Ejercicio 1

Una asociación de defensa del consumidor argumenta que el contenido de las latas de atún de una determinada marca es inferior a los 250 g que se indican en el paquete. Para contrastarlo se coge una muestra de nueve de estas latas, obteniéndose:

$$\overline{X} = 242gr. \qquad S_X = 12gr.$$

¿Qué conclusión obtendremos trabajando con un nivel de significación del 5%? ¿Y si trabajamos al 1%?

Por otra parte, el fabricante afirma que el contenido medio de sus latas es igual al del resto de latas del mercado. Para contrastar esta nueva afirmación se coge también una muestra de 25 de las "otras" latas, obteniéndose:

$$\overline{Y} = 245gr. \qquad S_Y = 10gr.$$

¿Qué conclusión extraeremos trabajando con un nivel de significación del 5%?

Resolució:

a) 1) Ho : $\mu = 250$
H1: $\mu < 250$

2) Estadístico de contraste = $\dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}} = \dfrac{242 - 250}{\dfrac{12}{\sqrt{9}}} = -2$

3) Valor crítico correspondiente a t-student con 8 grados de libertad y un nivel de significación del 5% : -1.860

Si el nivel de significación es de 1% el valor crítico es: -2.896

4) Conclusión: el estadístico de contraste cae dentro de la región crítica y, por tanto, rechazamos la hipótesis nula para un nivel de significación del 5 %, i.e.: admitiremos que el contenido de las latas es inferior a 250 gr.

En cambio, si el nivel de significación es del 1%, deberemos aceptar la hipótesis nula

b) Hay que realizar un contraste sobre las diferencias de las medias:

$\overline{x} = 242$ $\qquad$ $\overline{y} = 245$
$s_x = 12$ $\qquad$ $s_y = 10$
$n_x = 9$ $\qquad$ $n_y = 25$

1) Ho: $\mu_x - \mu_y = 0$
H1: $\mu_x - \mu_y \neq 0$

2) Estadístico de contraste = $\dfrac{242 - 245}{\sqrt{\dfrac{8*12^2 + 24*10^2}{9+25-2} \cdot \left(\dfrac{1}{9} + \dfrac{1}{25}\right)}} = -0{,}73$

3) El valor crítico corresponderá a una t-student con un nivel de significación del 5% y 32 grados de libertad es –2,04.

4) El estadístico de contraste cae fuera de la región de *no rechazo* y, por lo tanto, no rechazaremos la hipótesis nula de igualdades de medias poblacionales.

Ejercicio 2

La siguiente tabla recoge (figuradamente) las horas asignadas por la TV3 durante la campaña electoral a las distintas opciones políticas que se presentaron a las últimas elecciones al Parlament de Catalunya, y el número de diputados que han obtenido estas formaciones políticas.

Horas dedicadas	8	7	5	4	2
Diputados	CiU = 48	PSC = 37	ERC = 21	PP = 14	IC-V = 12

Encontrad la recta de regresión por mínimos cuadrados del número de diputados sobre las horas de presencia en TV.
Encontrad el coeficiente de correlación a partir de los datos anteriores.
Encontrad el coeficiente de determinación R cuadrado e interpretad los resultados.
Si dos formaciones políticas tuvieran 6 y 3 horas de permanencia *ceteris paribus*, ¿cuántos diputados podrían esperar?

RESOLUCIÓN

a) Para encontrar la recta de regresión de Y sobre X, tenemos que encontrar los valores de la ecuación Y = aX + b. Para encontrarlo, primero realizamos los cálculos necesarios y tenemos:

a) **Encontrad la recta de regresión por mínimos cuadrados del número de diputados sobre las horas de presencia en TV.**

$\bar{x} = 5{,}20$
$\bar{y} = 15$
$s_x = 2{,}388$
$s_y = 11{,}747$

X	**x_i-x**	**Y**	**Y_i-y**	**$(x_i-x)(y_i-y)$**	**$(y_1-y)^2$**	**$(x_i-x)^2$**
8	2’8	33	18	50,4	324	7,84
7	1’8	19	4	7,2	16	3,24
5	-0’2	13	-2	0,4	4	0,04
4	-1’2	7	-8	9,6	64	1,44
2	-3,2	3	-12	38,4	144	10,24
				106	**552**	**22,8**

$$m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{106}{22.8} = 4.65$$

$$b = \bar{y} - m\bar{x} = 15 - 4.65 \times 5.2 = -9.18$$

$$y = 4.65x - 9.18$$

Un poco más ajustada sería: y = 4,649x - 9,174.

2. Encontrad el coeficiente de correlación a partir de los datos anteriores.
Utilizando las fórmulas del manual, pág. 208 y sig., tenemos:

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$\sigma_y = \sqrt{\frac{\sum (y_i - y)^2}{n-1}}$$

Si en estas fórmulas efectuamos las sustituciones pertinentes, tendremos:

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \text{cov}(x, y) = \frac{1}{4}.112 = 28$$

$$\sigma_y = \sqrt{\frac{\sum (y_i - y)^2}{n-1}} = \sigma_y = \sqrt{\frac{522}{4}} = 14'42$$

$$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sigma_x = \sqrt{\frac{26}{4}} = 2'55$$

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = r = \frac{28}{2'55.14'42} = 0'76$$

Pero se puede utilizar eso: (con los datos ya corregidos)
Utilizaremos la fórmula para el coeficiente de correlación:
r = cov (x, y) / Sx Sy
r = (106/4) /11,747*2,388
r = 26,5 / 28,051

r = 0,945

3. El coeficiente de determinación
Dado que se trata de una regresión simple, el R^2 es igual al r^2, es decir, $R^2 = r^2$.

Entonces, tendremos:

$R^2 = 0'945^2$; $R^2 = 0'893$

Interpretación:
La R^2 nos informa sobre la bondad del ajuste. Nos dice qué parte de la variable dependiente es explicada por la variable explicativa.
En nuestro caso concreto, nos dice que las horas de permanencia de las distintas opciones políticas en TV explica el 89,3% de su voto. Si R^2 fuera inferior a 0,5 diríamos que no tiene influencia o que no es significativo; en este caso no lo podemos decir, pero vemos que estamos bordeando el límite de la no significación y que el ajuste es bastante malo. Eso era de esperar teniendo en cuenta que tenemos una muestra muy pequeña y que, mientras que los valores de la variable dependiente son reales, los valores de la variable independiente son totalmente arbitrarios.

4. Si dos formaciones políticas tuvieran 6 y 3 horas de permanencia *ceteris paribus*, ¿cuántos diputados podrían esperar?

En este caso, se trataría de sustituir en la recta de regresión a la "x" por los valores que nos da el enunciado.

$y = 4.65 \times 6 - 9.18 = 18.72$. Para 6 h de permanencia podríamos esperar entre 18 y 19 diputados.
$y = 4.65 \times 3 - 9.18 = 4.77$. Para 3 h de permanencia podríamos esperar entre 4 y 5 diputados.

También podemos utilizar este proceso:

Utilizaremos de nuevo la fórmula de la recta de regresión $\hat{y} = m\,\bar{x} + b$.
$\hat{y}=m\,\bar{x}+b$
$\hat{y} = 4{,}649\,\bar{x} - 9{,}174$
$\hat{y} = 4{,}649*6 - 9{,}174$
$\hat{y} = 27{,}894 - 9{,}174$
$\hat{y}$ = 18,72

Según la fórmula, con 6 horas, podrían esperar 18 diputados.

$\hat{y}=m\,\bar{x}+b$
$\hat{y} = 4{,}649\,\bar{x} - 9{,}174$
$\hat{y} = 4{,}649*3 - 9{,}174$
$\hat{y} = 13{,}947 - 9{,}174$

$\hat{y}$ = 4,773

Según los resultados obtenidos, con 3 horas, podrían esperar a 4 diputados.

Vemos, pues, cómo las predicciones no son nada fiables, especialmente la primera, ya que para 6 h de permanencia nos da un resultado superior al observado para 7 h.

Según los resultados obtenidos, con 3 horas, podrían esperar a 4 diputados.

Vemos, pues, cómo las predicciones no son nada fiables, especialmente la primera, ya que para 6 h de permanencia nos da un resultado superior al observado para 7 h.

Ejercicio 3

El número de libros prestados a estudiantes y profesores de cinco departamentos de una universidad es el siguiente:

Departamento	Estudiantes	Profesores
Antropología	1.122	340
Biología	311	273
Física	1.446	704
Informática	557	233
Psicología	1.153	495

a) Encontrad el coeficiente de correlación de lineal de Pearson y el R^2 e interpretadlos.
b) Calculad la recta de regresión entre ambas variables
c) Si un departamento universitario presta 1.000 libros a los estudiantes, ¿cuántos se puede esperar que preste a los profesores?

RESOLUCIÓ
RESOLUCIÓN

a) La tabla con las operaciones para encontrar el coeficiente r de Pearson es la siguiente:

Departamento	Estudiantes	Profesores	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
Antropología	1.122	340	204,20	-69,00	-14.089,80	41.697,64	4.761,00
Biología	311	273	-606,80	-136,00	82.524,80	368.206,24	18.496,00
Física	1.446	704	528,20	295,00	155.819,00	278.995,24	87.025,00
Informática	557	233	-360,80	-176,00	63.500,80	130.176,64	30.976,00
Psicología	1.153	495	235,20	86,00	20.227,20	55.319,04	7.396,00

Media	917,8	409,0					
				Sumatorio	307.982,00	874.394,80	148.654,00
					r =	0,85	
					m=	0,35	
					b =	85,73	

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

r = 0,85

Una r de 0,85 muestra una fuerte relación positiva entre ambas variables.

b) Las operaciones para encontrar los parámetros de la recta de regresión son los siguientes:

$$\hat{m} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad \hat{b} = \overline{Y} - \hat{m}\overline{X}$$

Y = mx + b
m = 0'35
b=87'73

y= 0'35x+87'73

c) Si los préstamos a los estudiantes (x) son de 1.000, entonces:

Y= 0'35*1000+87'73 = 437'95

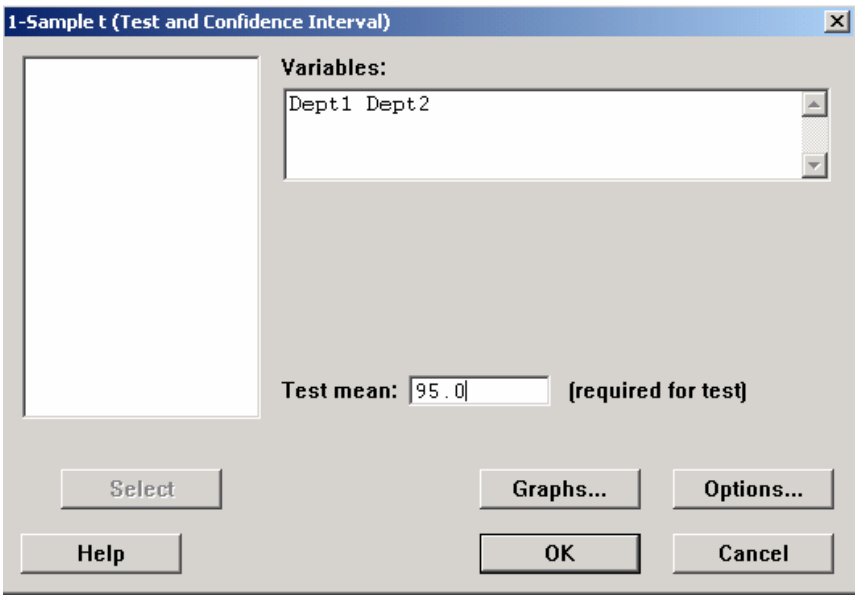
1. En una determinada empresa de producción editorial se ha implantado un nuevo sistema informático destinado a agilizar la revisión y corrección del material editorial. Pedimos a dos departamentos de la empresa que evalúen este nuevo sistema. Las calificaciones obtenidas son las que aparecen en la tabla siguiente:

Dept. 1	Dept. 2
5	6.25
7.5	5.75
6	5
2.5	4.75
8	8
9	9
7	7.5
6	8
4	9
3.75	10
9	
10	
8.25	
9	
6	

a) Calcula un intervalo de confianza para cada una de las dos poblaciones al nivel de confianza del 95%. Comentar los resultados.

Como, en este caso, el valor de la varianza de la población es desconocida, utilizaremos la opción correspondiente a la distribución t-Student.

Para calcular el intervalo de confianza seleccionaremos: Stat> Basic Statistic> 1-Sample t , obteniendo los siguientes resultados:



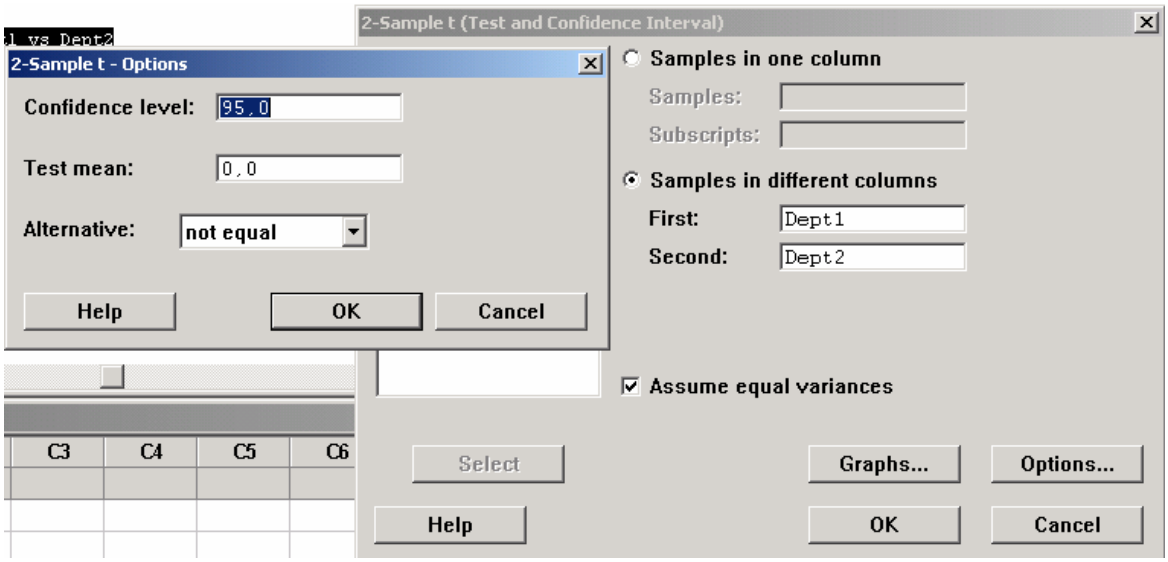
One-Sample T: Dept1; Dept2

```
Test of mu = 95 vs mu not = 95

Variable          N      Mean     StDev   SE Mean
Dept1            15     6,733     2,229     0,576
Dept2            10     7,325     1,807     0,571

Variable             95,0% CI            T      P
Dept1         (    5,499;   7,968)  -153,37  0,000
Dept2         (    6,032;   8,618)  -153,45  0,000
```

Si nos fijamos en los dos intervalos de confianza, éstos se solapan. Esto implica que si estamos interesados en comparar las medias de ambas poblaciones, estas medias pertenecen a intervalos con datos en común, lo cual hace pensar que estas medias poblacionales, es decir, las medias del dept1 y del dept2 pueden ser iguales.

b) Calcular un intervalo de confianza para la diferencia de medias. Utilizando este intervalo contrastar la hipótesis de que las medias en los dos grupos no difieren.

Seleccionamos Stat > Basic Statistics > 2-Sample t:



Two-Sample T-Test and CI: Dept1; Dept2

```
Two-sample T for Dept1 vs Dept2

        N      Mean     StDev   SE Mean
Dept1  15      6,73      2,23      0,58
Dept2  10      7,33      1,81      0,57

Difference = mu Dept1 - mu Dept2
Estimate for difference:  -0,592
95% CI for difference: (-2,343; 1,160)
T-Test of difference = 0 (vs not =): T-Value = -0,70  P-Value = 0,492  DF = 23
Both use Pooled StDev = 2,07
```

c) ¿Qué error de equivocarnos, si concluimos que hay diferencias entre las poblaciones, deberíamos estar dispuestos a asumir?

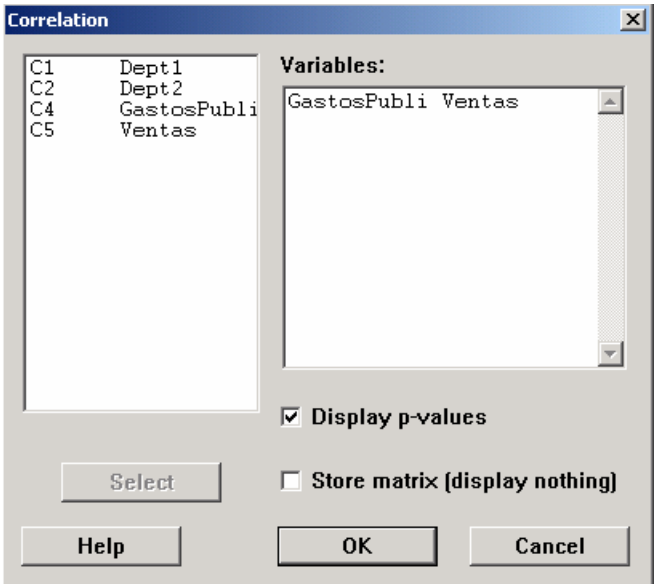
Si observamos por ejemplo el caso en el cual consideramos las varianzas iguales en las dos poblaciones, el error de equivocarnos al rechazar la hipótesis de igualdad de medias es de 0,49. Este error es muy alto, por lo que debemos concluir que no podemos rechazar la hipótesis nula de igualdad de medias.

2. En una determinada empresa de venta de libros on-line, se quiere estudiar la relación entre las ventas realizadas y la cantitad gastada en publicidad. Los datos, en millones de €, de los 4 últimos meses son los siguientes:

Gastos en publicidad	2	1	3	4
Ingreso por ventas	7	3	8	10

a) Determina el coeficiente de correlación entre las dos variables. Calcula y representa también la recta de regresión.

Para calcular el coeficiente de correlación, seleccionamos *Stat > Basic Statistics > Correlation:*

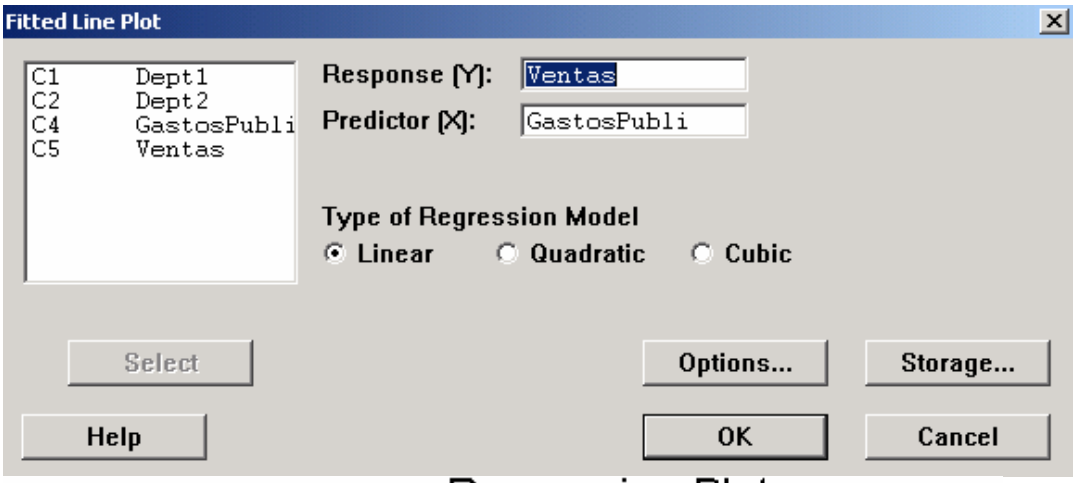


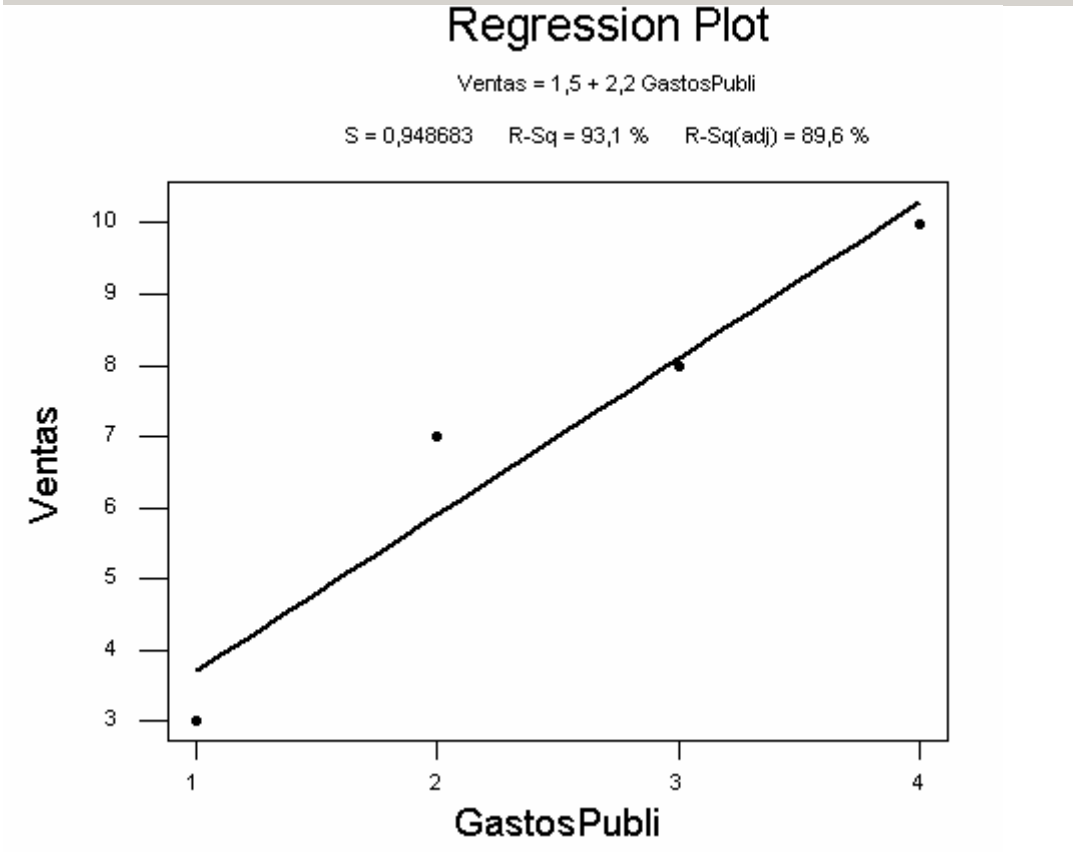
Correlations: GastosPubli; Ventas

```
Pearson correlation of GastosPubli and Ventas = 0,965
P-Value = 0,035
```

El coeficiente de correlación r = 0.965 nos indica que hay una fuerte correlación entre los gastos invertidos en publicidad y la cantidad de ventas conseguida.

Para representar la recta de regresión, utilizamos la opción *Stat > Regresión > Fitted Line Plot* :





Regression Analysis: Ventas versus GastosPubli

The regression equation is
Ventas = 1,5 + 2,2 GastosPubli

S = 0,948683 R-Sq = 93,1 % R-Sq(adj) = 89,6 %

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	24,2	24,2	26,8889	0,035
Error	2	1,8	0,9		
Total	3	26,0			

b) Estimar la cantidad de ventas obtenidas si se destinan 3 millones de € a publicidad.

La recta de regresión es Ventas=1.5 + 2.2*GastosPubli

Por tanto, si se invierten 3 millones de € en publicidad, se obtiene:

Ventas = 1.5 + 2.2*3 = 8.1

Es decir, se estima que se obtendran 8.1 millones de €

c) ¿Crees que la muestra anterior presenta suficiente evidencia, a un nivel de significación de 0,05, como para rechazar la hipótesis nula sobre la pendiente (H_0: pendiente de la recta es cero)?

En el output anterior podemos ver que el p-valor asociado al contraste de hipótesis anterior es 0,035. Como este valor es menor que α = 0,05, debemos rechazar la hipótesis nula, es decir, podemos concluir que la pendiente de la recta es distinta de cero o, lo que es lo mismo (ver GES 8), que el coeficiente de correlación poblacional es no nulo (es decir, que ambas variables están correlacionadas y que, por tanto, el modelo estudiado tiene sentido).