=media_x2-media_x^2
var_y=media_y2-media_y^2
cov_xy<-media_xy-media_x*media_y


(r=round(cov_xy/sqrt(var_x*var_y),3))
```
La respuesta correcta es: -0.656

Pregunta **4**

Correcta

Se puntúa 0,75 sobre 0,75

La siguiente figura muestra un diagrama de puntos. ¿Cuáles de los siguientes son ciertos?



Seleccione una o más de una:

- [x] a. Para $X = -19.7$, Y se espera que tenga un valor aproximado de 101 . ✔ Verdadero. La recta de regresión en $X = -19.7$ implica un valor aproximado de $Y = 101$.
- [ ] b. El valor absoluto del coeficiente de correlación es menor que 0.8.
- [ ] c. La pendiente de la recta de regresión asociada es aproximadamente 1.
- [x] d. La media de Y es mayor que 30. ✔ Verdadero. La media de Y es aproximadamente 100 y por lo tanto es más grande que 30.

a. Verdadero. La recta de regresión en $X = -19.7$ implica un valor aproximado de $Y = 101$.
b. Falso. En el diagrama se observa una fuerte dependencia entre las variables. El valor absoluto del coeficiente de correlación lineal es cercano a 1 y por lo tanto mayor que 0.8.
c. Falso. La pendiente de la recta de regresión viene dada por $r \cdot s_y / s_x$ y por lo tanto no es aproximadamente 1.
d. Verdadero. La media de Y es aproximadamente 100 y por lo tanto es más grande que 30.

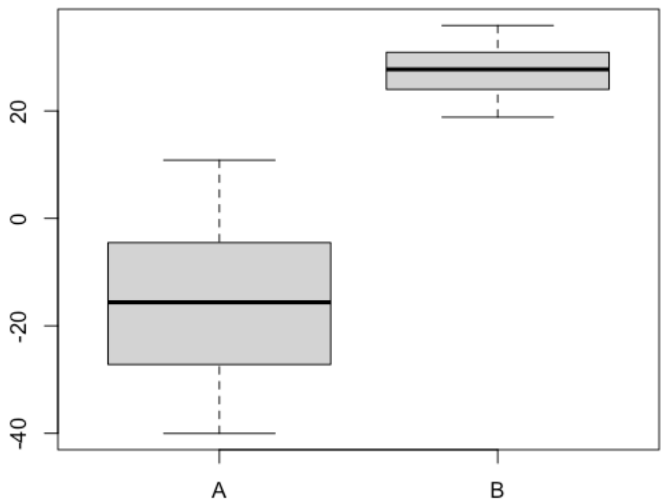
Las respuestas correctas son: Para \(X = -19.7\), \(Y\) se espera que tenga un valor aproximado de 101 .
, La media de \(Y\) es mayor que \(30\).

Pregunta **5**

Parcialmente correcta

Se puntúa 0,50 sobre 0,75

El siguiente gráfico muestra los diagramas de cajas de dos distribuciones de datos (A y B). ¿Cuál de las siguientes afirmaciones es correcta? *(Nota: Las afirmaciones son aproximadamente verdaderas o evidentemente falsas.)*



Seleccione una o más de una:

- [ ] a. El centro de las dos distribuciones es aproximadamente el mismo.
- [ ] b. Ninguna de las distribuciones contiene atípicos.
- [x] c. La variabilidad en la muestra A es obviamente mayor que en la muestra B. ✔ Verdadero. El rango intercuartílico de la muestra A es claramente mayor que la B.
- [x] d. La simetría de las dos muestras es similar. ✔ Verdadero. La simetría de las dos muestras es similar, las dos son aproximadamente simétricas.
- [ ] e. La distribución A es asimétrica a la derecha.

a. Falso. Distribución B tiene valores mayores, en general, que la distribución A.
b. Verdadero. Ninguna de las dos distribuciones tiene observaciones que se desvíen más de 1.5 veces el rango intercuartílico de la caja.
c. Verdadero. El rango intercuartílico de la muestra A es claramente mayor que la B.
d. Verdadero. La simetría de las dos muestras es similar, las dos son aproximadamente simétricas.
e. Falso. La distribución A es aproximadamente simétrica.

Las respuestas correctas son: Ninguna de las distribuciones contiene atípicos., La variabilidad en la muestra A es obviamente mayor que en la muestra B., La simetría de las dos muestras es similar.

Pregunta **6**

Parcialmente correcta

Se puntúa 1,36 sobre 1,50

Nota: la tolerancia para las respuestas numéricas es 0.1, las respuestas verdadero/falso son correctas o claramente incorrectas.) Un técnico está llevando a cabo un estudio comparativo del rendimiento de dos tipos de procesadores:Apple M2 Ultra y i9-13900k. Para ello se ha probado a los procesadores en diferentes escenarios, y se ha anotado el tiempo que tarda en realizar una determinada tarea. Con los datos obtenidos (expresados en segundos) ha rellenado la siguiente tabla:

```
         Apple M2 Ultra i9-13900k
Media               5.6       8.0
Varianza            3.1       1.6
Mínimo              2.3       5.4
Q1                  4.2       7.6
Q2                  5.4       7.8
Q3                  7.3       8.8
Máximo              7.9      10.0
```

- ¿Existe algún atípico en el i9-13900k? Existe algún valor atípicamente bajo ✔
- Tras realizar este resumen con R, el técnico ha repasado los datos del procesador **i9-13900k** y se ha dado cuenta de que el 5.4 que había tecleado era en realidad un 4.4. ¿Cómo variarán los siguientes estadísticos tras corregir esta errata?
    - Media: Disminuirá ✔
    - Mediana: Son iguales ✔
    - Rango intercuartílico: Son iguales ✔
    - Varianza: Aumentará ✔
    - Amplitud o rango: Aumentará ✔
- Revisando sus datos sobre el procesador **Apple M2 Ultra**, el técnico se ha percatado de que inicializó los contadores por lo que todas las mediciones de este tipo de procesador son en realidad 1 unidad menor a lo que ha tecleado. Indicar cómo variarán los siguientes estadísticos tras restar 1 segundo a todos los datos del Apple M2 Ultra.
    - Media: Disminuirá en un segundo ✔
    - Mediana: Disminuirá en un segundo ✔
    - Rango intercuartílico: No cambia ✔
    - Varianza: Disminuirá en un segundo ✖
    - Amplitud o rango: No cambia ✔

Para los datos del procesador i9-13900k, el rango intercuartílico es:

$RI = Q3 - Q1 = 8.8 - 7.6= 1.3$

Multiplicando esta cantidad por 1.5 se obtiene

$1.5 \times RI = 1.5 \times 1.3 = 1.95$

luego, la barrera que establece qué valores de la duración son atípicamente elevados está en

$Q3 + 1.5 \times RI = 8.8 + 1.95 = 10.75$

El valor máximo del procesador i9-13900k es 10, un valor menor o igual a esta barrera, 10 <= 10.75, y en consecuencia puede afirmarse que niguna del procesador i9-13900k presenta una duración atípicamente alta.

Análogamente, estudiamos si hay algún valor atípicamente bajo:
La barrera que establece qué valores de la duración son atípicamente bajos está en

$Q1 - 1.5 \times RI = 7.6 - 1.95 = 5.65$

El valor mínimo del procesador i9-13900k es 5.4, un valor menor a esta barrera, 5.4 < 5.65, y en consecuencia puede afirmarse que al menos una del procesador i9-13900k presenta una duración atípicamente baja.

Si el técnico sustituye el mínimo de los datos del i9-13900k (que era 5.4 unidades) por 4.4 (que es un valor menor) se tendrá que:

- La media disminuirá su valor, y será por tanto inferior a 8 unidades, ya que es la suma de todos los datos dividida entre el número de observaciones.
- La mediana (segundo cuartil) no variará, pues es el dato que ocupa la posición central. Tras la sustitución de 5.4 unidades por 4.4 unidaes, el valor que ocupa la posición central en la muestra ordenada sigue siendo el mismo. Por tanto la mediana se mantiene en 7.8 unidades.
- El rango intercuartílico mantendrá su valor en 1.3 unidades, ya que (por la misma razón que en el caso de la mediana) el primer y el tercer cuartil no se verán modificados tras la sustitución de 5.4 por 4.4.
- La varianza aumentará su valor. La sustitución de 5.4 unidades por 4.4 unidades hace que las desviaciones a la media sean mayores. Al elevar al cuadrado dichas desviaciones y sumarlas se obtendrá un valor superior a 1.6 unidades.
- El rango aumentará su valor, al haber más distancia entre los valores máximos y mínimos.

Si se les resta 1 unidad a todos los datos del Apple M2 Ultra:

- La media disminuirá su valor en 1 segundo, es decir, pasará a ser de 4.6 unidades.
- La mediana también disminuirá en un segundo, pasando a ser 4.4 unidades.
- El rango intercuartílico mantendrá su valor. El primer y el tercer cuartil disminuirán en un segundo, por lo que su diferencia seguirá siendo de 3.1 unidades.
- La varianza no modificará su valor. Al restar a cada observación 1 segundo, la media disminuye en un segundo, por lo que las desviaciones a la media son las mismas que para los datos originales. En consecuencia la varianza se mantendrá en 3.1 unidades.
- El rango se mantendrá igual, ya que al restar a todos los valores una misma cantidad, la diferencia entre el máximo y el mínimo no cambia.

Pregunta 7

Parcialmente correcta

Se puntúa 2,06 sobre 2,50

Nota: la tolerancia es 0.1.

En un laboratorio de informática, se han realizado pruebas exhaustivas para evaluar el rendimiento de distintos sistemas operativos. La tabla siguiente recoge los datos correspondientes a dos métricas clave: la carga de CPU utilizada por el sistema (en porcentaje) y el tiempo de respuesta del sistema (en milisegundos), tal y como se registraron en 45 pruebas llevadas a cabo en el laboratorio:

```
carga 16 13 8 17 11 20 19 17 17 14 18 10 15 12 13 9 10 9 19 12 14 16 11 14 18 12 18 8 20 17 11 16 12 19 12 8 12 13 12 20 13 15 19 19
11
tiempo 108 121 142 106 125 90 95 103 105 117 102 134 114 123 120 133 134 139 96 127 114 106 128 116 100 121 101 141 91 107 129 105
127 94 121 139 125 118 122 89 119 113 93 93 127
```

Sea \(X\) la carga de CPU utilizada por el sistema e \(Y\) el tiempo de respuesta del sistema:

- Calcular el vector de medias:

  \(\begin{pmatrix} \bar{x}\\ \bar{y} \end{pmatrix}=\)

  14.2 ✔

  114.9556 ✔

- Calcular la matriz de varianzas y convarianzas:

  \(\begin{pmatrix} s_x^2 & s_{xy}\\ s_{xy} & s_y^2 \end{pmatrix}=\)

  13.04889 ✔ -53.79111 ✔

  -53.79111 ✔ 225.2869 ✔

- El coeficiente de correlación entre la carga de CPU utilizada por el sistema y el tiempo de respuesta del sistema es -0.9921007 ✔ . Este valor indica que la dependencia lineal entre las dos variables es: fuerte y negativa ✔ .

  Es decir, el tiempo de respuesta disminuye ✖ a medida que la carga de CPU utilizada por el sistema disminuye.

- La recta de regresión, obtenida mediante mínimos cuadrados, que explica el tiempo de respuesta del sistema en función de la carga de CPU utilizada es:

  \(y=\) 173.4919 ✔ \(+\) -4.122275 ✔ \(\cdot x\)

- La pendiente de la recta indica que

  si no hay carga de CPU, el tiempo medio de respuesta es de 173.49

  ✖ .

- El término independiente indica que

  por cada unidad que aumenta la carga de CPU utilizada, el tiempo medio de respuesta se incrementa en aproximadamente 173

  ✖ .

- El tiempo de respuesta del sistema para una carga de CPU de 14\(\%\) es: 115.78 ✔ .

- El tiempo de respuesta del sistema para una carga de CPU de 5\(\%\) es: no se puede determinar ✔ .

- La carga de CPU estimada cuando el tiempo de respuesta del sistema es de 127 milisegundos es: 11.2782 ✔ .

- El porcentaje de la variabilidad observada en el tiempo de respuesta del sistema que puede explicarse a partir de la carga de CPU es: 98.42638 ✔ \(\%\).

- Llamemos \(X\) a la carga de CPU utilizada por el sistema (en porcentaje) e \(Y\) al tiempo de respuesta del sistema (en milisegundos).

  Se tiene que:

$\begin{aligned} \bar{x}=\frac{639}{45}=14.2\\ \frac{\sum_{i=1}^{45}x_i^2}{45}\frac{9661}{45}=214.69\\ s_x^2=\frac{1}{n}\sum_{i=1}^{45}x_i^2- \bar{x}^2= 214.69-14.2^2=13.05\\ \bar{y}=\frac{5173}{45}=114.96\\ \frac{\sum_{i=1}^{45}y_i^2}{45}\frac{604803}{45}=13440.07\\ s_y^2=\frac{1}{n}\sum_{i=1}^{45}y_i^2- \bar{y}^2= 13440.07-114.96^2=225.29\\ \frac{\sum_{i=1}^{45}x_iy_i}{45}\frac{71036}{45}=1578.58\\ s_{xy}=\frac{1}{n}\sum_{i=1}^{45}x_iy_i- \bar{x}\bar{y}= 1578.58-14.2\cdot114.96=-53.79\\ \end{aligned}$

Podemos resumir estos valores indicando el vecor de medias y la matriz de varianzas y covarianzas de estos datos:

$\begin{pmatrix} \bar{x}\\ \bar{y} \end{pmatrix}=\begin{pmatrix} 14.2\\ 114.96 \end{pmatrix}$

$\begin{pmatrix} s_x^2&s_{xy}\\ s_{xy}&s_y^2 \end{pmatrix}=\begin{pmatrix} 13.05&-53.79\\ -53.79&225.29 \end{pmatrix}$

- El coeficiente de correlación entre X e Y es

  $r_{xy}=\frac{s_{xy}}{\sqrt{s_x^2s_y^2}}=\frac{-53.79}{\sqrt{13.05\cdot 225.29}}=-0.992$ es

  Puesto que se trata de un coeficiente de correlación negativo y con un módulo próximo a 1, indica que existe una fuerte a 1 dependencia lineal negativa entre X e Y. el tiempo de respuesta aumenta a medida que la carga de CPU disminuye.

- Debemos obtener la ecuación de la recta de regresión que explica Y en función de X: $\hat{y}=a+b\cdot x$

  El valor de la pendiente de esta recta es :

  $b=\frac{s_{xy}}{s_x^2}=\frac{-53.79}{-53.79}=-4.122$

  En cuanto al intercepto, su valor es:

  $a=\bar{y}-b \cdot\bar{x}=114.96--4.122\cdot 14.2=173.492$

  En consecuencia, la recta de regresión de Y sobre X es: $\hat{y}=173.492+-4.122\cdot x$

- Como acabamos de ver, la ecuación de la recta de regresión que explica Y en función de X es $\hat{y}=173.492+-4.122\cdot x$

  La pendiente de esta recta es $b=-4.122$, lo cual indica que si la carga de CPU utilizada por el sistema disminuye en un 1$\%$, entonces el incremento medio del número de muertes por cáncer es de 4.122 casos por cada 10.000 habitantes.

  Por su parte, el intercepto $a=173.492$ indicaría que, si no hay carga de CPU (es decir, es 0), el tiempo medio de respuesta sería de 173.492 milisegundos. Sin embargo, en la práctica no existen situaciones sin carga de CPU.

- Hemos visto que la ecuación de la recta de regresión que permite predecir el tiempo de respuesta a partir de la carga de CPU es $\hat{y}=173.492+-4.122\cdot x$. Por tanto, la predicción que proporciona la recta para el tiempo de respuesta del sistema por cáncer en una localidad situada a 14 de la refinería es $\hat{y}=173.492+-4.122\cdot 14=115.78$

  Es decir, 115.78 milisegundos.

- La recta de regresión no puede utilizarse para hacer predicciones usando valores de X que estén fuera del rango de los datos observados (es decir, fuera del rango de valores de X utilizados para construir la recta). Por tanto, no podemos realizar ninguna predicción fiable sobre el tiempo de respuesta del sistema en un sistema cuya carga de CPU es 5$\%$.

- Para predecir la carga de CPU utilizada por un sistema en el que se ha registrado un tiempo de respuesta de 127 milisegundos no podemos utilizar la recta antes calculada. En efecto, esta recta está diseñada para predecir el tiempo de respuesta del sistema en función del porcentaje de uso de CPU y no al revés. Debemos por tanto construir la recta de regresión que explica la carga de CPU utilizada por el sistema a partir de el tiempo de respuesta del mismo.

  La pendiente de esta recta es:

  $b2=\frac{s_{xy}}{s_y^2}=\frac{-53.79}{225.29}=-0.239$

  En cuanto al intercepto, su valor es:

  $a2=\bar{x}-b2 \cdot\bar{y}=14.2--0.239\cdot 114.96=41.648$

  En consecuencia, la recta de regresión de X sobre Y es:

  $\hat{x}=41.648+-0.239\cdot y$

  Luego la predicción que puede hacerse sobre la carga de CPU utilizada por un sistema en el que se ha registrado un tiempo de respuesta de 127 milisegundos será:

  $\hat{x}=41.648+-0.239\cdot 127=11.295$

- El coeficiente de determinación entre la carga de CPU utilizada por el sistema y el tiempo de respuesta del mismo es:

  $R^2=r_{xy}^2=-0.992^2=0.9843$

  Este valor indica que el porcentaje de la variabilidad observada en el tiempo de respuesta del sistema que puede explicarse a partir de la carga de CPU utilizada es de un 98.43$\%$.

Podemos usar el código R:

x=scan() ...datos de x...

y=scan() ...datos de y...

(n=length(x))

(mean_x=mean(x))

```
(mean_y=mean(y))
(varianza_x=var_x=var(x)*(n-1)/n)
(varianza_y=var(x)*(n-1)/n)
(covarianza_xy=cov(x,y)*(n-1)/n)
(r=cor(x,y))
ml=lm(y x)
summary(ml)
(a=ml$coefficients[1])
(b=ml$coefficients[2])
(est=a+b*12)
ml2=lm(x y)
summary(ml2)
(a2=ml2$coefficients[1])
(b2=ml2$coefficients[2])
(est2=a+b*139)
(R2=r^2)
```

Pregunta **8**

Parcialmente correcta

Se puntúa 1,00 sobre 1,50

Nota: la tolerancia para las respuestas numéricas es 0.001.

Un analista de datos en el mercado de valores utiliza dos test diferentes para evaluar oportunidades de inversión. Si una inversión es rentable, el primer test lo reconoce con una probabilidad del 0.6, mientras que el segundo test lo hace con una probabilidad de 0.8.

1. Supongamos que los tests son independientes. Si una inversión es rentable,
   - ¿Con qué probabilidad lo detecta únicamente uno de los dos tests? [0.44] ✔ .
   - ¿Con qué probabilidad no lo detecta ninguno de los tests? [0.08] ✔ .
2. Supongamos ahora que el segundo test sólo se aplica en los casos que el primer test ha dado como NO rentable. Si una inversión ha sido identificada como rentable, ¿con qué probabilidad ha sido clasificada por el primer test? [0.6] ✖ .

1. Sea P1 el suceso *la inversión rentable es detectada por el primer test* y P2 el suceso *la inversión rentable es detectada por el segundo test*. Entonces, $P(P1)=0.6$ y $P(P2)=0.8$.

   Si una inversión es rentable, la probabilidad de que lo detecten ambos tests es la probabilidad de la intersección y como son sucesos independientes: $P(P1\cap P2)=P(P1)\cdot P(P2) = 0.6\times0.8 = 0.48$

   y la probabilidad de que lo detecte al menos uno de los tests, es la probabilidad de la unión: $P(P1\cup P2)=P(P1)+P(P2)-P(P1\cap P2) = 0.6+0.8-0.48 =0.92$

   - Si una inversión es rentable, ¿con qué probabilidad lo detecta únicamente uno de los tests?

     Es la probabilidad de que lo detecte al menos uno de los tests, menos la probabilidad de que lo detecten los dos, esto es la probabilidad de la unión menos la probabilidad de la intersección: $P(P1\cup P2)-P(P1\cap P2)=0.92- 0.48 = 0.44$
   - Si una inversión es rentable, ¿con qué probabilidad no lo detecta ninguno de los tests?

     Es la probabilidad de que no ocurra ninguno de los dos sucesos, es decir, la probabilidad de que ocurra el suceso complementario de la unión: $1-P(P1\cup P2)=1-0.92=0.08$
2. Supongamos que el segundo test sólo se aplica en las inversiones que el primer test ha dado como NO rentable.

   Si $P1n$ y $P2n$ son los sucesos *la inversión se detecta en el primer y segundo test de revisión*, respectivamente.

   Si una inversión es rentable, la probabilidad de que lo detecte el primer test, es directamente la probabilidad inicial. $P(P1n)=P(P1)=0.6$.

   El segundo test lo detecta si el primero no lo detecta y el segundo sí, luego la probabilidad de que lo detecte el segundo test $P(P2n) = P(P1^c)\cdot P(P2)=(1-P(P1))\cdot P(P2) = (1-0.6)\cdot 0.8 = 0.32$ donde $P1^c$ es el suceso complementario de $P1$.

   Por lo tanto, si una inversión ha sido detectada como rentable, podemos usar el teorema de Bayes, para calcular la probabilidad de que haya sido detectada por el primer test:

   Si P1n es el suceso *el dato se detecta en el primer test* y PD es el suceso *la inversión rentable se detecta como tal*, tenemos que: $P(PD) = P(P1n) + P(P2n) = 0.6+0.32 = 0.92$ Por tanto, la probabilidad solicitada es la probabilidad de que ocurra $P1n$ condicionado a que ocurre PD: $P(P1n/PD) = \frac{P(P1n\cap PD)}{P(PD)} = \frac{P(P1n)}{P(PD)} = \frac{0.6}{0.92}=0.6522.$