Juan Menduiña

Trabajo Práctico N° 2:
Regresión Lineal.

Ejercicio 1.

Suponer que (x_1, y_1), ... , (x_n, y_n) son pares observados generados por los siguientes modelos y deducir los estimadores de mínimos cuadrados de β_1 y β_0.

(a) $y= \beta_1 x + \varepsilon$.

$f(\beta_1)= \sum_{i=1}^{n}(y_i - \beta_1 x_i)^2$.

$\frac{\partial f(\beta_1)}{\partial \beta_1}= 0$
$\sum_{i=1}^{n} 2(y_i - \beta_1 x_i)(-x_i)= 0$
$-2 \sum_{i=1}^{n}(y_i - \beta_1 x_i)\, x_i= 0$
$\sum_{i=1}^{n}(y_i x_i - \beta_1 x_i^2)= \frac{0}{-2}$
$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \beta_1 x_i^2= 0$
$\sum_{i=1}^{n} y_i x_i - \beta_1 \sum_{i=1}^{n} x_i^2= 0$
$\beta_1 \sum_{i=1}^{n} x_i^2= \sum_{i=1}^{n} y_i x_i$
$\hat{\beta}_1= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$.

(b) $y= \beta_1 (ax + c) + \beta_0 + \varepsilon$.

$f(\beta_0, \beta_1)= \sum_{i=1}^{n}\{y_i - [\beta_1(ax_i + c) + \beta_0]\}^2$
$f(\beta_0, \beta_1)= \sum_{i=1}^{n}[y_i - \beta_1(ax_i + c) - \beta_0]^2$.

$\frac{\partial f(\beta_0)}{\partial \beta_0}= 0$
$\sum_{i=1}^{n} 2\,[y_i - \beta_1(ax_i + c) - \beta_0]\,(-1)= 0$
$-2 \sum_{i=1}^{n}[y_i - \beta_1(ax_i + c) - \beta_0]= 0$
$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_1(ax_i + c) - \sum_{i=1}^{n} \beta_0= \frac{0}{-2}$
$n\overline{y} - \beta_1 \sum_{i=1}^{n}(ax_i + c) - n\beta_0= 0$
$n\overline{y} - \beta_1 (\sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} c) - n\beta_0= 0$
$n\overline{y} - \beta_1 (a \sum_{i=1}^{n} x_i + nc) - n\beta_0= 0$
$n\overline{y} - \beta_1 (an\overline{x} + nc) - n\beta_0= 0$
$n\overline{y} - n\beta_1 (a\overline{x} + c) - n\beta_0= 0$
$n\,[\overline{y} - \beta_1 (a\overline{x} + c) - \beta_0]= 0$
$\overline{y} - \beta_1 (a\overline{x} + c) - \beta_0= \frac{0}{n}$
$\overline{y} - \beta_1 (a\overline{x} + c) - \beta_0= 0$
$\hat{\beta}_0= \overline{y} - \hat{\beta}_1 (a\overline{x} + c)$.

$\frac{\partial f(\beta_1)}{\partial \beta_1}= 0$

$\sum_{i=1}^{n} 2\,\{y_i - \beta_1\,(ax_i + c) - [\overline{y} - \beta_1\,(a\overline{x} + c)]\}\,[-(ax_i + c)] = 0$
$-2\sum_{i=1}^{n}[y_i - \beta_1\,(ax_i + c) - \overline{y} + \beta_1\,(a\overline{x} + c)]\,(ax_i + c) = 0$
$\sum_{i=1}^{n}\{(y_i - \overline{y}) - \beta_1\,[(ax_i + c) - (a\overline{x} + c)]\}\,(ax_i + c) = \frac{0}{-2}$
$\sum_{i=1}^{n}\{(y_i - \overline{y})(ax_i + c) - \beta_1\,[(ax_i + c) - (a\overline{x} + c)](ax_i + c)\} = 0$
$\sum_{i=1}^{n}(y_i - \overline{y})(ax_i + c) - \sum_{i=1}^{n}\beta_1\,[(ax_i + c) - (a\overline{x} + c)](ax_i + c) = 0$
$\sum_{i=1}^{n}(y_i - \overline{y})(ax_i + c) - \beta_1\sum_{i=1}^{n}[(ax_i + c) - (a\overline{x} + c)](ax_i + c) = 0$
$\beta_1\sum_{i=1}^{n}[(ax_i + c) - (a\overline{x} + c)](ax_i + c) = \sum_{i=1}^{n}(y_i - \overline{y})(ax_i + c)$
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(ax_i + c)}{\sum_{i=1}^{n}[(ax_i + c) - (a\overline{x} + c)](ax_i + c)}$
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})[(ax_i + c) - (a\overline{x} + c)]}{\sum_{i=1}^{n}[(ax_i + c) - (a\overline{x} + c)]^2}$
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(ax_i + c - a\overline{x} - c)}{\sum_{i=1}^{n}(ax_i + c - a\overline{x} - c)^2}$
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(ax_i - a\overline{x})}{\sum_{i=1}^{n}(ax_i - a\overline{x})^2}$
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})[a(x_i - \overline{x})]}{\sum_{i=1}^{n}[a(x_i - \overline{x})]^2}$.

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, con y= gastos; x= ingresos.

(a) *Estimar 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
$\hat{y}$= 595000 - 18000
$\hat{y}$= 577000.

Por lo tanto, los gastos en alimentos en un mes para una familia de 4 miembros con un ingreso de $700.000 son $577.000.

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

La respuesta sería que la estimación fue realizada con familias con ingresos netos entre $688.000 y $820.000, por lo que la recta de regresión no es adecuada para ingresos fuera de este rango considerado en la investigación.

Ejercicio 3.

La empresa META quiere pronosticar el precio de sus acciones en función de los días en el período 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", hacer los cálculos necesarios y responder. Sugerencia: Realizar dos análisis diferentes y, para una de ellos, desestimar los datos del período de actualización.

Estimación para período 03/09/23-01/02/24:

```
Linear regression                               Number of obs     =        106
                                                F(1, 104)         =     432.68
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8279
                                                Root MSE          =      11.99

------------------------------------------------------------------------------
             |               Robust
   PrecioUSD | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        Dias |   .8513341   .0409277    20.80   0.000     .7701729    .9324954
       _cons |   285.4107   2.142445   133.22   0.000     281.1622    289.6592
------------------------------------------------------------------------------
```

Estimación para período 25/04/24-30/08/24:

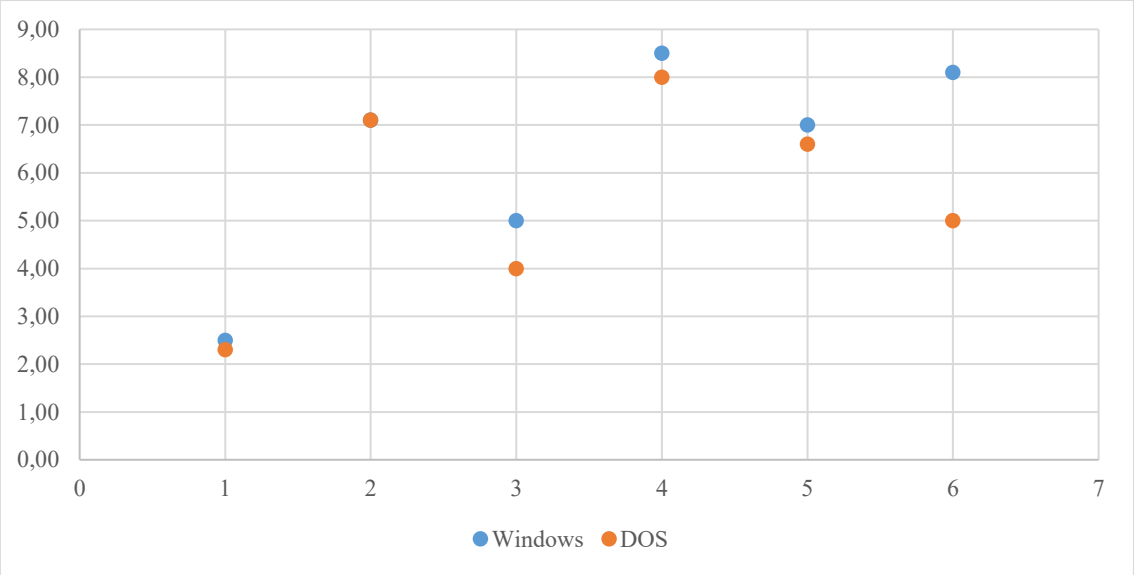
```
Linear regression                               Number of obs     =        106
                                                F(1, 104)         =     432.68
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8279
                                                Root MSE          =      11.99

------------------------------------------------------------------------------
             |               Robust
   PrecioUSD | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        Dias |   .8513341   .0409277    20.80   0.000     .7701729    .9324954
       _cons |   285.4107   2.142445   133.22   0.000     281.1622    289.6592
------------------------------------------------------------------------------
```

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:

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

(a) *Realizar el gráfico de dispersión de los puntos.*



(b) *Si un programa tarda 6 segundos en ejecutarse en Windows, ¿cuánto tardará en ejecutarse en DOS?*

Recta de regresión:

DOS= 0,28 + 0,82 Windows.

Si un programa tarda 6 segundos en ejecutarse en Windows, tardará 5,2 segundos en ejecutarse en DOS.

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

Nueva recta de regresión:

DOS= 0,28 + 0,91 Windows.

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).*

```
Linear regression                               Number of obs     =         15
                                                F(1, 13)          =    6050.94
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9974
                                                Root MSE          =     .38961

------------------------------------------------------------------------------
             |               Robust
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |   .4940731   .0063516    77.79   0.000     .4803514    .5077948
       _cons |   .0691115   .1932975     0.36   0.726    -.3484823    .4867054
------------------------------------------------------------------------------
```

(b) *Comentar los resultados siguientes:*

- *Recta de regresión del rendimiento del sistema informático frente al número de buffers e interpretación de los coeficientes.*
- *Contraste de hipótesis sobre la pendiente de la recta.*
- *Coeficiente de determinación y correlación lineal.*

Recta de regresión:

y= 0,691115 + 0,4940731 x.

Contraste de hipótesis sobre la pendiente de la recta:

La pendiente de la recta es estadísticamente significativa al 1%.

Coeficiente de determinación y correlación lineal:

R^2= 0,9974.
r= 0,9987.

Ejercicio 6.

Determinar si las siguientes relaciones son posibles o no y justificar la respuesta:

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

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

$\hat{\sigma}^2 = \frac{SCE}{n-2}$
$\hat{\sigma}^2 = \frac{20}{102-2}$
$\hat{\sigma}^2 = \frac{20}{100}$
$\hat{\sigma}^2 = 0,2.$

Por lo tanto, es una relación posible.

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

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
$\hat{\beta}_0 = 64 - 7 * 10$
$\hat{\beta}_0 = 64 - 70$
$\hat{\beta}_0 = -6.$

Por lo tanto, no es una relación posible.

(c) $\hat{\beta}_0 = 10,073;\ \hat{\beta}_1 = -2,06;\ \bar{x} = 8,5;\ \bar{y} = 8,325.$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
$\hat{\beta}_0 = 8,325 - (-2,06) * 8,5$
$\hat{\beta}_0 = 8,325 - (-17,51)$
$\hat{\beta}_0 = 8,325 + 17,51$
$\hat{\beta}_0 = 25,835.$

Por lo tanto, no es una relación posible.

Ejercicio 7.

Indicar si las siguientes afirmaciones son correctas o no. Justificar la respuesta:

(a) $SS_R = S_{yy} - \hat{\beta}_0 \, S_{xy}$.

SCR= $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

SCR= STC - SCE
SCR= $\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
SCR= $S_{yy} - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

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* $x^* = \bar{x}$ *e igual (1 -* α*).*

Considerando $x^* = \bar{x}$, se tiene:

$$\frac{SE(y)}{SE(\hat{y})} = \frac{\sqrt{\hat{\sigma}^2(1+\frac{1}{n})}}{\sqrt{\hat{\sigma}^2\frac{1}{n}}}$$

$$\frac{SE(y)}{SE(\hat{y})} = \frac{\sqrt{\hat{\sigma}^2}\sqrt{1+\frac{1}{n}}}{\sqrt{\hat{\sigma}^2}\sqrt{\frac{1}{n}}}$$

$$\frac{SE(y)}{SE(\hat{y})} = \frac{\sqrt{\frac{n+1}{n}}}{\frac{\sqrt{1}}{\sqrt{n}}}$$

$$\frac{SE(y)}{SE(\hat{y})} = \frac{\frac{\sqrt{n+1}}{\sqrt{n}}}{\frac{1}{\sqrt{n}}}$$

$$\frac{SE(y)}{SE(\hat{y})} = \frac{\sqrt{n+1}}{1}$$

$$\frac{SE(y)}{SE(\hat{y})} = \sqrt{n+1}.$$

La afirmación es CORRECTA.

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

La afirmación es INCORRECTA, ya que R^2 indica la proporción de variación observada de la variable dependiente que puede ser explicada por la variación de la variable independiente y el coeficiente de correlación lineal (r) es el que indica el grado de relación lineal que existe entre la variable independiente y la dependiente y.

(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.*

La afirmación es INCORRECTA, ya que el principio de mínimos cuadrados consiste en minimizar la suma de los residuos al cuadrado considerando la distancia vertical (no perpendicular) entre el valor observado y el estimado.

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 medían 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;* $\hat{\beta}_0$ *= 27,3275;*
$\sum_{i=1}^{n} x_i^2$ *= 424,350;* $\sum_{i=1}^{n} x_i y_i$ *= 183,760;* $\sum_{i=1}^{n} y_i^2$ *= 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.*

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{S_{xx}}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i}{S_{xx}}.$$

$$\hat{\beta}_1 = \frac{183{,}76 - 194 * \sum_{i=1}^{n} y_i}{47{,}99}$$
$$\hat{\beta}_1 = \frac{183{,}76 -}{47{,}99}$$
$$\hat{\beta}_1 = \frac{\quad}{47{,}99}$$
$$\hat{\beta}_1 = .$$

$y = \hat{\beta}_0 + \hat{\beta}_1 x$
$y = 27{,}3275 + \hat{\beta}_1 x.$

Por lo tanto, $\hat{\beta}_0$= 27,3275 indica que el tiempo (en milisegundos) de transmisión del fichero, en promedio, independientemente de la longitud (bytes) de éste, es de 27,3275 milisegundos, mientras que $\hat{\beta}_1$= XXX indica que un aumento de 1 byte en la longitud del fichero aumenta, en promedio, en XXX milisegundos el tiempo de transmisión del fichero.

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

$$R^2 = 1 - \frac{SCE}{SCT}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = 1 - \frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{S_{yy}}$$

$$R^2 = 1 - \frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{S_{yy}}$$

$$R^2 = 1 - \left[1 - \frac{(S_{xy})^2 S_{yy}}{S_{xx}}\right]$$

$$R^2 = 1 - 1 + \frac{(S_{xy})^2 S_{yy}}{S_{xx}}$$

$$R^2 = \frac{(S_{xy})^2 S_{yy}}{S_{xx}}$$

$$R^2 = \frac{(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}{S_{xx}}$$

$$R^2 = \frac{(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i)^2 (\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}{S_{xx}}$$

$$R^2 = \frac{(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i)^2 [\sum_{i=1}^{n} y_i^2 - n(\frac{\sum_{i=1}^{n} y_i}{n})^2]}{S_{xx}}$$

$$R^2 = \frac{(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i)^2 [\sum_{i=1}^{n} y_i^2 - n\frac{(\sum_{i=1}^{n} y_i)^2}{n^2}]}{S_{xx}}$$

$$R^2 = \frac{(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i)^2 [\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}]}{S_{xx}}.$$

$$R^2 = \frac{(183{,}76 - 194 \sum_{i=1}^{n} y_i)^2 [81{,}715 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}]}{47{,}99}$$

$$R^2 = \frac{(183{,}76-)^2 (81{,}715 - \frac{}{n})}{47{,}99}$$

$$R^2 = \frac{()^2 (81{,}715-)}{47{,}99}$$

$$R^2 = \frac{*}{47{,}99}$$

$$R^2 = \frac{}{47{,}99}$$

$$R^2 = .$$

$$r = \sqrt{R^2}$$

$$r = \sqrt{\blacksquare}$$

$$r = .$$

Por lo tanto, el coeficiente de determinación es $R^2 =$ XXX y el coeficiente de correlación lineal es r= XXX.

(c) *Estudiar la significación del modelo.*

$$\hat{\sigma}^2 = \frac{SCE}{n-2}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

$$\hat{\sigma}^2 = \frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{n-2}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n})^2}{S_{xx}}}{n-2}$$

$$\hat{\sigma}^2 = \frac{(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2) - \frac{(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i)^2}{S_{xx}}}{n-2}$$

$$\hat{\sigma}^2 = \frac{[\sum_{i=1}^{n} y_i^2 - n(\frac{\sum_{i=1}^{n} y_i}{n})^2] - \frac{(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i)^2}{S_{xx}}}{n-2}$$

$$\hat{\sigma}^2 = \frac{[\sum_{i=1}^{n} y_i^2 - n\frac{(\sum_{i=1}^{n} y_i)^2}{n^2}] - \frac{(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i)^2}{S_{xx}}}{n-2}$$

$$\hat{\sigma}^2 = \frac{[\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}] - \frac{(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i)^2}{S_{xx}}}{n-2}.$$

$$\hat{\sigma}^2 = \frac{[81{,}715 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}] - \frac{(183{,}76 - 194\sum_{i=1}^{n} y_i)^2}{47{,}99}}{n-2}$$

$$\hat{\sigma}^2 = \frac{(81{,}715 - \frac{\quad}{n}) - \frac{(183{,}76-)^2}{47{,}99}}{n-2}$$

$$\hat{\sigma}^2 = \frac{(81{,}715-) - \frac{()^2}{47{,}99}}{n-2}$$

$$\hat{\sigma}^2 = \frac{-\overline{47{,}99}}{\underline{n}-2}$$

$$\hat{\sigma}^2 = \frac{\quad}{n-2}$$

$$\hat{\sigma}^2 = \frac{\quad}{n-2}$$

$$\hat{\sigma}^2 = .$$

- Test de hipótesis sobre β_0:

Hipótesis:

H_0: $\beta_0 = 0$.
H_1: $\beta_0 \neq 0$.

Estadístico de prueba:

$$T = \frac{\hat{\beta}_0 - (\beta_0)_0}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - (\beta_0)_0}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \sim t_{n-2} \text{ bajo } H_0.$$

Valor observado:

$$t_0 = \frac{27{,}3275 - 0}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{194^2}{47{,}99})}}$$

$$t_0 = \frac{27{,}3275}{\sqrt{\hat{\sigma}^2(+\frac{37636}{47{,}99})}}$$

$$t_0 = \frac{27{,}3275}{\sqrt{\hat{\sigma}^2(+784{,}2467)}}$$

$t_0 = \frac{27{,}3275}{\sqrt{\hat{\sigma}^2{}_*}}$
$t_0 = \frac{27{,}3275}{\sqrt{\quad}}$
$t_0 = \frac{27{,}3275}{\quad}$
$t_0 =$.

Valor crítico:

$t_{\frac{\alpha}{2},n-2} = t_{\frac{0{,}05}{2},n-2}$
$t_{\frac{\alpha}{2},n-2} = t_{0{,}025,n-2}$
$t_{\frac{\alpha}{2},n-2} =$.

Conclusión:

Por lo tanto, con un nivel de significancia de $\alpha = 0{,}05$, estos datos no aportan evidencia suficiente para indicar que la constante del modelo es estadísticamente significativa, ya que $|t_0| = |\quad| =$ XXX $< t_{\frac{\alpha}{2},n-2} =$ XXX.

- Test de hipótesis sobre β_1:

Hipótesis:

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

Estadístico de prueba:

$T = \frac{\hat{\beta}_1 - (\beta_1)_0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - (\beta_1)_0}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$ bajo H_0.

Valor observado:

$t_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{47{,}99}}}$
$t_0 = \frac{\hat{\beta}_1}{\sqrt{\quad}}$
$t_0 = \frac{\hat{\beta}_1}{\quad}$
$t_0 =$.

Valor crítico:

$t_{\frac{\alpha}{2},n-2} = t_{\frac{0{,}05}{2},n-2}$
$t_{\frac{\alpha}{2},n-2} = t_{0{,}025,n-2}$
$t_{\frac{\alpha}{2},n-2} =$.

Conclusión:

Por lo tanto, con un nivel de significancia de α= 0,05, estos datos no aportan evidencia suficiente para indicar que la pendiente del modelo es estadísticamente significativa, ya que $|t_0|$= | |= XXX $<$ $t_{\frac{\alpha}{2},n-2}$= XXX.

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

$IC_{\beta_1}^{95\%}$= [$\hat{\beta}_1$ - $t_{0,025,n-2}$ SE ($\hat{\beta}_1$); $\hat{\beta}_1$ + $t_{0,025,n-2}$ SE ($\hat{\beta}_1$)]
$IC_{\beta_1}^{95\%}$= [$\hat{\beta}_1$ - $t_{0,025,n-2}$ $\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$; $\hat{\beta}_1$ + $t_{0,025,n-2}$ $\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$].

$IC_{\beta_1}^{95\%}$= [$\hat{\beta}_1$ - $t_{0,025,n-2}$ $\sqrt{\frac{\hat{\sigma}^2}{47,99}}$; $\hat{\beta}_1$ + $t_{0,025,n-2}$ $\sqrt{\frac{\hat{\sigma}^2}{47,99}}$]
$IC_{\beta_1}^{95\%}$= [$\hat{\beta}_1$ - $t_{0,025,n-2}$ $\sqrt{\ }$; $\hat{\beta}_1$ + $t_{0,025,n-2}$ $\sqrt{\ }$]
$IC_{\beta_1}^{95\%}$= [$\hat{\beta}_1$ - $t_{0,025,n-2}$ * ; $\hat{\beta}_1$ + $t_{0,025,n-2}$ *]
$IC_{\beta_1}^{95\%}$= [$\hat{\beta}_1$ - ; $\hat{\beta}_1$ +]
$IC_{\beta_1}^{95\%}$= [;].

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

y= 27,3275 + $\hat{\beta}_1$ * 250
y= 27,3275 +
y= .

Por lo tanto, el tiempo de transmisión para un fichero que tiene una longitud de 250 bytes será XXX.

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 (β_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 $\hat{\beta}_0$ 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 $\hat{\beta}_0$.*

$\frac{L}{2} = \frac{-0,4248-(-0,4348)}{2}$
$\frac{L}{2} = \frac{-0,4248+0,4348}{2}$
$\frac{L}{2} = \frac{0,01}{2}$
$\frac{L}{2}$ = 0,005.

Por lo tanto, el error (máximo) que se comete al estimar la verdadera ordenada al origen con $\hat{\beta}_0$ es 0,005.

(b) *La recta de regresión estimada.*

$\hat{\beta}_0 = \frac{-0,4248+(-0,4348)}{2}$
$\hat{\beta}_0 = \frac{-0,4248-0,4348}{2}$
$\hat{\beta}_0 = \frac{-0,8596}{2}$
$\hat{\beta}_0$ = -0,4298.

$\hat{\beta}_1 = 12 \frac{L}{2}$
$\hat{\beta}_1$ = 12 * 0,005
$\hat{\beta}_1$ = 0,06.

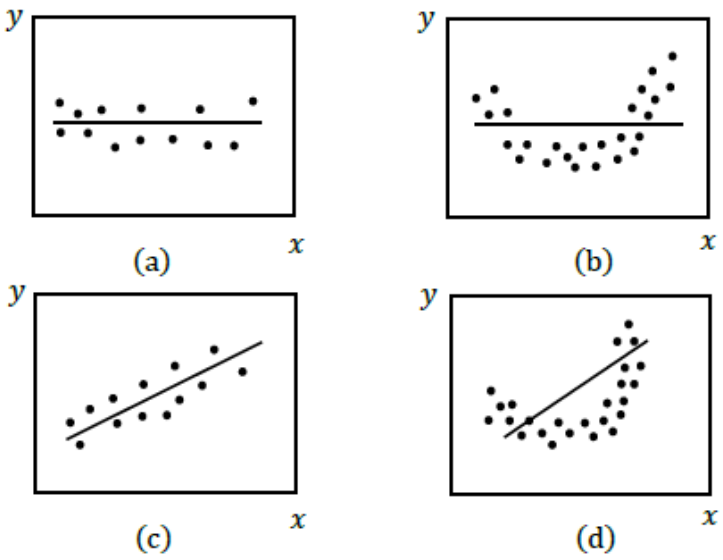
y= $\hat{\beta}_0 + \hat{\beta}_1$x
y= -0,4298 + 0,06x.

(c) *La bondad del ajuste.*

$R^2 = 1 - \frac{SCE}{SCT}$
R^2 = 1 - 0,02
R^2 = 0,98.

Ejercicio 10.

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



En (a) y (b), no se rechaza H_0 (no existe suficiente evidencia para indicar que hay una relación lineal entre x e y), mientras que, en (c) y (d), sí se rechaza H_0 (existe suficiente evidencia para indicar que hay una relación lineal entre x e y). Sin embargo, en (b) y (d), hay una posible relación no lineal.

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?

Hipótesis:

H_0: β_1= 4,0.
H_1: $\beta_1 \neq$ 4,0.

Estadístico de prueba:

T= $\frac{\widehat{\beta}_1-(\beta_1)_0}{SE(\widehat{\beta}_1)} \sim t_{16}$, bajo H_0.

Valor observado:

t_0= $\frac{4,3-4}{0,17}$
t_0= $\frac{0,3}{0,17}$
t_0= 1,765.

Valor crítico:

$t_{\frac{\alpha}{2},n-2}$= $t_{\frac{0,05}{2},16}$
$t_{\frac{\alpha}{2},n-2}$= $t_{0,025,16}$
$t_{\frac{\alpha}{2},n-2}$= 2,12.

Conclusión:

Por lo tanto, con un nivel de significancia de α= 0,05, estos datos no aportan evidencia suficiente para indicar que la pendiente verdadera ha cambiado, ya que $|t_0|$= |1,765|= 1,765 < $t_{\frac{\alpha}{2},n-2}$= 2,12.

Ejercicio 12.

Un horticultor inventó una escala para medir la frescura de rosas que fueron empacadas y almacenadas durante períodos 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?*

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

```
Linear regression                               Number of obs     =         10
                                                F(1, 8)           =     425.18
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9840
                                                Root MSE          =     .76305

------------------------------------------------------------------------------
             |               Robust
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |      -.758   .0367607   -20.62   0.000    -.8427703   -.6732297
       _cons |      20.47   .7174111    28.53   0.000     18.81565    22.12435
------------------------------------------------------------------------------
```

Por lo tanto, hay suficiente evidencia para indicar que la frescura está linealmente relacionada con el tiempo de almacenaje (p-value= 0,000).

(b) *Estimar, mediate un intervalo de 98%, el descenso de frescura de las rosas por cada día que pasa.*

$IC_{\beta_1}^{98\%} = [\hat{\beta}_1 - t_{0,01,n-2}\ \text{SE}(\hat{\beta}_1);\ \hat{\beta}_1 + t_{0,01,n-2}\ \text{SE}(\hat{\beta}_1)]$
$IC_{\beta_1}^{98\%} = [-0{,}758 - 2{,}896 * 0{,}0367;\ -0{,}758 + 2{,}896 * 0{,}0367]$
$IC_{\beta_1}^{98\%} = [-0{,}758 - 0{,}106;\ -0{,}758 + 0{,}106]$
$IC_{\beta_1}^{98\%} = [-0{,}864;\ -0{,}652]$.

(c) *Estimar, mediante un intervalo de 98%, la frescura de las rosas cuando no han sido almacenadas ni empacadas.*

$IC_{\beta_0}^{98\%} = [\hat{\beta}_0 - t_{0,01,n-2}\ \text{SE}(\hat{\beta}_0);\ \hat{\beta}_0 + t_{0,01,n-2}\ \text{SE}(\hat{\beta}_0)]$
$IC_{\beta_0}^{98\%} = [20{,}47 - 2{,}896 * 0{,}717;\ 20{,}47 + 2{,}896 * 0{,}717]$
$IC_{\beta_0}^{98\%} = [20{,}47 - 2{,}078;\ 20{,}47 + 2{,}078]$

$IC_{\beta_0}^{98\%}$= [18,392; 22,548].

(d) *Estimar la medición de frescura media para un tiempo de almacenaje de 14 días con un intervalo de confianza de 95%.*

$IC_{\beta_0+\beta_1 x^*}^{95\%}$= [$\hat{\beta}_0 + \hat{\beta}_1 x^*$ - $t_{0,025,n-2}$ SE ($\hat{\beta}_0 + \hat{\beta}_1 x^*$); $\hat{\beta}_0 + \hat{\beta}_1 x^* + t_{0,025,n-2}$ SE ($\hat{\beta}_0 + \hat{\beta}_1 x^*$)]
$IC_{\beta_0+\beta_1 x^*}^{95\%}$= [9,858 - 2,306 * 0,2437; 9,858 + 2,306 * 0,2437]
$IC_{\beta_0+\beta_1 x^*}^{95\%}$= [9,858 - 0,562; 9,858 + 0,562]
$IC_{\beta_0+\beta_1 x^*}^{95\%}$= [9,296; 10,42].

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) *Desarrollar 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.*

$vida\_litio_i = \beta_0 + \beta_1 horas_i + \varepsilon_i.$
$vida\_alcalina_i = \beta_0 + \beta_1 horas_i + \varepsilon_i.$

(b) *¿Cuál de las dos estimaciones anteriores se ajusta mejor a los datos?*

Estimación con "vida_litio":

```
Linear regression                               Number of obs     =          4
                                                F(1, 2)           =    1575.00
                                                Prob > F          =     0.0006
                                                R-squared         =     0.9973
                                                Root MSE          =      .0866

------------------------------------------------------------------------------
             |               Robust
  vida_litio | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       horas |       -2.1   .0529151   -39.69   0.001    -2.327675   -1.872325
       _cons |        7.3   .1000001    73.00   0.000     6.869734    7.730266
------------------------------------------------------------------------------
```

Estimación con "vida_alcalina":

```
Linear regression                               Number of obs     =          4
                                                F(1, 2)           =     204.13
                                                Prob > F          =     0.0049
                                                R-squared         =     0.9836
                                                Root MSE          =     .05916

------------------------------------------------------------------------------
             |               Robust
vida_alcal~a | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       horas |       -.58   .0405956   -14.29   0.005    -.7546688   -.4053313
       _cons |       2.45   .0754984    32.45   0.001     2.125157    2.774843
------------------------------------------------------------------------------
```

(c) *Encontrar 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?*

Intervalo de confianza para “vida_litio”:

$IC^{90\%}_{\beta_0+\beta_1 horas^*}$= [$\hat{\beta}_0 + \hat{\beta}_1 x^*$ - $t_{0,025,n-2}$ SE ($\hat{\beta}_0 + \hat{\beta}_1 x^*$); $\hat{\beta}_0 + \hat{\beta}_1 x^* + t_{0,025,n-2}$ SE ($\hat{\beta}_0 + \hat{\beta}_1 x^*$)]
$IC^{90\%}_{\beta_0+\beta_1 horas^*}$= [4,675 - 2,92 * 0,0433; 4,675 + 2,92 * 0,0433]
$IC^{90\%}_{\beta_0+\beta_1 horas^*}$= [4,675 - 0,1264; 4,675 + 0,1264]
$IC^{90\%}_{\beta_0+\beta_1 horas^*}$= [4,5486; 4,8014].

Intervalo de confianza para “vida_alcalina”:

$IC^{90\%}_{\beta_0+\beta_1 horas^*}$= [$\hat{\beta}_0 + \hat{\beta}_1 x^*$ - $t_{0,025,n-2}$ SE ($\hat{\beta}_0 + \hat{\beta}_1 x^*$); $\hat{\beta}_0 + \hat{\beta}_1 x^* + t_{0,025,n-2}$ SE ($\hat{\beta}_0 + \hat{\beta}_1 x^*$)]
$IC^{90\%}_{\beta_0+\beta_1 horas^*}$= [1,725 - 2,92 * 0,0296; 1,725 + 2,92 * 0,0296]
$IC^{90\%}_{\beta_0+\beta_1 horas^*}$= [1,725 - 0,0864; 1,725 + 0,0864]
$IC^{90\%}_{\beta_0+\beta_1 horas^*}$= [1,6386; 1,8114].

Por lo tanto, la compañía puede asegurar, con un 90% de confianza, que la batería que la batería que proporciona la vida más larga es la de litio.

(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 el vida aproximada (en meses).*

$horas_i = \beta_0 + \beta_1 vida_i + \varepsilon_i$.

```
Linear regression                               Number of obs     =          4
                                                F(1, 2)           =    1028.00
                                                Prob > F          =     0.0010
                                                R-squared         =     0.9953
                                                Root MSE          =     .05395

------------------------------------------------------------------------------
             |               Robust
       horas | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        vida |  -.7427937   .0231671   -32.06   0.001    -.8424737   -.6431137
       _cons |    3.62694   .0617199    58.76   0.000     3.361381    3.892499
------------------------------------------------------------------------------
```

(e) *Mejorar la estimación utilizando los dos tipos de batería juntas que por separado. Explicar.*

$horas_i = \beta_0 + \beta_1 vida\_litio_i + \beta_2 vida\_alcalina_i + \varepsilon_i$.

```
Linear regression                               Number of obs     =          4
                                                F(2, 1)           =    2789.87
                                                Prob > F          =     0.0134
                                                R-squared         =     0.9997
                                                Root MSE          =      .0208

------------------------------------------------------------------------------
             |               Robust
       horas | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   vida_litio |  -.7698965   .0522349   -14.74   0.043    -1.433603   -.1061896
vida_alcalina |   1.064015    .173853     6.12   0.103    -1.144996    3.273027
        _cons |    3.01384   .0595881    50.58   0.013     2.256701    3.770979
------------------------------------------------------------------------------
```

Ejercicio 14.

Una empresa de desarrollo de software 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, se representa Y (Ventas miles de S/.) y X (N° pedidos de Sistemas), W (N° de pedidos de Aplicaciones Educativas) y Z (N° 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, estimar la ecuación de regresión múltiple para cumplir con el requerimiento de la empresa.*

$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \beta_3 z_i + \varepsilon_i$.

```
Linear regression                               Number of obs     =         10
                                                F(3, 6)           =       2.98
                                                Prob > F          =     0.1179
                                                R-squared         =     0.4498
                                                Root MSE          =     22.558

------------------------------------------------------------------------------
             |               Robust
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |    .681508   .8131731     0.84   0.434    -1.308255    2.671271
           w |  -.4462964   .9988882    -0.45   0.671    -2.890488    1.997895
           z |  -6.252611   3.414964    -1.83   0.117    -14.60873    2.103504
       _cons |   934.8084   324.3938     2.88   0.028     141.0453    1728.571
------------------------------------------------------------------------------
```

(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 se recomendarían? Calcular los mismos y comentar.*

Por un lado, se puede observar que ninguna variable de pedidos de los tipos de software es estadísticamente significativa. Por otro lado, se observa que el test F de significatividad conjunta indica que las variables explicativas no son, en conjunto, estadísticamente significativas (Prob $>$ F= 0,1179). Por último, se ve que el coeficiente de determinación es R^2= 0,4498.

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.*

$$php_i = \beta_0 + \beta_1 algoritmos_i + \beta_2 basededatos_i + \beta_3 programacion_i + \varepsilon_i.$$

```
Linear regression                               Number of obs     =         15
                                                F(3, 11)          =      12.31
                                                Prob > F          =     0.0008
                                                R-squared         =     0.6970
                                                Root MSE          =     .86126

------------------------------------------------------------------------------
             |               Robust
         php | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  algoritmos |   .5826896   .2270632     2.57   0.026     .0829268    1.082452
 basededatos |   .3734826   .2100417     1.78   0.103    -.0888161    .8357813
programacion |  -.2415261    .310188    -0.78   0.453    -.9242452    .4411931
       _cons |   2.551474   2.400433     1.06   0.311    -2.731843    7.834791
------------------------------------------------------------------------------
```

(b) *Si más del 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 sería la respuesta?*

La respuesta sería que, dado que menos del 80% del aprendizaje del curso de PHP puede ser explicado mediante las notas obtenidas por las asignaturas de Algoritmos, Base de Datos y Programación (R^2= 0,697), es necesario destinar más recursos a estas asignaturas para obtener mejores resultados, ya que éstas no están contribuyendo lo suficiente al aprendizaje de PHP.