



## Ejercicios resueltos del capítulo 6: MODELOS DE REGRESIÓN

### 1. Problema 1

Para hacer un modelo de regresión necesitamos lápiz (o bolígrafo), folios y una calculadora elemental. Nada más.

En las prácticas era suficiente con introducir los datos relativos a  $x$  y a  $y$ . Sin embargo, para hacer las cosas sin ordenador hay que trabajar un poquito más. Por ese motivo vamos a hacer ejercicios con pocos datos.

La idea es escribir una tabla como la siguiente:

	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$	
	0.8	1	0.64	1	0.8	$\bar{X} = \frac{4,3}{4} = 1,075$
	1	2	1	4	2	$\bar{Y} = \frac{11}{4} = 2,75$
	1.2	3	1.44	9	3.6	$S_{XY} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X}\bar{Y} = \frac{12,9}{4} - 1,075 \cdot 2,75 \approx 0,26875$
	1.3	5	1.69	25	6.5	$S_X^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2 = \frac{4,77}{4} - 1,075^2 \approx 0,0369$
Suma	4.3	11	4.77	39	12.9	

En dicha tabla, además de introducir los valores de  $x$  e  $y$ , nos ayudamos de la calculadora para hacer el resto de columnas y las sumas finales de cada una de ellas. A partir de esta tabla, y conociendo las fórmulas de la varianza y la covarianza, las calculamos tal y como aparecen a la derecha de la tabla.

A partir de las medias, las varianzas y la covarianza se calculan los coeficientes de la recta de regresión de  $y$  sobre  $x$ . Recordemos que en la recta de regresión  $y = a + bx$ , los coeficientes  $a$  y  $b$  están dados por las siguientes fórmulas:

$$b = \frac{S_{XY}}{S_X^2} \approx \frac{0,26875}{0,0369} \approx 7,283 \quad \text{y} \quad a = \bar{Y} - b\bar{X} \approx 2,75 - 7,283 \cdot 1,075 \approx -5,0847.$$

Por lo tanto, la recta es  $y = -5,0847 + 7,283x$ .

Esta recta es la que mejor predice el comportamiento de la variable  $y$  en función de la variable  $x$ . Así, para calcular lo que podemos esperar que cueste un automóvil de 1,1 Tm, basta sustituir en la recta de regresión la  $x$  por 1,1:  $y(1,1) = -5,0847 + 7,283 \cdot 1,1 = 2,9266$  millones. Éste es el valor esperado (o valor que predice) nuestra regresión lineal para  $x = 1,1$ .

Para saber si la predicción es fiable (si el ajuste es bueno), calculamos el coeficiente de correlación lineal  $r$ :

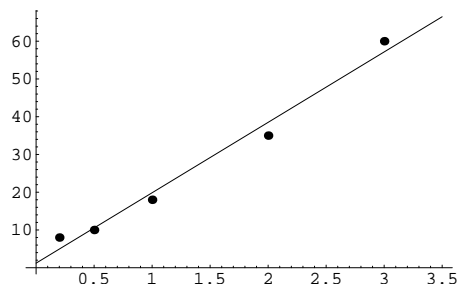
$$S_Y^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{Y}^2 = \frac{39}{4} - 2,75^2 \approx 2,1875, \quad \text{luego} \quad r = \frac{S_{XY}}{S_X S_Y} \approx \frac{0,26875}{\sqrt{0,0369} \sqrt{2,1875}} \approx 0,9459,$$

que es bastante próximo a 1. Por tanto, los resultados se pueden considerar fiables.

## 2. Problema 4

Si representamos los datos como puntos de coordenadas  $(x_i, y_i)$  en el plano vemos que, efectivamente, éstos podrían ajustarse a una recta, lo que nos indica que la velocidad de reacción aumenta “linealmente” con la concentración de glucogenasa.

Al igual que en el problema anterior, debemos elaborar una tabla con los valores observados de las variables  $x$  e  $y$ , a partir de ellos, completar las columnas siguientes ayudados de la calculadora.



	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
	0.2	8	0.04	64	1.6
	0.5	10	0.25	100	5
	1	18	1	324	18
	2	35	4	1225	70
	3	60	9	3600	180
Suma	6.7	131	14.29	5313	274.6

A partir de aquí, hacemos también el cálculo de los estadísticos descriptivos más sencillos: medias, varianzas y covarianza.

$$\begin{aligned} \bar{X} &= \frac{6,7}{5} = 1,34 & \bar{Y} &= \frac{131}{5} = 26,2 & S_{XY} &= \frac{274,6}{5} - 35,108 = 19,812 \\ S_X^2 &= \frac{14,29}{5} - 1,34^2 = 1,0624 & S_Y^2 &= \frac{5313}{5} - 26,2^2 = 376,18 \end{aligned}$$

A continuación, calculamos los coeficientes  $a$  y  $b$  de la recta de regresión  $y = a + bx$ :

$$b = \frac{S_{XY}}{S_X^2} = \frac{19,812}{1,0624} = 18,648343 \quad \text{y} \quad a = \bar{Y} - b\bar{X} = 26,2 - 18,648343 \cdot 1,34 = 1,2112204.$$

La recta de regresión es  $y = 1,2112204 + 18,648343x$ ; en la figura se ve cómo se ajustan los datos a ella.

Para calcular la velocidad de reacción a una concentración de 2,5 milimoles/litro, basta sustituir  $x$  por 2,5 en la recta de regresión:  $y(2,5) = 1,2112204 + 18,648343 \cdot 2,5 = 47,832078$  micromoles/minuto.

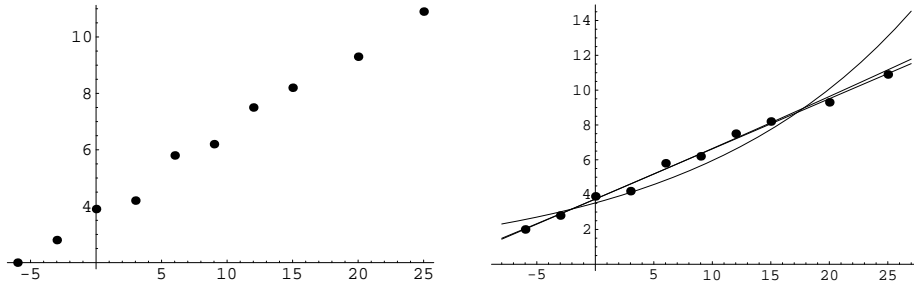
Finalmente, vemos si el ajuste lineal es bueno calculando el coeficiente de correlación lineal  $r$ :

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{19,812}{\sqrt{1,0624} \sqrt{376,18}} \approx 0,9910555,$$

que es muy próximo a 1. Por tanto, la dependencia lineal es buena.

### 3. Problema 5

La representación de la nube de puntos nos da la idea de que un buen ajuste va a ser el lineal aunque tampoco debemos descartar el ajuste exponencial sólo por el dibujo.



(a) Ajuste por una función lineal:

Como es habitual, introducimos la tabla con las columnas siguientes:

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$	
-6	2	36	4	-12	
-3	2.8	9	7.84	-8.4	
0	3.9	0	15.21	0	
3	4.2	9	17.64	12.6	
6	5.8	6	33.64	34.8	
9	6.2	81	38.44	55.8	
12	7.5	144	56.25	90	
15	8.2	225	67.24	123	
20	9.3	400	86.49	186	
25	10.9	625	118.81	272.5	
Suma	81	60.8	1535	445.56	754.3

$$\bar{X} = \frac{81}{10} = 8,1 \quad \bar{Y} = \frac{60,8}{10} = 6,08$$

$$S_{XY} = \frac{754,3}{10} - 8,1 \cdot 6,08 = 26,182$$

$$S_X^2 = \frac{1565}{10} - 8,1^2 = 90,89$$

Luego:

$$b = \frac{S_{XY}}{S_X^2} = \frac{26,182}{90,89} = 0,288$$

$$a = \bar{Y} - b\bar{X} = 6,08 - 0,288 \cdot 8,1 = 3,7467$$

Por tanto, la recta de regresión es  $y = 3,7467 + 0,288x$ .

(b) Ajuste por una función exponencial: Como la función buscada es  $y = ae^{bx}$ , tomando logaritmos tenemos que  $\log y = \log a + bx$ . Llamamos a la nueva variable  $y' = \log y$  y también hacemos  $a' = \log a$ . Tenemos entonces que calcular entonces la recta de regresión  $y' = a' + bx$  para las nuevas variables  $y'$  y  $x$ .

$x_i$	$y_i$	$y'_i = \log y_i$	$x_i^2$	$(y'_i)^2$	$x_i y'_i$	
-6	2	0.693	36	0.480249	-4.158	
-3	2.8	1.03	9	1.0609	-3.09	
0	3.9	1.361	0	1.852321	0	
3	4.2	1.435	9	2.059225	4.305	
6	5.8	1.758	6	3.090564	10.548	
9	6.2	1.825	81	3.330625	16.425	
12	7.5	2.015	144	4.060225	24.18	
15	8.2	2.104	225	4.426816	31.56	
20	9.3	2.23	400	4.9729	44.6	
25	10.9	2.389	625	5.707321	59.725	
Suma	81	60.8	16.84	1535	31.041146	184.095

$$\bar{X} = 8,1 \quad \bar{Y}' = \frac{16,84}{10} = 1,684$$

$$S_{XY'} = \frac{184,1}{10} - 8,1 \cdot 1,684 = 4,7696$$

$$S_X^2 = 90,89$$

Luego:

$$b = \frac{S_{XY'}}{S_X^2} = \frac{4,7696}{90,89} = 0,0525$$

$$a' = \bar{Y}' - b\bar{X} = 1,684 - 0,0525 \cdot 8,1 = 1,259$$

El valor que buscamos para la función exponencial es  $a$ , no  $a'$ ; pero

$$a = e^{a'} = e^{1,259} = 3,522.$$

Luego la función exponencial que mejor se ajusta a las observaciones es  $y = 3,522e^{0,0525x}$ .

Finalmente, para decidir cuál es el mejor modelo de los dos utilizados, calculamos los errores típicos:  $e_t = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y(x_i))^2}$ .

(a) Modelo lineal:

$$\begin{array}{llll} y_1 - y(x_1) = 0,0187 & y_2 - y(x_2) = 0,0827 & y_3 - y(x_3) = 0,1533 & y_4 - y(x_4) = -0,4107 \\ y_5 - y(x_5) = 0,3253 & y_6 - y(x_6) = -0,1387 & y_7 - y(x_7) = 0,2973 & y_8 - y(x_8) = 0,1333 \\ y_9 - y(x_9) = -0,2067 & y_{10} - y(x_{10}) = -0,0467 & & \end{array}$$

Tomamos estas diferencias, hacemos sus cuadrados y las sumamos. Finalmente dividimos por 10 y hacemos la raíz cuadrada. Entonces,  $e_t = 0,218$ .

(b) Modelo exponencial:

$$\begin{array}{llll} y_1 - y(x_1) = -0,57 & y_2 - y(x_2) = -0,2088 & y_3 - y(x_3) = 0,378 & y_4 - y(x_4) = 0,0772 \\ y_5 - y(x_5) = 0,974 & y_6 - y(x_6) = 0,551 & y_7 - y(x_7) = 0,887 & y_8 - y(x_8) = 0,459 \\ y_9 - y(x_9) = 0,765 & y_{10} - y(x_{10}) = -2,186 & & \end{array}$$

Análogamente,  $e_t = 0,902$ .

El error menor es el correspondiente al modelo lineal, que será, por tanto, el más adecuado.