387 TEXTOS DOCENTS

REGRESIÓN LINEAL SIMPLE PARA EL DEPORTE Y EL EJERCICIO

Martín Ríos (coord.)
Daniel Ríos
Marta Cubedo

Departamento de Estadística

Universitat de Barcelona

REGRESIÓN LINEAL SIMPLE PARA EL DEPORTE Y EL EJERCICIO

Martín Ríos (coord.)
Daniel Ríos
Marta Cubedo

Departamento de Estadística

387 TEXTOS DOCENTS

Universitat de Barcelona
Publicacions i Edicions

Índice

Prólogo

Muchos científicos y estudiantes dedicados al estudio del deporte consideran la Estadística como una disciplina ajena y una asignatura de las menos importantes de su formación. Una de las razones para esta falta de motivación es el hecho de que los libros de texto de Estadística y Bioestadística generalmente no incluyen ejemplos relevantes relacionados con el deporte y el ejercicio físico.

En los últimos años se ha producido un importante incremento de la investigación en el campo de la actividad física y el deporte, lo que se manifiesta en el aumento de consultas sobre estadística planteadas a los autores.

Estas razones nos han animado a elaborar esta colección de cuadernos de problemas de Estadística, basados en la experiencia de los últimos años. En ellos hemos utilizado datos reales, lo que hace que el tema sea más motivador y comprensible para el lector.

Para seguir los temas tratados solo se necesita un conocimiento muy básico de Estadística, aunque diseñamos y analizamos los problemas de una manera formal y rigurosa.

En cada uno de los problemas que resolvemos, los cálculos se realizan «a mano» y con el software libre R.

Finalmente, queremos agradecer a los profesores José María Oller y Antonio Miñarro sus sugerencias en algunas cuestiones.

REGRESIÓN Y CORRELACIÓN

En muchos casos, dadas dos variables aleatorias X e Y, resulta interesante relacionar linealmente la Y y la variable X. Para ello debemos encontrar la combinación lineal de X, $\alpha+\beta X$, de modo que se ajuste de la mejor forma posible a Y.

El criterio para obtener la combinación lineal es el de los mínimos cuadrados:

$$F(\alpha,\beta)=E(Y-\alpha-\beta X)^2=\min$$

Obteniéndose:

$$\beta=\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)} \quad \alpha=E(Y)-\beta E(X)$$

El grado de relación lineal entre X e Y viene cuantificado por el coeficiente de correlación ρ cuyo valor viene dado por la expresión:

$$\rho^2=\frac{\operatorname{cov}^2(X,Y)}{\operatorname{var}(X)\operatorname{var}(Y)}$$

que se considera el porcentaje de variabilidad de la Y que depende de la X.

Las propiedades de este parámetro son:

1. $-1\leq\rho\leq 1$
2. Si $\rho^2=1$, se cumple que $Y=\alpha+\beta X$.
3. Si $\rho=0$ se dice que las variables están incorrelacionadas. En particular, si son independientes, entonces $\rho=0$. El recíproco en general no es cierto.
4. El coeficiente de correlación es invariante por transformaciones lineales de las variables, es decir:

$$\rho=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}}$$

Si hacemos un cambio de escala: $\alpha X+\beta$; $\alpha' Y+\beta'$

$$\rho'=\frac{\operatorname{cov}(\alpha X+\beta,\alpha' Y+\beta')}{\sqrt{\operatorname{var}(\alpha X+\beta)}\sqrt{\operatorname{var}(\alpha' Y+\beta')}}$$

$$\rho=\rho'$$

En la práctica para analizar la relación lineal entre la X y la Y partimos de una muestra de n pares de observaciones de la variable (X,Y).

Observaciones	X	Y
1	x_1	y_1
2	x_2	y_2
	$\vdots$	$\vdots$
n	x_n	y_n

Tanto los parámetros de la regresión, coeficientes α, β, como el coeficiente de correlación ρ, deben ser estimados a partir de la muestra. Sus estimaciones son a, b, r respectivamente.

$$b = \frac{S_{xy}}{S_x^2};\ a = \bar{y} - b\,\bar{x};\ r = \frac{S_{xy}}{S_x S_y}$$

siendo:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i; \quad S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad S_y^2 = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \bar{y}\big)^2; \quad S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\big(y_i - \bar{y}\big)$$

En general, interesa conocer si las variables analizadas están incorrelacionadas. Para ello nos planteamos realizar el siguiente contraste de hipótesis:

$$H_0 : \rho = 0 \text{ frente a } H_1 : \rho \neq 0$$

Si H_0 es cierta, el estadístico

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

Sigue una t de *Student* con $n-2$ grados de libertad, donde r es la estimación de ρ y n es el tamaño muestral.

Rechazamos H_0 a nivel de significación ε si $|t| > k$. Siendo k un valor real que cumple

$$p\,(T > k) = \frac{\varepsilon}{2},$$

T es una variable aleatoria que se distribuye según una t de *Student* con $n-2$ grados de libertad.

PROBLEMAS

1. La presión parcial de oxígeno de la sangre de los escaladores disminuye cuando ascienden una montaña. Los datos de la siguiente tabla nos muestran las presiones parciales de oxígeno arteriales según la altura alcanzada. H: altura en m y p_{O_2}: presión parcial en mmHg.

H	Nivel del mar	1700	2000	2238	3070	3450	3846	4240	5610
p_{O_2}	100	69,5	65,3	62	51,5	47,1	42,8	38,6	25,7

a) ¿Cuál es la variable dependiente?

b) Representar los datos de la tabla en un sistema de coordenadas cartesianas. Analiza las características del gráfico que consideres más interesantes.

c) Calcular el coeficiente de correlación muestral. Analizar, a un nivel de significación del 0,1%, si hay relación lineal entre la altura y la p_{O_2}. Explicar los valores obtenidos.

d) Estimar la recta de regresión que relaciona la altura con el p_{O_2}.

e) Dar una estimación de la p_{O_2} en los siguientes lugares: Bogotá (2,640 km) y Everest (8,848 km). Comentar los resultados.

f) Dar un intervalo de la predicción realizada para Bogotá (2,640 km) al 95%.

g) Entre qué valores se encuentra la p_{O_2} de la población de habitantes de Bogotá con probabilidad 0,95.

Solución

a) La p_{O_2}.

Representamos en el eje de abscisas la altura (H) y en el eje de ordenadas p_{O_2}. A medida que la altura aumenta la p_{O_2} disminuye. Véase la figura 1.1.

b)

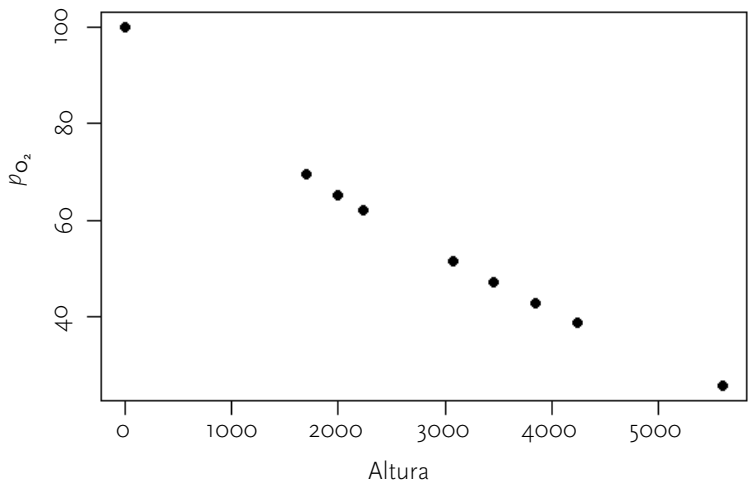


Figura 1.1. Representación gráfica de la altura (H) en metros y p_{O_2} en mmHg.

c)
$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)\right)}} = -0{,}988$$

$$\sum x = 26154; \quad \sum x^2 = 97467460; \quad \sum y = 502{,}5; \quad \sum y^2 = 31791{,}29; \quad \sum xy = 1180555{,}8$$

$$n = 9$$

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} = -16{,}839$$

Consultando la tabla de la *T-Student* con 7 g.l., se tiene

$$p(|t| > 12{,}924) = 0{,}001$$

Puesto que $|-16{,}839| \geq 12{,}924$ podemos aceptar que a un nivel de significación del 0,1%, hay relación lineal entre la altura y la p_{O_2}.

El coeficiente de regresión es negativo, lo que confirma que a medida que la altura aumenta la p_{O_2} disminuye. El valor absoluto de t es muy superior al obtenido en las tablas, por lo que podemos considerar que la relación lineal es fuerte.

d)
$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)} = -0{,}013;$$

$$a = \bar{y} - b\,\bar{x} = 93{,}703$$

$$y = 93{,}703 - 0{,}013x$$

e) Bogotá: $93{,}703 - 0{,}013 * 2640 = 59{,}3$ mmHg
Everest: $93{,}703 - 0{,}013 * 8848 = -21{,}6$ mmHg

El resultado para Everest no tiene sentido, no se debe considerar.

f) El intervalo de confianza para el verdadero valor de la estimación de Bogotá es:

$$59{,}3 \pm t(0{,}05)\sqrt{\frac{n}{n-2}}\,\tilde{s}_y\sqrt{1 + \frac{1}{n} + \frac{(2640-\bar{x})^2}{n s_x^2}}$$

Como $t(0{,}05) = 2{,}306$, el intervalo de confianza es:

$$I = (59{,}3 \pm 8{,}95) = (50{,}35 : 68{,}25)$$

g) La variable aleatoria $Y/X = 2640$, al ser combinación lineal de la X, se distribuye según una normal aproximadamente de media $\bar{x} = 2640$ y desviación típica, $\tilde{s}_y = s_y\sqrt{1-r^2} = 1{,}464$, por lo que aproximadamente el 95% de los valores de la p_{O_2} de la población de habitantes de Bogotá estarán comprendidos entre $(59{,}3 \pm 1{,}96 * 1{,}464) = (56{,}432 : 62{,}167)$.

Cálculos en R

```
> x <-c(0, 1700, 2000, 2238, 3070, 3450, 3846, 4240, 5610)
> y<-c(100, 69.5, 65.3, 62, 51.5, 47.1, 42.8, 38.6, 25.7)
> plot(x,y, xlab="Altura",ylab="pO2",pch=19)
> modelo <-lm(y~x)
> cor(x,y)
-0.9878806
> cor.test(x,y)
        Pearson's product-moment correlation
data:  x and y
t = -16.839, df = 7, p-value = 6.377e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.997542 -0.941359
> qt(0.9995, 3, lower.tail=TRUE)
[1] 12.92398
sample estimates:
       cor
-0.9878806
> summary(modelo)
Call:
lm(formula = y ~ x)
Residuals:
    Min      1Q  Median      3Q     Max
-2.5384 -2.1962 -1.6441  0.1508  6.2969
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 93.7031320  2.5467650   36.79 2.85e-09 ***
x           -0.0130316  0.0007739  -16.84 6.38e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.585 on 7 degrees of freedom
Multiple R-squared: 0.9759,    Adjusted R-squared: 0.9725
F-statistic: 283.6 on 1 and 7 DF,  p-value: 6.377e-07
> new.data<-as.data.frame(c(2640,8848))
> new <- data.frame(x = c(2640,8848))
> predict(modelo,new, interval='prediction')
       fit      lwr      upr
1  59.29974  50.34982 68.249656
2 -21.60037 -35.67518 -7.525561
```

2. Se supone que la relación entre dos variables x e Y es lineal, donde Y toma un valor constante e igual a cero cuando $x = 0$. Es decir $E(Y/x) = \beta x$. Determinar β por el método de mínimos cuadrados.

Para analizar el efecto que sobre la temperatura corporal tiene una determinada manta térmica usada en los rescates de montaña, se midió el incremento de la temperatura de la zona central del cuerpo y la de la piel en 10 personas, a $-10\,^{\circ}\mathrm{C}$ y con una velocidad del viento de 2,7 ms^{-1}, durante 25 minutos cada 5 minutos. Los promedios de los resultados vienen dados en la siguiente tabla:

	5 min	**10 min**	**15 min**	**20 min**	**25 min**	**30 min**
Zona central	0,07	−0,07	−0,13	−0,17	−0,27	−0,38
Piel	−2,81	−3,53	−4,08	−4,53	−4,92	−5,17

a) Representar los datos de las temperaturas en un diagrama.

b) Proponer un modelo para los datos, explicando las razones de su elección.

c) Estimar los parámetros del modelo escogido. Analizar la bondad del modelo con un nivel de significación del 5%

d) Calcular los incrementos de temperatura predichos para los seis tiempos utilizados en el estudio.

Solución

Sean las n observaciones:

X	x_1	x_2	·	·	·	x_n
Y	y_1	y_2	·	·	·	y_n

El valor de β será el que minimice la expresión:

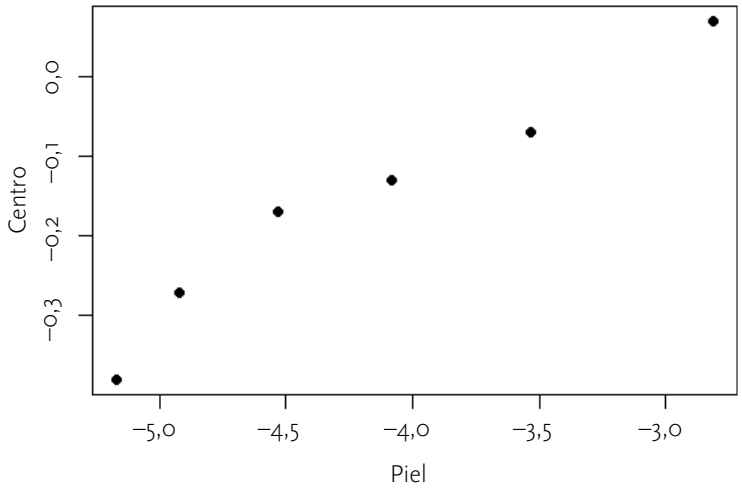
$$\emptyset(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

por lo tanto se ha de cumplir que:

$$\frac{\partial \emptyset}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - \beta x_i) = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

a)



b) La variación de la temperatura en la zona central del cuerpo es proporcional a la temperatura de la piel y además consideramos que a incremento cero en piel, en el centro el incremento también es nulo:

$$\Delta \textit{ temperatura en Centro} = \beta \, \Delta \textit{ temperatura en Piel}$$

c)

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = 0{,}043;$$

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n \sum x_i^2 - \left(\sum x_i\right)^2 \left(n \sum y_i^2 - \left(\sum y_i\right)^2\right)\right)}} = 0{,}975$$

$$t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} = 8{,}844$$

Consultando la tabla de la *T-Student* con 4 g.l., se tiene

$$p(|t| > 2{,}776) = 0{,}05$$

Puesto que 8,844 > 2,776 podemos aceptar que a un nivel de significación del 0,05, hay relación lineal entre el incremento de temperatura en la piel y en el centro del cuerpo.

d) $$\Delta t\,(centro) = 0{,}043 \Delta t\,(piel)$$

Zona central observada	0,07	−0,07	−0,13	−0,17	−0,27	−0,38
Zona central predicha	−0,12	−0,15	−0,17	−0,19	−0,21	−0,22
Piel	−2,81	−3,53	−4,08	−4,53	−4,92	−5,17

Cálculos en R

```
> x<-c(-2.81, -3.53, -4.08,-4.53,-4.92, -5.17)
> y <-c(0.07, -0.07, -0.13, -0.17, -0.27, -0.38)
> plot(x,y, xlab="Piel",ylab="Centro",pch=19)
> modelo <-lm(y~x-1)
> summary(modelo)
Call:
lm(formula = y ~ x - 1)
Residuals:
      1       2       3       4       5       6
 0.19032  0.08114  0.04469  0.02396 -0.05934 -0.15864
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x  0.04282    0.01169   3.664   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1217 on 5 degrees of freedom
Multiple R-squared: 0.7286,    Adjusted R-squared: 0.6743
F-statistic: 13.42 on 1 and 5 DF,  p-value: 0.01454
> cor(x,y)
[1] 0.9753701
> cor.test(x,y)
        Pearson's product-moment correlation
data:  x and y
t = 8.8439, df = 4, p-value = 0.0009025
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7859279 0.9974094
sample estimates:
      cor
0.9753701
> qt(0.975, 4, lower.tail = TRUE)
[1] 2.776445
> predict(modelo)
         1          2          3          4          5          6
-0.1203154 -0.1511435 -0.1746928 -0.1939604 -0.2106590 -0.2213632
```

3. Para estudiar la posible relación entre las horas de clase de Educación Física de los alumnos de bachillerato y el rendimiento en las pruebas de salto de longitud, se seleccionaron diez centros de enseñanza secundaria de la ciudad de Barcelona. Para cada centro se consideró el número x de horas de clase de Educación Física que se impartían a la semana y el promedio Y de las marcas en metros, obtenidas por sus alumnos varones no repetidores que cursaban primero de bachillerato. Aquí x es una medida controlada (no tiene errores) pero Y es una variable aleatoria, observada por el valor y. Los resultados fueron los siguientes:

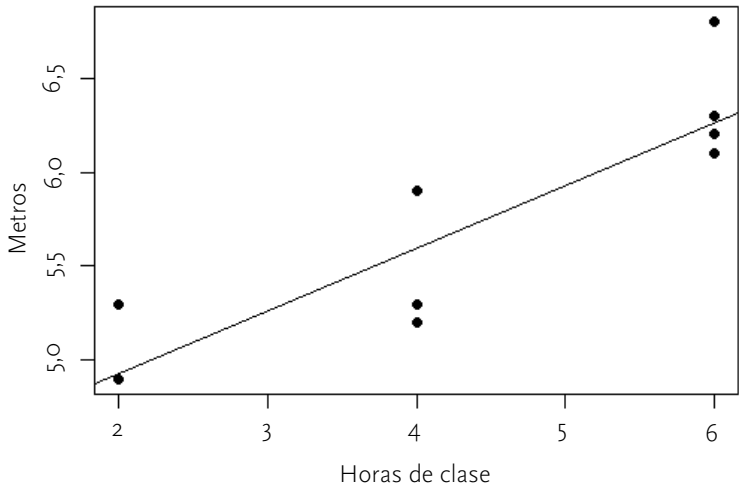
x	4	6	2	6	4	6	2	6	4	4
y	5,2	6,3	5,9	6,1	5,3	6,2	5,3	6,8	5,9	5,3

Se propone como modelo que relaciona x, Y: $E(Y \mid x_i) = \alpha_0 + \beta(x_i - \bar{x})$.

Dibujar un diagrama y analizar si el modelo es aceptable con un nivel de significación del 5%.

Estimar por mínimos cuadrados α_0 y β e interpretar ambos parámetros.

Solución



$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)\right)}} = 0{,}860$$

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} = 4{,}757$$

Consultando la tabla de la *T-Student* con 8 g.l., se tiene

$$p(|t| > 2{,}306) = 0{,}05$$

Puesto que 4,757 > 2,306 podemos aceptar que a un nivel de significación del 0,05, hay relación lineal entre las horas de clase de Educación Física y el rendimiento en el salto de longitud.

La recta de regresión Metros $= \alpha + \beta *$ Horas, viene estimada por: Metros $= a + b *$ Horas, donde:

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)} = 0{,}334;\ a = \bar{y} - b\,\bar{x} = 4{,}261$$

Es decir: Metros $= 4{,}261 + 0{,}334 *$ *Horas*.

Por lo tanto la estimación del parámetro β del modelo propuesto es: $b = 0{,}334$ y representa el rendimiento en metros de un alumno por hora de clase.

La estimación de α_0 es $4{,}261 + 0{,}334\,\bar{x} = 5{,}596$, es decir el valor predicho de los metros para el promedio de horas de clase de Educación Física en todos los centros ya que:

$$E(Y \mid \bar{x}) = \alpha_0 + \beta(\bar{x} - \bar{x}) = \alpha_0$$

Cálculos en R

```
> x<-c(4,6,2,6,4,6,2,6,4,4)
> y<-c(5.2,6.3,4.9,6.1,5.3,6.2,5.3,6.8,5.9,5.3)
> plot(x, y, xlab="horas de clase", ylab="metros",pch=19)
> abline(lm(y~x))
> cor(x,y)
[1] 0.859518
> cor.test(x,y)
        Pearson's product-moment correlation
data:  x and y
t = 4.7565, df = 8, p-value = 0.001433
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5010447 0.9662395
sample estimates:
     cor
0.859518
> qt(0.975, 8, lower.tail = TRUE)
[1] 2.306004
> modelo <-lm(y~x)
>  summary(modelo)
Call:
lm(formula = y ~ x)
Residuals:
     Min       1Q   Median       3Q      Max
-0.39643 -0.26339 -0.04643  0.23661  0.53571
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.2607     0.3263  13.058 1.12e-06 ***
x             0.3339     0.0702   4.757  0.00143 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3323 on 8 degrees of freedom
Multiple R-squared: 0.7388,    Adjusted R-squared: 0.7061
F-statistic: 22.62 on 1 and 8 DF,  p-value: 0.001433
> mean(x)
[1] 4.4
> new.data<-as.data.frame(c(2640,8848))
> new <- data.frame(x = c(2640,8848))
> predict(modelo,new, interval='prediction')
       fit      lwr      upr
1 5.596429 4.790215 6.402642
```

4. Un complejo de rehabilitación de Villena (Alicante) está interesado en el aprovechamiento de la energía solar para calentar el agua de su piscina. Para ello encarga un estudio que relacione las horas de sol mensual con la energía producida por la radiación solar medida en $\text{kW}\cdot\text{h/m}^2/\text{día}$.

Los resultados observados fueron los siguientes:

Horas	196	164	252	183	330	323	332	289	250	188	190	190
$\text{kW}\cdot\text{h/m}^2/\text{día}$	2,36	3,33	4,53	5,25	6,42	6,89	7,17	6,25	5,08	3,86	2,72	2,11

a) Representar los datos en un sistema de coordenadas.

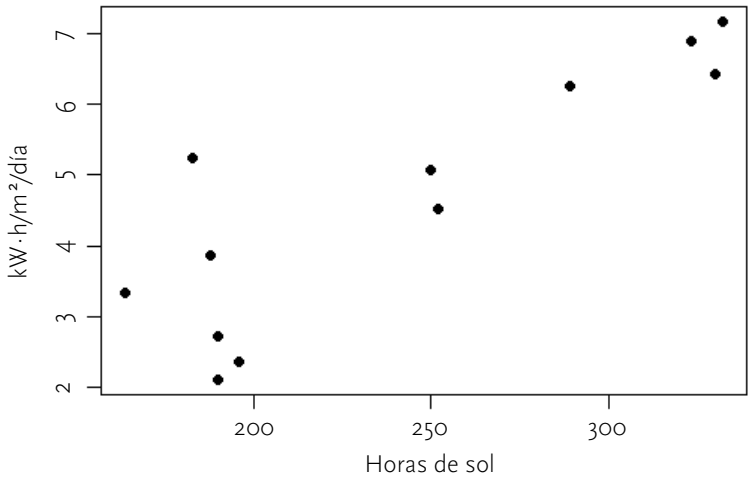
b) Calcular la recta de regresión estimada que relaciona ambas variables y dibujarla en el diagrama anterior.

c) Analizar la adecuación del modelo con un nivel de significación de *0,05*.

d) Estimar la radiación solar de una localidad próxima, en un mes que tuviera 300 horas de sol.

Solución

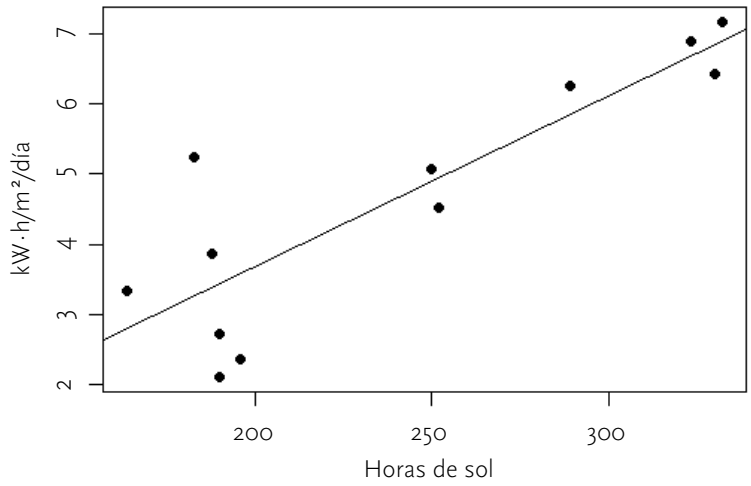
a)



b)

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)} = 0{,}024; \qquad a = \bar{y} - b\,\bar{x} = -1{,}193$$

$$Radiación = -1{,}193 + 0{,}024 * Horas$$



c)

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)\right)}} = 0{,}866$$

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} = 5{,}467$$

Consultando la tabla de la *T-Student* con 4 g.l., se tiene

$$p(|t| > 2{,}228) = 0{,}05$$