UNIDAD TEMÁTICA 9

REGRESIÓN LINEAL Y CORRELACIÓN

ENUNCIADO 1

La siguiente tabla muestra la nota final en los exámenes de estadística (E) e investigación operativa (IO) de una muestra aleatoria de 20 alumnos. Supongamos que las notas finales están conjuntamente distribuidas de modo normal.

E	86	75	69	75	90	94	83	86	71	65	84	71	62	90	83	75	71	76	84	97
IO	80	81	75	81	92	95	80	81	76	72	85	72	65	93	81	70	73	72	80	98

(A) *Estima el error que se ha cometido en la evaluación de dichos exámenes si se sugiere un modelo lineal que explique la nota en "IO" como función de la nota alcanzada en "E" con un nivel de confianza* α *= 99 %.* **(B)** *¿Qué tanto por ciento se puede imputar al azar a partir de la muestra presentada?* **(C)** *¿Existe relación lineal entre ambas variables? Comenta los resultados que se obtienen.*

Resolución:

De los datos que proporciona la tabla de frecuencias se deduce que

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} = 0.8234$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 14.76$$

con lo que la recta de regresión es $IO = 0.8234E + 14.76$ ($E = 0.9919IO - 0.02804$).

(**A**) Una estimación del error que se ha cometido en la prueba es

$$\hat{\sigma} = \sqrt{\frac{SS_E}{n-2}} = 3.9127$$

Tratándose de un problema de desviaciones típicas el modelo de probabilidad que hay que utilizar es el modelo χ^2 con $\nu = n-2$ grados de libertad, que para este caso serán (estimación intercalar)

$$\chi_1^2 = \chi^2_{\alpha=0.5\%,\nu=18\ gdl} = 6.26$$
$$\chi_2^2 = \chi^2_{\alpha=99.5\%,\nu=18\ gdl} = 37.16$$

De la desigualdad

$$\chi_1^2 \le \chi^2 = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \le \chi_2^2$$

se obtienen los límites del intervalo de confianza buscados

$$\frac{SS_E}{\chi_2^2} \le \sigma^2 \le \frac{SS_E}{\chi_1^2}$$

que para el caso concreto que se plantea implica que

$$\sqrt{7.425533711} = 2.723147758 \le \sigma^2 \le \sqrt{44.01936624} = 6.634709206$$

(**B**) Dado que

$$r^2 = \frac{SS_{YY} - SS_E}{SS_{YY}} = \frac{SS_R}{SS_{YY}} = 0.816022678 \Rightarrow 1 - r^2 = 19.3977322\%$$

Es decir, la suma de los cuadrados de las desviaciones de Y respecto de los valores estimados se redujo en un 81.60 % al utilizar $\widehat{IO} = E\left[IO \mid E\right]$ en lugar de $\bar{y}\big|_{IO}$ para predecir $Y\big|_{IO}$.

(**C**) Se trata de un contraste bilateral (de dos colas) de hipótesis sobre β_1 donde las hipótesis a trabajar son:

$$\begin{cases} H_0: & \beta_1\big|_0 = 0 \\ H_a: & \beta_1\big|_0 \neq 0 \end{cases}$$

El error probable que se obtiene en la estimación del parámetro β_1 es

$$\sigma^2_{\hat{\beta}_1} = \frac{\hat{s}^2}{SS_{XX}} = \frac{SS_E}{(n-2)SS_{XX}} = 0.08492944$$

El estadístico del contraste se obtiene a partir de

$$t = \frac{\hat{\beta}_1 - \beta_1|_0}{\sigma^2_{\hat{\beta}_1}} = \frac{0.8234}{0.09215717} = 8.93521939$$

Y la región de admisibilidad está delimitada por:

$$t_1 = t_{\forall \alpha, \nu = n-2 = 20-2 = 18\ gdl} > t_{contraste} = 8.93521939$$

En consecuencia, como $t > |t_1|$ a partir de la muestra seleccionada existe evidencia estadística para rechazar la hipótesis nula (es decir, existe relación lineal entre ambas variables).

ENUNCIADO 2

Dos montañeros han utilizado dos instrumentos GPS, A y B, para medir los tiempos invertidos en una ruta de montaña con los mismos puntos de control, aunque han salido en horas diferentes. Se les ha indicado que ambos aparatos habían sido bien regulados. Dichos datos se muestran en la tabla siguiente:

PUNTO DE CONTROL	TIEMPO MEDIDO POR EL GPS_A (min)	TIEMPO MEDIDO POR EL GPS_B (min)
1	22.33	27.35
2	10.21	17.43
3	35.55	40.12
4	31.16	36.47
5	17.23	15.34
6	29.34	27.63
7	25.01	31.14
8	18.77	21.35

(1º) *Calcula la recta de regresión,* $E(Y) = \beta_0 + \beta_1 X$ *, del tiempo del GPSB (Y) en función del tiempo del GPSA (X).* **(2º)** *¿Calcula la reducción de la suma de los cuadrados de las desviaciones de los valores de Y respecto de sus valores estimados* $\hat{Y}$ *, que se pueden atribuir a una relación lineal entre Y y X?* **(3º)** *Para un tiempo t = 20 min invertido en un tramo por el montañero A estima el tiempo invertido por el montañero B.*

Resolución:

(**1º**) El diagrama de dispersión, junto con la tabla de frecuencias, viene dado por:

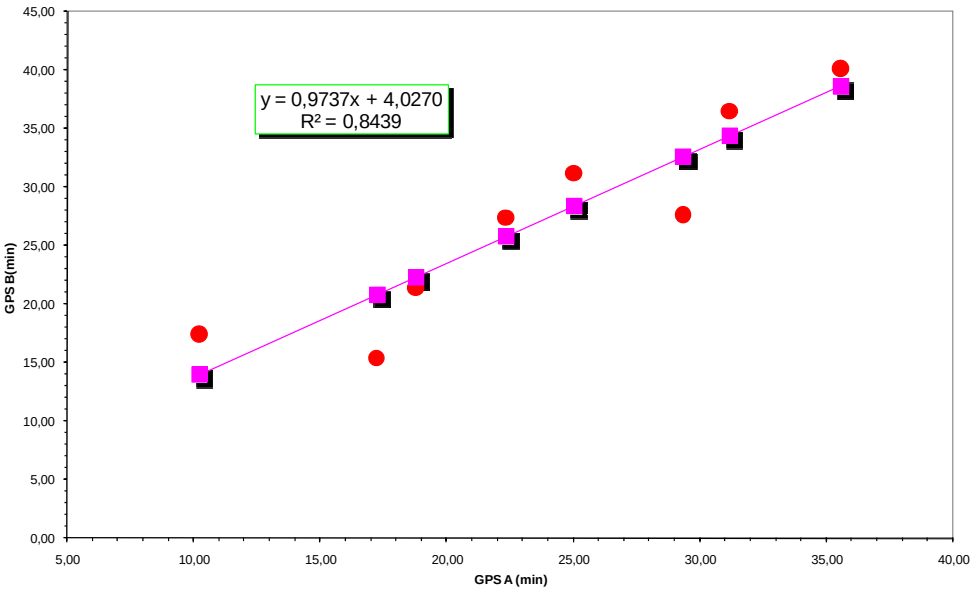
	GPSA	GPSB					
n	X	$Y_{observada}$	X_i^2	X_iY_i	Y_i^2	$Y_{teórica}$	$Residuo^2$
1	22,33	27,35	498,6289	610,7255	748,0225	25,7698	2,4971
2	10,21	17,43	104,2441	177,9603	303,8049	13,9685	11,9819
3	35,55	40,12	1263,8025	1426,266	1609,6144	38,6421	2,1841
4	31,16	36,47	970,9456	1136,4052	1330,0609	34,3676	4,4202
5	17,23	15,34	296,8729	264,3082	235,3156	20,8039	29,8542
6	29,34	27,63	860,8356	810,6642	763,4169	32,5954	24,6555
7	25,01	31,14	625,5001	778,8114	969,6996	28,3793	7,6215
8	18,77	21,35	352,3129	400,7395	455,8225	22,3034	0,9090
Sumas	8	189,6	216,83	4973,1426	5605,8803	6415,7573	84,12347807

β_1	0,9737
β_0	4,02702
σ^2	14,0206
σ	3,74441

SS_{XX}	479,6226
SS_{YY}	538,851187
SS_{XY}	467,0093
SS_E	84,1234781
SS_R	454,727709

$\hat{\sigma}^2$	18,17218947
$\hat{\sigma}$	0,029232525
r	0,918631403
r^2	0,843883655

Modelo de regresión lineal (mínimos cuadrados)



De la correspondiente tabla de frecuencias de las dos variables, siendo X = GPS_A e Y = GPS_B, se deducen los siguientes valores numéricos:

$$\sum_{i=1}^{n} x_i = 189.6 \text{ min}; \sum_{i=1}^{n} y_i = 216.83 \text{ min};$$

$$\sum_{i=1}^{n} x_i^2 = 4973.14 \text{ min}^2; \sum_{i=1}^{n} y_i^2 = 5605.88 \text{ min}^2; \sum_{i=1}^{n} x_i y_i = 6415.76 \text{ min}^2$$

siendo la suma de los cuadrados de los errores:

$$SS_{XX} = \sum_{\forall i}(x_i - \overline{x})^2 = \sum_{\forall i} x_i^2 - \frac{1}{n}\left(\sum_{\forall i} x_i\right)^2 = 479.62 \text{ min}^2$$

$$SS_{YY} = \sum_{\forall i}(y_i - \overline{y})^2 = \sum_{\forall i} y_i^2 - \frac{1}{n}\left(\sum_{\forall i} y_i\right)^2 = 538.85 \text{ min}^2$$

$$SS_{XY} = \sum_{\forall i}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{\forall i} x_i y_i - \frac{1}{n}\sum_{\forall i} x_i \sum_{\forall i} y_i = 467.00 \text{ min}^2$$

de donde se pueden deducir los valores estimados de la pendiente y de la ordenada en origen de la recta de regresión en el sentido de mínimos cuadrados:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} = 0.9737$$

$$\hat{\beta}_0 = \frac{1}{n}\left(\sum_{\forall i} y_i - \hat{\beta}_1 \sum_{\forall i} x_i\right) = 4.027 \text{ min}$$

Entonces, la recta de regresión es:

$$GPS_B = 0.9737 GPS_A - 4.0270$$

(**2º**) Análogamente, el coeficiente de correlación es

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX}}\sqrt{SS_{YY}}} \quad \text{(adimensional)} = 0.9186$$

y el coeficiente de determinación:

$$r^2 = \frac{SS_{YY} - SS_E}{SS_{YY}} = \frac{SS_{XY}^2}{SS_{XX} SS_{YY}} = 0.8439$$

que indica que existe un 25.61 % que no es explicado por el modelo lineal propuesto (y que es debido al azar, a otras relaciones que no se han puesto de manifiesto en el modelo, etc.).

(**3º**) La interpolación que se pide es:

$$GPS_B\big|_{t=20 \text{ min}} = 0.9737 \times 20 + 4.0270 = 23.5010 \text{ min}$$

ENUNCIADO 3

Una empresa química conoce que el coste de fabricar un determinado detergente responde a una ley lineal del tipo $y = a + b.x$, *donde* y = *Coste en €,* x = *Detergente fabricado en* m^3, a = *Coste fijo en €,* b = *Coste adicional de la producción en €/*m^3. *Los datos recogidos en 7 plantas de la empresa en el último ejercicio son los siguientes:*

Planta	**1**	**2**	**3**	**4**	**5**	**6**	**7**
x **(m^3)**	2200	3400	4600	5500	8000	9100	10000
y **(€)**	141000	190000	240000	300000	450000	450000	530000

(A) *Obtener, a partir de estos datos, la recta de regresión.*
(B) *De acuerdo con el resultado obtenido en (A) decir cuál es la estimación del coste fijo en €, así como el coste adicional en €/m3, de la producción.*
(C) *Calcular el coeficiente de correlación de Pearson entre las variables* x *e* y.
(D) *Estimar cuál será el gasto en € de una planta que pretenda producir 6000 m3 de detergente.*