TP2 – Regresión Lineal

Agustina Sol Rojas y Antonio Felix Glorioso Ceretti

Ejercicio 1.

Suponga que (x1, y1), … , (xn, yn) son pares observados generados por los siguientes modelos y deduzca los estimadores de mínimos cuadrados de β1 y β0 .

a) $Y = \beta 1x + \varepsilon$

Se utilizará el método de los mínimos cuadrados:

$$f(\beta 1) = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (yi - \beta 1xi)^2$$

$$\frac{\partial f(\beta 1)}{\partial \beta 1} = \sum_{i=1}^{n} 2 * (yi - \beta 1xi) * (-xi) = 0$$

$$= 2 \sum_{i=1}^{n} (-yixi + \beta 1xi^2) = 0$$

$$= \sum_{i=1}^{n} -yixi + \sum_{i=1}^{n} (\beta 1xi^2) = 0$$

$$= -\sum_{i=1}^{n} yixi + \beta 1 \sum_{i=1}^{n} xi^2 = 0$$

$$= \beta 1 \sum_{i=1}^{n} xi^2 = \sum_{i=1}^{n} yixi$$

$$\widehat{\beta 1} = \frac{\sum_{i=1}^{n} yixi}{\sum_{i=1}^{n} xi^2}$$

La recta de regresión estimada es $\hat{y}i = \widehat{\beta 1}x$

b) $Y = \beta 1(ax + c) + \beta 0 + \varepsilon$

$$Y = \beta 1X + \beta 0 + \varepsilon$$

$$\hat{y} = \widehat{\beta 1}x + \widehat{\beta 0}$$

$$y = \beta 1(ax + c) + \beta 0 + \varepsilon$$

$$y = a\beta 1x + \beta 1c + \beta 0 + \varepsilon$$

$$\beta 1' \qquad \beta 0'$$

$$y = \beta 1'x + \beta 0' + \varepsilon$$

$$\widehat{\beta 1} = \frac{Sxy}{Syy}$$

$$\widehat{\beta 0} = \hat{y} - \widehat{\beta 1}\bar{x}$$

$$\widehat{\beta 1}' = a\widehat{\beta 1}$$

$$\widehat{\beta 0}' = \widehat{\beta 1}c + \widehat{\beta 0}$$

La recta de regresión estimada es $\hat{y}' = \widehat{\beta 1}'x + \widehat{\beta 0}'$

Ejercicio 2.

Una cadena de supermercados financia un estudio sobre los gastos mensuales en alimentos, de familias de 4 miembros. La investigación se limitó a familias con ingresos netos entre $688.000 y $820.000, con lo cual se obtuvo la siguiente recta de estimación $\hat{y} = 0{,}85x - 18.000$.

y = gastos ; x = Ingresos

a) Estime los gastos en alimentos en un mes, para una familia de 4 miembros con un ingreso de $700.000

$$\hat{y} = 0{,}85 * 700000 - 18000 = 577000$$

b) Uno de los directivos de la compañía se preocupa por el hecho de que la ecuación aparentemente indica que para una familia que tiene un ingreso de $12.000 no gastaría nada en alimentos ¿Cuál sería su respuesta?

No se pueden estimar los gastos porque el ingreso está fuera del rango.

Ejercicio 3.

La empresa META quiere pronosticar el precio de sus acciones en función de los días en el periodo del 03/09/23 al 30/08/24, pero durante las fechas del 02/02/24 al 24/04/24

implementaron una serie de actualizaciones en sus distintas plataformas que dispararon el precio de sus acciones y querían saber en qué porcentaje afectaron dichas actualizaciones al ajuste y a la linealidad.

Utilizando los datos proporcionados en el archivo “META” haga los cálculos necesarios y responda.

Sugerencia: Realice dos análisis diferentes y para una de ellas desestimar los datos del periodo de actualización.

Sin actualización:

Cálculos auxiliares:

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i) * (\sum_{i=1}^{n} y_i)}{n} = 14896028.792634 - \frac{31878 * 106855.010034}{252}$$

$$= 14896028.792634 - 13517158.7693 = 1378870.02333$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 5366130 - \frac{1016206884}{252} = 5366130 - 4032567.0$$

$$= 1333563.0$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 47061434.07557989 - \frac{11417993169.366241}{252}$$

$$= 47061434.0755798 - 45309496.70383429 = 1751937.3717456013$$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{1378870.02333}{1333563.0} = 1.0339744154066948$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 424.0278175952381 - 1.0339744154066948 * 126.5$$

$$= 424.0278175952381 - 130.79776354894688$$

$$= 293.23005404629123$$

$$SS_R = Syy - \widehat{\beta_1} S_{xy} = 1751937.3717456013 - 1.0339744154066948 * 1378870.02333$$

$$= 1751937.3717456013 - 1425716.3262975523 = 326221.045448049$$

Ajuste y linealidad:

$$R^2 = 1 - \frac{SS_R}{Syy} = 1 - \frac{326221.045448049}{1751937.3717456013} = 1 - 0.18620588310357677$$
$$= 0.8137941168964232 * 100 = 81.37941\%$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{1378870.02333}{\sqrt{1333563.0 * 1751937.3717456013}} = \frac{1378870.02333}{\sqrt{2336318857277.179}}$$
$$= \frac{1378870.02333}{1528502.1613583604} = 0.9021053801504696$$

<u>Con actualización:</u>

Cálculos auxiliares:

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i) * (\sum_{i=1}^{n} y_i)}{n} = 1325517.5464050155$$
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 1312812.9538461538$$
$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 1397233.2607275099$$
$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = 1.009677382083747$$
$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 278.98103545163457$$

$$SS_R = Syy - \widehat{\beta_1}S_{xy} = 58888.17456722213$$

Ajuste y linealidad:

$$R^2 = 1 - \frac{SS_R}{Syy} = 0.9578537269170359 * 100 = 95.78537\%$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.9787000188602408$$

Dichas actualizaciones afectaron al ajuste un 14.40596% y a la linealidad un 7.65946387%

Ejercicio 4.

Los siguientes datos corresponden a los tiempos relativos en segundos que tardaron en ejecutarse seis programas elegidos al azar en el entorno Windows y en DOS:

	Programas					
Windows	2,5	7,1	5	8,5	7	8,1
DOS	2,3	7,1	4	8	6,6	5

a) Realizar el grafico de dispersión de los puntos

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i) * (\sum_{i=1}^{n} y_i)}{n} = 20.759999999999962$$

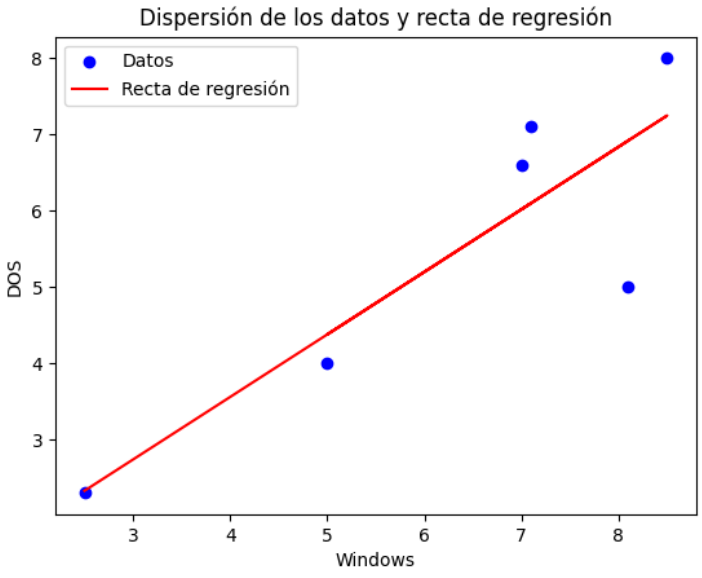
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 25.313333333333276$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 22.75999999999999$$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{20.759999999999962}{25.313333333333276} = 0.8201211482749543$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 5.5 - 0.8201211482749543 * 6.3666 = 5.5 - 5.22138330261$$
$$= 0.2785620226494574$$

Grafico



$$\hat{y} = \widehat{\beta_1}x + \widehat{\beta_0} = 0.8201211482749543x + 0.2785620226494574$$

b) Si un programa tarda 6 segundos en ejecutarse en Windows, ¿Cuánto tardara en ejecutarse en DOS?

Para saber cuánto tarada en ejecutar un programa de DOS si el de Windows tarda 6 segundos utilizaremos nuestra función de regresión linear:

$$\hat{y} = \widehat{\beta_1}x + \widehat{\beta_0} = 0.8201211482749543x + 0.2785620226494574$$

Reemplazaremos x con el valor 6 y resolveremos para $\hat{y}$

$$\hat{y} = 0.8201211482749543 * 6 + 0.2785620226494574 = 5.199288912299183$$

Por lo tanto, en DOS, se tardará unos 5.199 segundos

c) Se estima que los tiempos de Windows mejoraran reduciéndose en un 10% en los próximos años, estime la recta de regresión considerando esta mejora. Suponga que los tiempos DOS no se modifican.

Para calcular la mejora sobre los tiempos de Windows se deberá multiplicar un 0.9 a $\widehat{\beta_1}x$

$$\hat{y} = 0.9 * \widehat{\beta_1}x + \widehat{\beta_0} = 0.9 * 0.8201211482749543x + 0.2785620226494574$$
$$= 0.73810903344x + 0.2785620226494574$$

Ejercicio 5.

En la tabla siguiente, se muestran la variable y, rendimiento de un sistema informático, respecto a la variable x, numero de buffer:

x	5	10	15	20	25	5	10	15	20	25	5	10	15	20	25
y	9.6	20.1	29.9	39.1	50.0	9.6	19.4	29.7	40.3	49.9	10.7	21.3	30.7	41.8	51.2

A partir de la tabla anterior, se quiere ajustar la variable y como función de x.

a) Realizar el análisis de regresión de los datos (Estimación de la recta, Test de Hipótesis, Indicadores).

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i) * (\sum_{i=1}^{n} y_i)}{n} = 1514.000000000001$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 750.0$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 3064.324000000006$$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{1514.000000000001}{750.0} = 2.0186666666666677$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 30.22 - 2.0186666666666677 * 15 = -0.06000000000002004$$

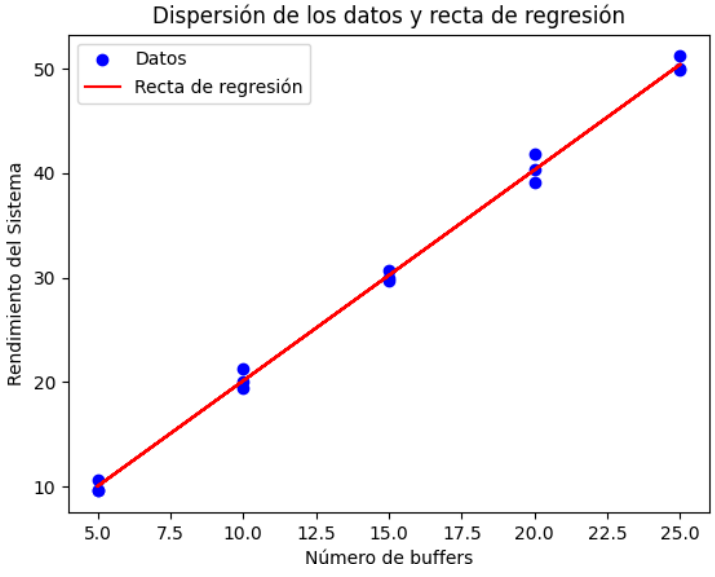
$$SS_R = Syy - \widehat{\beta_1} S_{xy} = 8.062666666669429$$

$$\widehat{\sigma^2} = \frac{SS_R}{n-2} = 0.6202051282053407$$

$$R^2 = 1 - \frac{SS_R}{Syy} = 0.9973688596027478$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.9986835632985795$$

<u>Grafico</u>



Estimación de la recta:

$$\hat{y} = \widehat{\beta_1}x + \widehat{\beta_0} = 2.0186666666666677x + (-0.06000000000002004)$$

Test de Hipótesis sobre β_1:

Deseamos probar la hipótesis de que la pendiente β_1 es igual a 0. Entonces supongamos las hipótesis:

$$H_0: \beta_1 = 0 \qquad H_1: \beta_1 \neq 0$$

Estadístico de prueba:

$$T = \frac{\widehat{\beta_1} - \theta}{\sqrt{\frac{\widehat{\sigma^2}}{S_{xx}}}} \sim t_{n-2}$$

Como regla de decisión se usará:

- Se rechaza H_0 si $|T| > t_{\frac{\alpha}{2},n-2}$
- Se acepta en caso contrario.

Se usará un nivel de significancia $\alpha = 0.05$.

$$T = \frac{2.0186666666666677-0}{\sqrt{\frac{0.6202051282053407}{750}}} = \frac{2.0186666666666677}{\sqrt{\frac{0.6202051282053407}{750}}} = \frac{2.0186666666666677}{\sqrt{0.00082694017}} = 70.1984571704$$

$$t_{\frac{\alpha}{2},n-2} = t_{\frac{0.05}{2},15-2} = t_{0.025,13} = 2{,}16037$$

Como $|70.1984571704| > 2{,}16037$ se rechaza H_0.

b) Comentar los resultados siguientes:

- Recta de regresión del rendimiento del sistema informático frente al número de buffers e interpretación de los coeficientes.

La recta de regresión se ajusta de manera casi perfecta a los valores dados, esto se puede ver en el grafico como también en el valor de R^2 (es muy cercano a 1). También se puede ver que r es muy cercano a 1 por lo que hay una gran correlación entre x e y.

- Contraste de hipótesis sobre la pendiente de la pendiente de la recta.

La hipótesis de la pendiente igualada a 0 fue rechazada por lo que se puede asumir que x tiene importancia al explicar la variabilidad en Y. También puede significar que el modelo lineal es adecuado, o que, aunque existe efecto lineal pueden obtenerse mejores resultados agregando términos polinomiales de mayor grado en x.

Ejercicio 6.

Determine si las siguientes relaciones son posibles o no y justifique su respuesta:

a. $\widehat{\sigma^2} = 0{,}2\,;\ n = 102\,;\ R^2 = 0{,}8\,;\ S_{yy} = 100$

$$R^2 = 1 - \frac{SS_R}{S_{yy}}$$
$$0{,}8 = 1 - \frac{SS_R}{100}$$
$$0{,}8 - 1 = -\frac{SS_R}{100}$$
$$20 = SS_R$$

$$\widehat{\sigma^2} = \frac{SS_R}{n-2} = \frac{20}{100} = 0{,}2$$

Si es posible ya que todos los valores están correctamente relacionados.

b. $\hat{y} = 7x + 4\,;\ \bar{x} = 10\,;\ \bar{y} = 64\,;\ r = -0{,}8$

No es posible ya que la pendiente $(7x)$ de la recta de regresión es positiva y r indica que la relación entre x e y es de índole negativa (es decir, la recta tendría que ser decreciente).

c. $\widehat{\beta_0} = 10{,}073\ ;\ \widehat{\beta_1} = -2{,}06\ ;\ \bar{x} = 8{,}5\ ;\ \bar{y} = 8{,}325$

$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 8{,}325 - (-2{,}06) * 8{,}5 = 25.835$

No es posible ya que el valor obtenido de $\widehat{\beta_0}$ mediante el uso de los otros indicadores es distinto al dado en el enunciado.

Ejercicio 7.

Indique si las siguientes afirmaciones son correctas o no. Justifique su respuesta:

a. $SS_R = S_{yy} - \widehat{\beta_0}S_{xy}$ $\qquad \frac{S_{xy}}{S_{xx}} = \widehat{\beta_1}$

$$SS_R = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = S_{yy} - \frac{S_{xy}}{S_{xx}} * S_{xy} = S_{yy} - \widehat{\beta_1}S_{xy}$$

La afirmación es incorrecta.

b. El error del intervalo de predicción es $\sqrt{n+1}$ veces mayor que el intervalo confianza para la respuesta media cuando $x^* = \bar{x}\ es\ igual\ (1 - \alpha)$.

$$ICM(\varepsilon) = t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}$$

$$IP(\varepsilon) = t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}$$

$x^* = \bar{x}\ es\ igual\ (1 - \alpha)$

$$\frac{ICM(\varepsilon)}{IP(\varepsilon)} = \frac{t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}}{t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}}$$

$$= \frac{\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}}{\sqrt{\widehat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}} = \frac{\sqrt{\frac{\widehat{\sigma^2}}{n}}}{\sqrt{\frac{\widehat{\sigma^2}}{n} * n + 1}}$$

$$= \frac{\sqrt{\frac{\widehat{\sigma^2}}{n}}}{\sqrt{\frac{\widehat{\sigma^2}}{n}} * \sqrt{n+1}} = \frac{1}{\sqrt{n+1}}$$

$$\frac{ICM(\varepsilon)}{IP(\varepsilon)} = \frac{1}{\sqrt{n+1}}$$

$$ICM(\varepsilon) = \frac{1}{\sqrt{n+1}} * IP(\varepsilon)$$

$$\sqrt{n+1} * ICM(\varepsilon) = \frac{1}{\sqrt{n+1}} * \sqrt{n+1} * IP(\varepsilon)$$

$$\sqrt{n+1} * ICM(\varepsilon) = IP(\varepsilon)$$

Esta afirmación es correcta ya que $IP(\varepsilon)$ siempre es $\sqrt{n+1}$ veces más grande que $ICM(\varepsilon)$ cuando $x^* = \bar{x}\ es\ igual\ (1 - \alpha)$.

c. El coeficiente de determinación R^2 indica el grado de relación lineal que existe entre la variable independiente y dependiente.

   Esta afirmación es incorrecta debido a que R^2 indica la relación entre la función de regresión lineal y nuestros datos. Lo que describe el grado de relación lineal que existe entre la variable independiente y dependiente es r (Coeficiente de Correlación Lineal).

d. El principio de mínimos cuadrados consiste en minimizar la suma de los residuos al cuadrado considerando la distancia perpendicular entre el valor observado y el estimado.

e. Esta afirmación es incorrecta debido a que el principio de mínimos cuadrados consiste en minimizar la suma de los residuos al cuadrado considerando la distancia vertical entre el valor observado y el estimado, no la perpendicular.

Ejercicio 8.

En un departamento de informática, un grupo de investigación dedicado al estudio de las comunicaciones por la red desea conocer la relación entre el tiempo de transmisión de un fichero y la información útil del mismo. Para ello se han hecho algunos experimentos en los que se enviaban paquetes de distintas longitudes (bytes) de información útil y se median los tiempos (en milisegundos) que tardaban desde el momento en que se enviaban hasta que llegaban al servidor. Los resultados del experimento se resumen en los siguientes estadísticos:

$$S_{xx} = 47.990\,;\ \bar{x} = 194\,;\ \widehat{\beta_0} = 27{,}3275$$
$$\sum x_i^2 = 424.350\,;\ \sum x_i y_i = 183.760\,;\ \sum y_i^2 = 81.715$$

Se pide estudiar la relación entre las variables tiempo (y) y longitud (x) de los ficheros. Para ello, se pide:

a. Obtener la recta de regresión del tiempo en función de la longitud de los ficheros. Interpretar los resultados obtenidos.

   Del enunciado se sabe:

$$S_{xx} = 47.990\,;\ \bar{x} = 194\,;\ \widehat{\beta_0} = 27{,}3275$$
$$\sum x_i^2 = 424.350\,;\ \sum x_i y_i = 183.760\,;\ \sum y_i^2 = 81.715$$

   La fórmula de S_{xx} se puede expresar de la siguiente manera.

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right) * \left(\sum_{i=1}^{n} x_i\right)}{n}$$
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)}{n} * \sum_{i=1}^{n} x_i$$
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \bar{x} * \sum_{i=1}^{n} x_i$$

Reemplazando valores se puede obtener $\sum_{i=1}^{n} x_i$

$$47.990 = 424.350 - 194 * \sum_{i=1}^{n} x_i$$

$$-47.990 + 424.350 = 194 * \sum_{i=1}^{n} x_i$$

$$\frac{376.36}{194} = \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i = 1.94$$

A partir de $\sum_{i=1}^{n} x_i$ y de $\bar{x}$, se puede obtener n:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Reemplazando valores

$$194 = \frac{1.94}{n}$$

$$n = 100$$

Continuara…

b. Indicar el valor que toma el coeficiente de determinación y correlación lineal. Interpretar los resultados.

c. Estudiar la significación del modelo.

d. Obtener el intervalo de confianza, al 95%, para la pendiente de la recta.

e. ¿Cuál será el tiempo de transmisión para un fichero que tiene una longitud 250 bytes?

Ejercicio 9.

De un análisis de regresión realizada sobre un Dataset, el cual consiste en un pequeño relevamiento del tiempo que demandan las llamadas a servicio técnico de una empresa

(x) y la cantidad de unidades de hardware reparadas (y), se sabe que el $IC(\beta_0) = (-0{,}4348\ ;\ -0{,}4248)$, que la estimación de la pendiente es 12 veces el error que se comete al estimar la verdadera ordenada al origen con $\widehat{\beta_0}$ y que la proporción de variación total observada no explicada por el modelo de regresión lineal es tan solo del 2%.

A partir de los datos proporcionados determinar:

a. El error que se comete al estimar la verdadera ordenada al origen con $\widehat{\beta_0}$.

$IC(\beta_0) = (-0.4348\ ;\ -0.4248)$

A partir de los valores dados por enunciado y siguiendo la fórmula de cálculo para los extremos del intervalo se tiene el siguiente sistema de ecuaciones.

1. $\widehat{\beta_0} - t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = -0{,}4348$

2. $\widehat{\beta_0} + t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = -0.4248$

Se busca el valor de $\widehat{\beta_0}$ en la ecuación 1. para reemplazarlo en la ecuación 2.

1. $\widehat{\beta_0} - t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = -0{,}4348$

   $\widehat{\beta_0} = -0{,}4348 + t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$

2. $\widehat{\beta_0} + t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = -0.4248$

   $-0.4348 + t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} + t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = -0.4248$

   $2 * t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = -0.4248 + 0.4348$

   $2 * t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = 0.01$

$$t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}=\frac{0.01}{2}=0.005$$

De esta forma, se afirma que el error que se comete al estimar la verdadera ordenada al origen vale 0.005

b. La recta de regresión estimada

Sabemos por el subpunto a. que $\widehat{\beta_0}$ es:

$$\widehat{\beta_0}=-0.4348+\ t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

Como ya sabemos el error, este se puede reemplazar:

$$\widehat{\beta_0}=-0.4348+0.005$$

$$\widehat{\beta_0}=-0.4298$$

Y por enunciado tenemos que:

$$\widehat{\beta_1}=12*\left(t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}\right)$$

$$\widehat{\beta_1}=12*0.005$$

$$\widehat{\beta_1}=0.06$$

Por lo tanto, la recta de regresión estimada es:

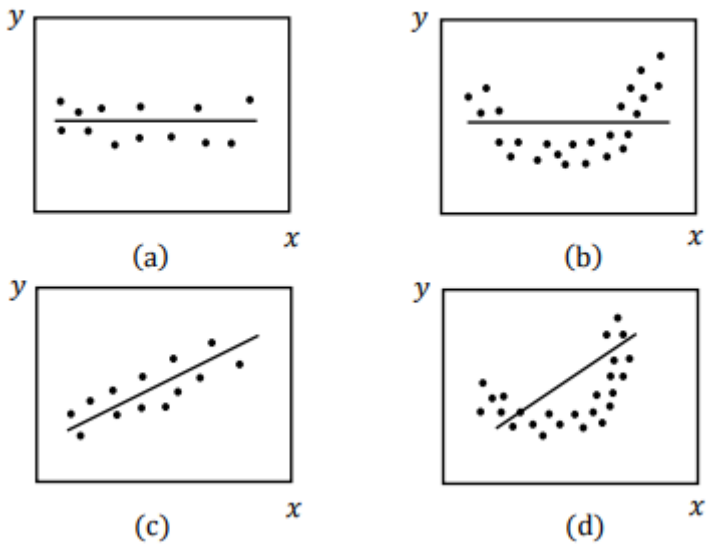
$$\hat{y}=\widehat{\beta_1}x+\widehat{\beta_0}=0.06x-0.4298$$

c. La bondad del ajuste

$$R^2=1-0.02=0.98$$

Ejercicio 10.

Observando los siguientes gráficos de regresión y considerando las hipótesis $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ Indique para cada una, si se acepta o no H_0 y la implicancia de esta.



En los gráficos de (a) y (c) se rechaza H_0 ya que x tiene importancia al explicar la variabilidad en y (hay una relación lineal entre x e y). Mientras que en los gráficos de (b) y (d) pueden obtenerse mejores resultados agregando términos polinomiales de mayor grado en x, es decir, se acepta H_0.

Ejercicio 11.

La autoridad aeronáutica argentina realizó un estudio de operaciones de aerolíneas, en 18 compañías, que reveló que la relación entre el número de pilotos empleados y el número de aviones en servicio tenía una pendiente de 4.3. Estudios anteriores indicaban que la pendiente de esta relación era 4.0. Si se calculó que la desviación estándar de la de pendiente de regresión es 0.17, ¿hay razones para creer, a un nivel de significancia de 0.05, que la pendiente verdadera ha cambiado?

Del enunciado se obtuvo:

$n = 18$

$\widehat{\beta_1} = 4.3$

$\widehat{\beta_{10}} = 4$

$\sqrt{\frac{\widehat{\sigma^2}}{S_{xx}}} = 0.17$

Se expresa la hipótesis que queremos plantear como:

$$H_0: \beta_1 = 4 \; vs \; H_1: \beta_1 \neq 4$$

Como estadístico de prueba utilizaremos:

$$T = \frac{\widehat{\beta_1} - \theta}{\sqrt{\frac{\widehat{\sigma^2}}{S_{xx}}}} \sim t_{n-2}$$

Regla de decisión:

- Se rechaza H_0 si $|T| > t_{\frac{\alpha}{2},n-2}$
- Se acepta en caso contrario.

Se calcula T y se usará un nivel de significancia $\alpha = 0.05$.

$T = \frac{4.3 - 4}{0.17} = 1.76470588235$

$t_{\frac{\alpha}{2},n-2} = t_{\frac{0.05}{2},18-2} = t_{0.025,16} = 2.11991$

Como $|1.76470588235| < 2.11991$ se acepta H_0. Es decir, no hay evidencia para saber si, a un nivel de significancia de 0.05, la pendiente verdadera ha cambiado.

Ejercicio 12.

Un horticultor inventó una escala para medir la frescura de rosas que fueron empacadas y almacenadas durante periodos variables antes de trasplantarlas. La medición y de

frescura y el tiempo x en días que la rosa está empacada y almacenada antes de trasplantarla, se dan a continuación.

x	5	5	10	10	15	15	20	20	25	25
y	15,3	16,8	13,6	13,8	9,8	8,7	5,5	4,7	1,8	1,0

a. ¿Hay suficiente evidencia para indicar que la frescura está linealmente relacionada con el tiempo de almacenaje?

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i) * (\sum_{i=1}^{n} y_i)}{n} = -379.0$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 500.0$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 291.93999999999994$$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{-379.0}{500.0} = -0.758$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 91.0 - (-0.758) * 150 = 20.47$$

$$SS_R = Syy - \widehat{\beta_1} S_{xy} = 4.657999999999959$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-379.0}{\sqrt{500.0 * 291.93999999999994}} = -0.9919902553515021$$

Si hay suficiente evidencia para para indicar que la frescura está linealmente relacionada con el tiempo de almacenaje ya que r es cercano a -1 → Correlación inversa perfecta.

b. Estime mediate un intervalo de 98% el descenso de frescura de las rosas por cada día que pasa.

Se utilizará α con valor:

$100\% * (1 - \alpha) = 98\%$

$\alpha = 0.02$

Se desea obtener un intervalo de confianza de 98% para la pendiente 1 de la línea de regresión verdadera por lo tanto se usará:

$$\left[\widehat{\beta_1} - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\widehat{\sigma^2}}{S_{xx}}}\ ;\ \widehat{\beta_1} + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\widehat{\sigma^2}}{S_{xx}}}\right]$$

Cálculos auxiliares:

$$\widehat{\sigma^2} = \frac{SS_R}{n-2} = \frac{4.657999999999959}{10-2} = 0.58225$$

$$t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\widehat{\sigma^2}}{S_{xx}}} = t_{\frac{0.02}{2},10-2}\sqrt{\frac{0.58225}{500.0}} = 2.89646 * 0.03412477106$$

$$= 0.09884103438$$

Así nos queda el IC con valores:

$$[-0.758 - 0.09884103438\ ;\ -0.758 + 0.09884103438]$$

$$= [-0.85684103438\ ;\ -0.65915896562]$$

c. Estime mediante un intervalo de 98% la frescura de las rosas cuando no han sido almacenada ni empacada.

El rango de valores que toma $Dom(x) = [5, 25]$ por lo tanto si queremos calcular un IC con $x^* = 0$ debemos utilizar el IP para futuras observaciones de Y^*. El cual es:

$$\left[\widehat{Y^*} - t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}\ ;\ \widehat{Y^*} + t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}\right]$$

Cálculos auxiliares:

$$\widehat{Y^*} = \widehat{\beta_0} + \widehat{\beta_1}x^*$$

$$\widehat{Y^*} = 20.47 - 0.758 * 0 = 20.47$$

$$t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]} =$$

$$2.89646 * \sqrt{0.58225 * \left[1 + \frac{1}{10} + \frac{(0 - 15)^2}{500.0}\right]} \approx$$

$$2.89646 * 0.94999 = 2.7516080354$$

Así nos queda el IP con valores:

$[20.47 - 2.7516080354\ ;\ 20.47 + 2.7516080354]$

$= [17.7183919646\ ;\ 23.2216080354]$

d. Estime la medición de frescura media para un tiempo de almacenaje de 14 días con un intervalo de confianza de 95%.

Si queremos calcular la medición de frescura media para un tiempo de almacenaje de 14 días con un intervalo de confianza de 95% se deberá usar el ICM con formula:

$$\left[\widehat{\beta_0} + \widehat{\beta_1}x^* - t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}\ ;\ \widehat{\beta_0} + \widehat{\beta_1}x^* + t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}\right]$$

Se utilizará α con valor:

$100\% * (1 - \alpha) = 95\%$

$\alpha = 0.05$

Cálculos auxiliares:

$\widehat{Y^*} = \widehat{\beta_0} + \widehat{\beta_1}x^*$

$\widehat{Y^*} = 20.47 - 0.758 * 14 = 9.858$

$$t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]} =$$

$$2.89646 * \sqrt{0.58225 * \left[\frac{1}{10} + \frac{(14 - 15)^2}{500.0}\right]} \approx$$

$2.3060 * 0.24369 = 0.56194914$

Así nos queda el ICM con valores:

$[9.858 - 0.56194914\ ;\ 9.858 + 0.56194914] = [9.29605086\ ;\ 10.41994914]$

Ejercicio 13.

Un fabricante de teléfonos celulares está probando dos tipos de baterías para ver cuánto duran con una utilización normal. La siguiente tabla contiene los datos provisionales:

Horas de uso diario	2	1,5	1	0,5
Vida aproximada (meses) Litio	3,1	4,2	5,1	6,3
Vida aproximada (meses) Alcalina	1,3	1,6	1,8	2,2

a. Desarrolle dos ecuaciones de estimación lineales, una para pronosticar la vida del producto basada en el uso diario con las baterías de litio y otra para las baterías alcalinas

Ecuación de estimación lineal para Litio:

$\hat{y} = \widehat{\beta_1}x + \widehat{\beta_0} = -2.1x + 7.3$

Ecuación de estimación lineal para Alcalina:

$\hat{y} = \widehat{\beta_1}x + \widehat{\beta_0} = -0.5799999999999997x + 2.4499999999999997$

b. ¿Cuál de las dos estimaciones anteriores se ajusta mejor a los datos?

El coeficiente de determinación para Litio:

$R^2 = 0.9972862957937605 * 100 = 99.72862\%$

El coeficiente de determinación para Alcalina:

$R^2 = 0.9836257309941506 * 100 = 98.36257\%$

La ecuación de estimación lineal para Litio se ajusta mejor a los datos ya que el coeficiente de determinación es mayor al de Alcalina.

c. Encuentre un intervalo para la estimación del 90% para la vida (en meses) con 1,25 horas de uso diario, para cada tipo de batería. ¿Puede la compañía asegurar algo respecto a qué batería proporciona la vida más larga según estos números?

¿Preguntar? Usar IP o ICM (Usamos ICM acá)

Si queremos calcular un intervalo para la estimación del 90% para la vida (en meses) con 1,25 horas de uso diario, para cada tipo de batería, se deberá usar el ICM con formula:

$$(\widehat{\beta_0} + \widehat{\beta_1}x^* - t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}\ ;\ \widehat{\beta_0} + \widehat{\beta_1}x^* + t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]})$$

Se utilizará α con valor:

$100\% * (1 - \alpha) = 90\%$

$\alpha = 0.1$

Cálculos auxiliares para Litio:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 1.25$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1.25$$

$$\widehat{\sigma^2} = \frac{SS_R}{n-2} = 0.007499999999994511$$

$$\widehat{Y^*} = \widehat{\beta_0} + \widehat{\beta_1}x^*$$

$$\widehat{Y^*} = -2.1 * 1.25 + 7.3 = 4.675$$

$$t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]} =$$

$$1.53321 * \sqrt{0.007499999999994511 * \left[\frac{1}{4} + \frac{(1.25 - 1.25)^2}{1.25}\right]} \approx$$

$1.53321 * 0.04330 = 0.066387993$

Cálculos auxiliares para Alcalina:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 1.25$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1.25$$

$$\widehat{\sigma^2} = \frac{SS_R}{n-2} = 0.0035000000000003084$$

$\widehat{Y^*} = \widehat{\beta_0} + \widehat{\beta_1} x^*$

$\widehat{Y^*} = -0.5799999999999997 * 1.25 + 2.4499999999999997 = 1.725$

$t_{\frac{\alpha}{2},n-2} \sqrt{\widehat{\sigma^2}\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]} =$

$1.53321 * \sqrt{0.0035000000000003084 * \left[\frac{1}{4} + \frac{(1.25 - 1.25)^2}{1.25}\right]} \approx$

$1.53321 * 0.02958 = 0.0453523518$

Así nos queda los ICM con valores:

Litio $= [4.675 - 0.066387993\ ;\ 4.675 + 0.066387993] =$

$[4.608612007\ ;\ 4.741387993]$

Alcalina $= [1.725 - 0.0453523518\ ;\ 1.725 + 0.0453523518] =$

$[1.679647648\ ;\ 1.770352352]$

La que proporciona la vida mas larga es la batería de Litio ya que dará un uso de vida diario (en meses) mayor al intervalo de la Alcalina.

d. El fabricante considera realizar una batería compuesta por los dos tipos de batería y pide para ello que se estime la ecuación lineal para pronosticar las horas de uso diario basada en la vida aproximada (en meses).

Se calculará una regresión lineal múltiple teniendo como x_1 a la vida aproximada (meses) Litio y x_2a la vida aproximada (meses) Alcalina. También tomares como y a las horas de uso diario.

El sistema de ecuaciones que se tendrá en cuenta para resolverlo será:

$$\left\{\begin{array}{c} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{1i}x_{2i} = \sum_{i=1}^{n} x_{1i}y_i \\ \beta_0 \sum_{i=1}^{n} x_{2i} + \beta_1 \sum_{i=1}^{n} x_{1i}x_{2i} + \beta_2 \sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} x_{2i}y_i \end{array}\right\}$$

Reemplazando valores queda:

$$\left\{\begin{array}{c} 4\beta_0 + \beta_1 18.7 + \beta_2 6.9 = 5 \\ \beta_0 18.7 + \beta_1 92.95 + \beta_2 33.79 = 20.75 \\ \beta_0 6.9 + \beta_1 33.79 + \beta_2 12.33 = 7.9 \end{array}\right\}$$

En forma matricial:

$$\begin{pmatrix} 4 & 18.7 & 6.9 \\ 18.7 & 92.95 & 33.79 \\ 6.9 & 33.79 & 12.33 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 20.75 \\ 7.9 \end{pmatrix}$$

Resolviendo el producto matricial:

$$\begin{pmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_2} \end{pmatrix} = \begin{pmatrix} 3.01384 \\ -0.76989 \\ 1.06401 \end{pmatrix}$$

Entonces la ecuación de regresión estimada será:

$\hat{y} = 3.01384 + -0.76989x_1 + 1.06401x_2$

e. ¿Mejora la estimación utilizando los dos tipos de batería juntas que por separado? Explique.

¿Preguntar? No se pueden comparar porque en a), b) y c) Estimamos la vida aproximada y en el d) estimamos el uso diario.

Ejercicio 14.

Una Empresa de desarrollo de software le pide relacionar sus Ventas en función del número de pedidos de los tipos de software que desarrolla (Sistemas, Educativos y Automatizaciones Empresariales), para atender 10 proyectos en el presente año. En la Tabla representa Y (Ventas miles de S/.) e X (Nº pedidos de sistemas), W (Nº de pedidos de Aplicaciones Educativas) y Z (Nº de pedidos de Automatizaciones empresariales).

y	440	455	470	510	506	480	460	500	490	450
x	50	40	35	45	51	55	53	48	38	44
w	105	140	110	130	125	115	100	103	118	98
z	75	68	70	64	67	72	70	73	69	74

a) Mediante un software a elección estime la ecuación de regresión múltiple para cumplir con el requerimiento de la empresa.

$x \rightarrow x_1$; $w \rightarrow x_2$; $z \rightarrow x_3$

Utilizando un código de Python que subiremos al git de Agus (:3) el resultado de la ecuación de regresión estimada es:

$\hat{y} = 934.80840 + 0.68150x_1 - 0.44629x_2 - 6.25261x_3$

b) La empresa quiere tener indicadores para asegurarse que la ecuación estimada se ajusta bien a los datos y si la relación lineal es la más correcta. ¿Cuáles recomendaría? Calcule las mismas y comente.

Se utilizará el coeficiente de determinación ajustado y el coeficiente de correlación:

$R_a^2 = 1 - (1 - R^2)\left(\frac{\mathrm{n} - \mathrm{k}}{n - k - 1}\right) = 0.3580668560486462$

$r_a = \sqrt{R_a^2} = 0.5983868782390254$

Se puede comentar que el modelo no se ajusta a los datos que se dan y que hay una linealización entre las variables x e y de carácter normal.

Ejercicio 15.

En la Facultad de Sistemas Informáticos se quiere entender los factores de aprendizaje de los alumnos que cursan la asignatura de PHP, para lo cual se escoge al azar una muestra de 15 alumnos y ellos registran notas promedios en las asignaturas correlativas de Algoritmos, Base de Datos y Programación como se muestran en el siguiente cuadro.

PHP	Algoritmos	Base de Datos	Programación
13	15	15	13
13	14	13	12
13	16	13	14
15	20	14	16
16	18	18	17
15	16	17	15
12	13	15	11
13	16	14	15
13	15	14	13
13	14	13	10
11	12	12	10
14	16	11	14
15	17	16	15
15	19	14	16
15	13	15	10

a. Construir un modelo para determinar la dependencia que exista de aprendizaje reflejada en las notas de la asignatura de PHP, conociendo las notas de las asignaturas Algoritmos, Base de Datos y Programación.

   Continuara...

b. Si más el 80% del aprendizaje del Curso de PHP no puede ser explicado mediante las notas obtenidas por las asignaturas de Algoritmos, Base de Datos y Programación, se destinarán más recursos a estas asignaturas para obtener mejores resultados. ¿Cuál es seria su respuesta?

   Continuara...