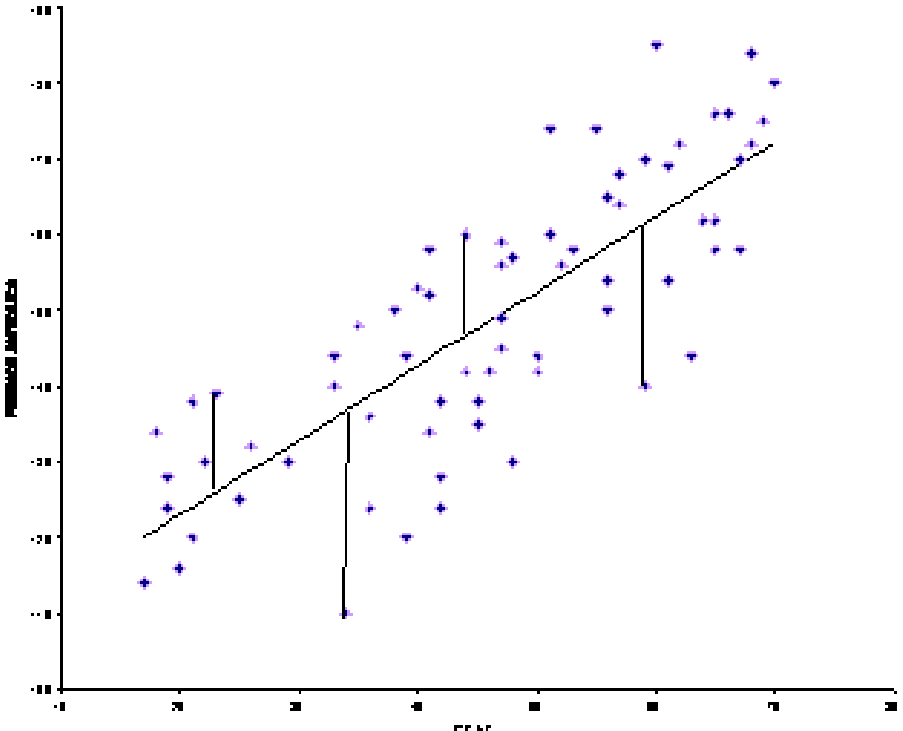
La recta de regresión

$(x_1, y_1), (x_2, y_2),...,(x_N, y_N)$: N pares de puntos observados

Hemos de encontrar una recta: $y = \alpha + \beta x$ que se ajuste “lo mejor posible” a nuestros puntos:



¿Cómo ajustar la recta?

- Queremos predecir la variable y en función de la variable x.
- Si usamos una recta $y = \alpha + \beta x$, entonces los residuos o errores de predicción son $r_i = y_i - \alpha - \beta x_i$ para i = 1,…,N.
- Intentamos minimizar el error.
- Usamos el criterio de mínimos cuadrados: elegimos la recta que minimiza Σr_i^2
- La recta de mínimos cuadrados es $y = a + bx$

  donde b es la pendiente de la recta y a es el intercepto:

$$b = \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x}$$

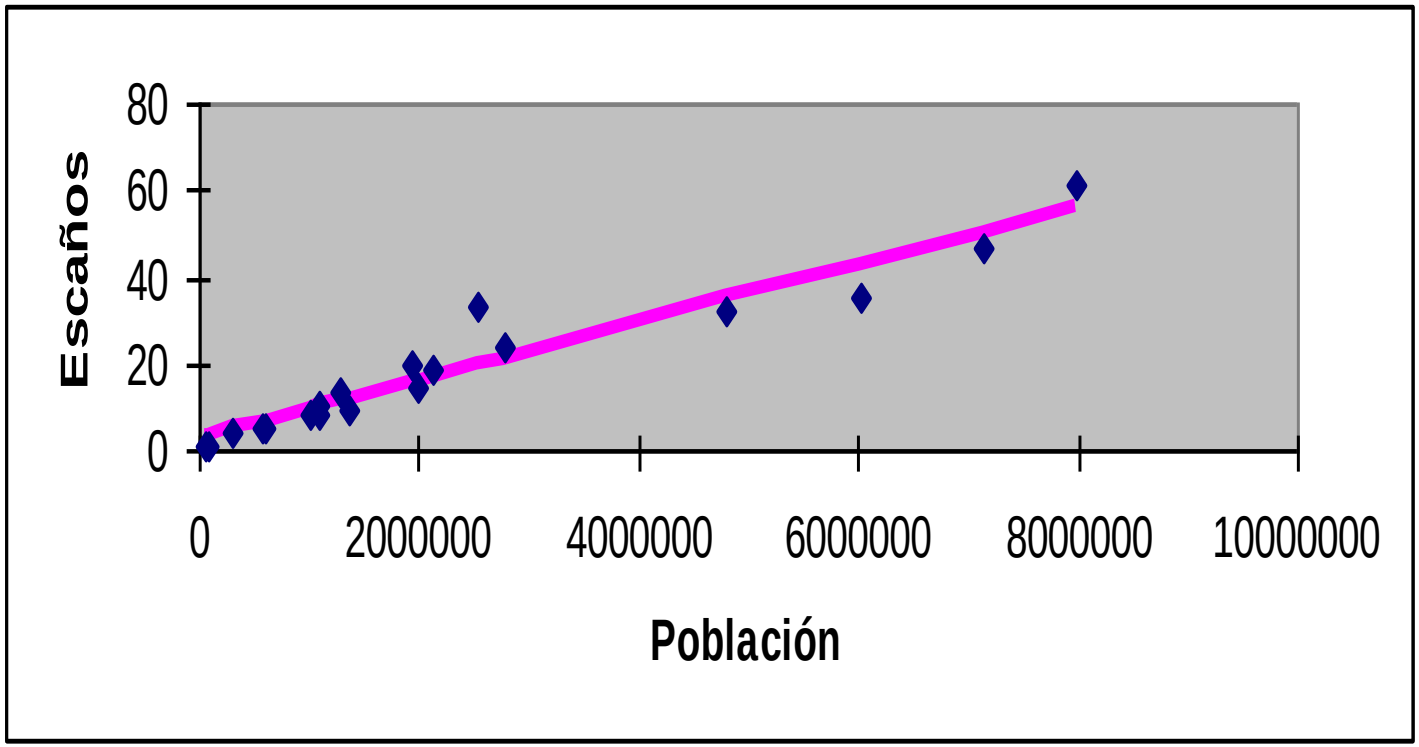
$$a = \bar{y} - b\bar{x}$$

Demostración

$$
\begin{aligned}
\sum_{i=1}^{N} r_i^2 &= \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2 \\
&= \sum_{i=1}^{N} (y_i - \bar{y} + \bar{y} - \alpha - \beta x_i + \beta\bar{x} - \beta\bar{x})^2 \\
&= \sum_{i=1}^{N} (y_i - \bar{y} - [\alpha - \bar{y} + \beta\bar{x}] - \beta[x_i - \bar{x}])^2 \\
&= \sum_{i=1}^{N} (y_i - \bar{y})^2 + (\alpha - \bar{y} + \beta\bar{x})^2 + (\beta[x_i - \bar{x}])^2 \\
&\quad -2(y_i - \bar{y})(\alpha - \bar{y} + \beta\bar{x}) - 2(y_i - \bar{y})\beta(x_i - \bar{x}) + 2(\alpha - \bar{y} + \beta\bar{x})\beta(x_i - \bar{x}) \\
&= N s_y^2 + N(\alpha - a)^2 + N\beta^2 s_x^2 - 2(\alpha - a)\sum_{i=1}^{N}(y_i - \bar{y}) - 2N\beta s_{xy} + 2(\alpha - a)\beta\sum_{i=1}^{N}(x_i - \bar{x}) \\
&= N s_y^2 + N(\alpha - a)^2 + N\beta^2 s_x^2 - 2N\beta s_{xy} \\
&= N s_y^2 + N(\alpha - a)^2 + N\beta^2 s_x^2 - 2N\beta s_{xy} + N\left(\frac{s_{xy}}{s_x}\right)^2 - N\left(\frac{s_{xy}}{s_x}\right)^2 \\
&= N s_y^2 + N(\alpha - a)^2 + N s_x^2\left(\beta - \frac{s_{xy}}{s_x^2}\right)^2 - N\left(\frac{s_{xy}}{s_x}\right)^2 \\
&= N(\alpha - a)^2 + N s_x^2 (\beta - b)^2 + N s_y^2 - N\left(\frac{s_{xy}}{s_x}\right)^2
\end{aligned}
$$

y se minimiza esta función con $\alpha = a$ y $\beta = b$.

Escaños y población:
La recta de regresión ajustada



Output de Excel

	Coeficientes
Intercepción	2,692069443
Variable X 1	6,68437E-06

La recta ajustada es y = 2,69+0,0000069x

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,96372808
Coeficiente de determinación R^2	0,928771813
R^2 ajustado	0,92458192
Error típico	4,544275594
Observaciones	19

¿Cómo predecimos el número de escaños en una comunidad de 1000000 de personas?

¿Y en una comunidad sin gente? ¿Tiene sentido la predicción?

Análisis de los residuos I: la media y varianza residual

Se puede demostrar que la media de los residuos es 0.

$$
\begin{aligned}
\sum_{i=1}^{N} r_i &= \sum_{i=1}^{N} (y_i - a - bx_i) \\
&= \sum_{i=1}^{N} (y_i - a - bx_i) \\
&= \sum_{i=1}^{N} (y_i - \bar{y} + b\bar{x} - bx_i) \\
&= \sum_{i=1}^{N} (y_i - \bar{y}) - b \sum_{i=1}^{N} (x_i - \bar{x}) \\
&= 0
\end{aligned}
$$

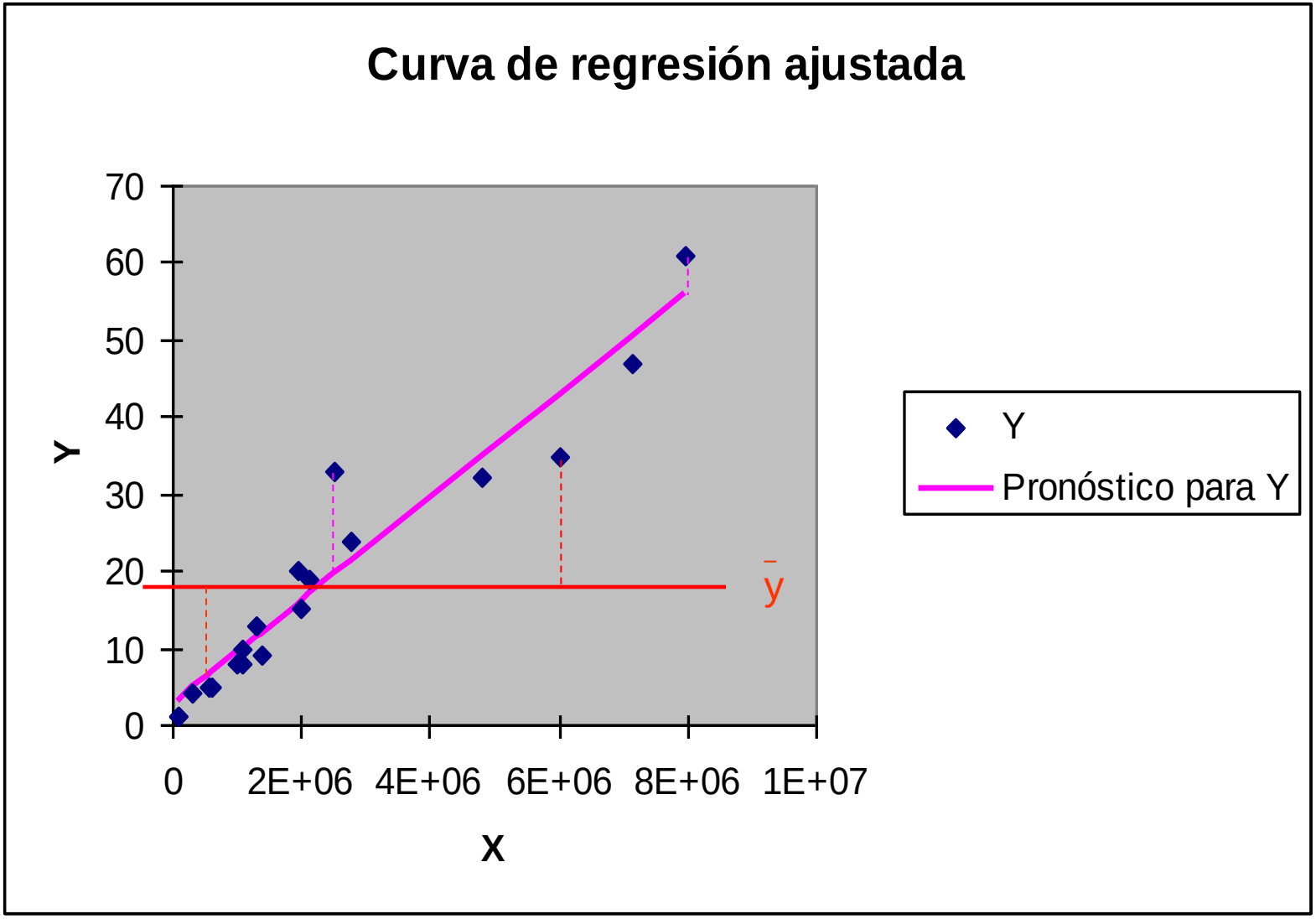
y se puede calcular la varianza residual

$$\begin{aligned}\sum_{i=1}^{N} r_i^2 &= N s_y^2 - N\left(\frac{s_{xy}}{s_x}\right)^2 \\ &= N s_y^2 \left(1 - \frac{s_{xy}}{s_x s_y}\right)^2 \\ &= N s_y^2 \left(1 - r_{xy}^2\right)\end{aligned}$$

Luego, la varianza residual es

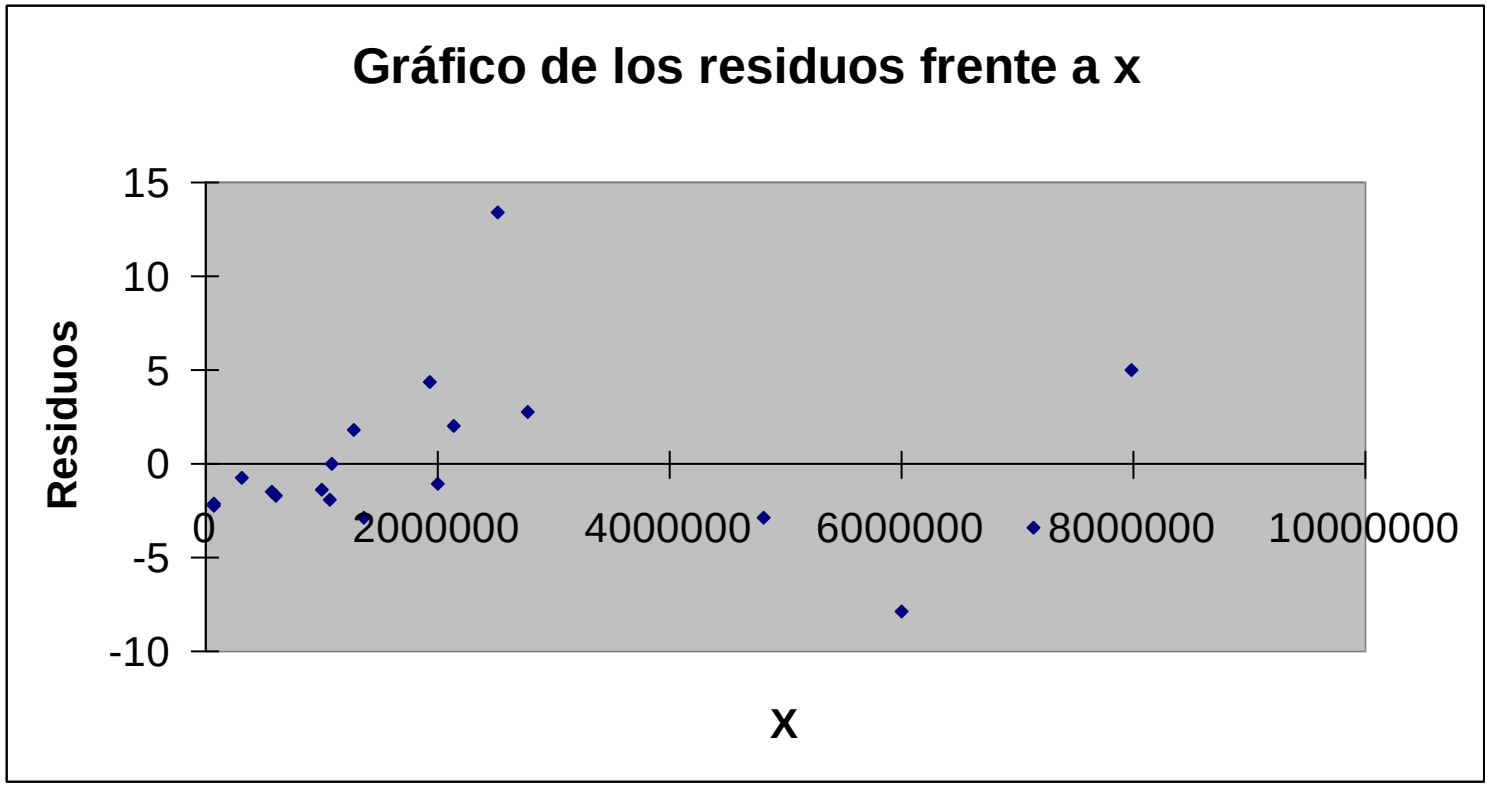
$$s_r^2 = \frac{1}{N}\sum_{i=1}^{N}(r_i - \bar{r})^2 = \frac{1}{N}\sum_{i=1}^{N} r_i^2 = s_y^2(1 - r_{xy}^2).$$

¿Cómo interpretamos esta expresión?


Curva de regresión ajustada
70
60
50
40
30
20
10
0
Y
0
2E+06
4E+06
6E+06
8E+06
1E+07
X
ȳ
Y
Pronóstico para Y

Análisis de los residuos II: gráficos

Si la recta de regresión se ajusta bien, los residuos deben aparecer como ruido aleatorio sin relación ninguna con x o y.



¿Parece bien el ajuste?

Ejercicio (Pregunta de Test)

En una oficina se desea conocer el grado de satisfacción de los empleados. Para ello se realiza un cuestionario de satisfacción a 10 de ellos y se les pide que valoren, en una escala continúa de 0 a 10, el ambiente en su puesto de trabajo. El valor 0 identifica un pésimo ambiente de trabajo y el 10 identifica un inmejorable ambiente de trabajo.
Además se recoge la edad de los empleados.
Asumiendo que la valoración depende de la edad se ha estimado la recta de regresión obteniéndose:

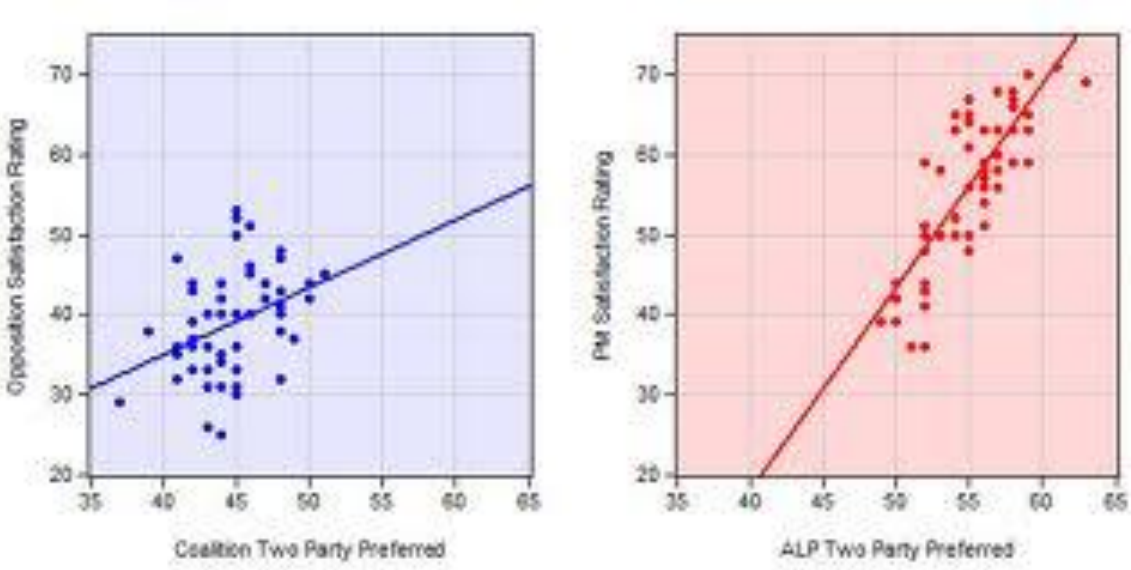
$$\hat{y}_i = 6.13 - 0.087 x_i$$

Ahora se desearía conocer cual es la valoración media para un nuevo trabajador cuya edad es 43 años. Di cual de las siguientes opciones es la correcta:

a) 2.19 puntos
b) 2.39 puntos
c) 4.69 puntos
d) -2.05 puntos

Ejercicio (Pregunta de Test)

Los siguientes gráficos muestran los niveles de satisfacción con el líder de la oposición (lado izquierdo) y el primer ministro (lado derecho) como función del voto preferido.

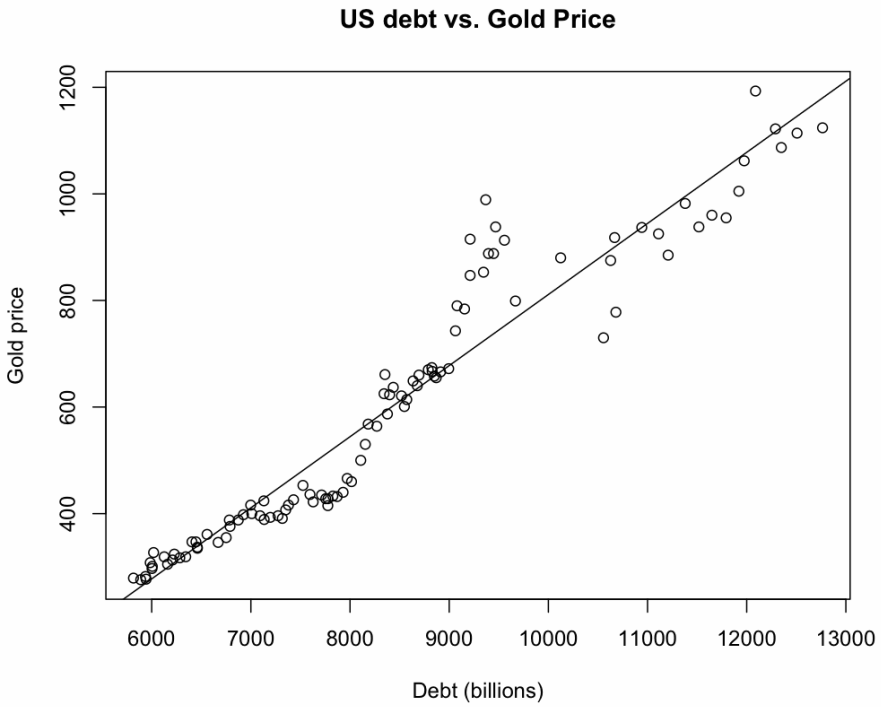


¿Cuál de las siguientes frases es la correcta?

a) En ambos casos, la correlación entre satisfacción y voto preferido es negativa.
b) La correlación con el voto preferido es más alta para el líder de la oposición.
c) La correlación es más alta en el caso del primer ministro.
d) El pendiente es igual para ambas rectas de regresión.

Ejercicio (Pregunta de Test)

El diagrama muestra el nivel de la deuda Americana como función del precio de oro.



La fórmula para la recta de regresión es:

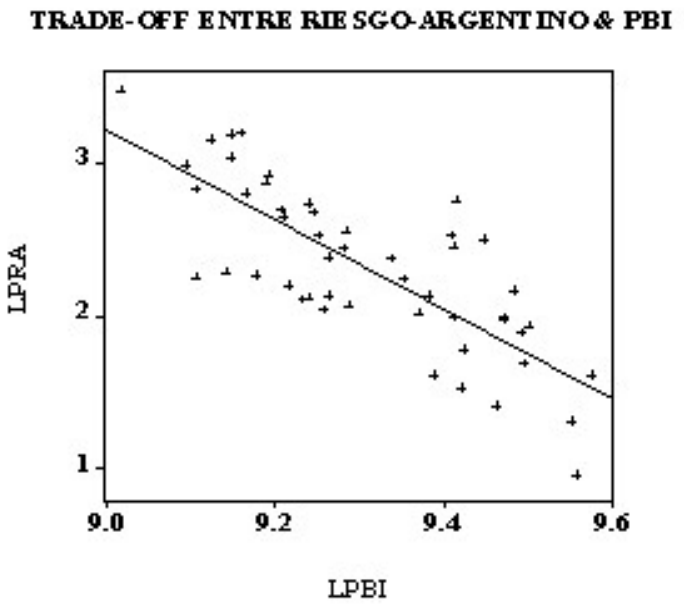
PRECIO DE ORO (nominal) = -522,86 + (0,1334 * deuda en $ billones)

Si la deuda Americana es de $19000 billones, calcular la predicción para el precio de oro.

a) 2011,74
b) 3057,46
c) 2933,14
d) -520,3254

Ejercicio (Pregunta de Examen)

El siguiente gráfico muestra la relación ente el riesgo argentino (LPRI) y el PBI (LPBI).



¿Señala cuál de los siguientes es lo correcto?

a) La línea de regresión es LPRA = 3,15+2,5 LPBI.
b) La correlación entre LPRA y LPBI es igual a cero.
c) La correlación entre LPRA y LPBI es negativa.
d) Ninguno de los anteriores.

Ejercicio (Pregunta de Examen)

El gráfico siguiente muestra los niveles de conocimiento de Griego y de Latín para 10 jueces. Llamamos Y al nivel de conocimiento de Griego y X al nivel de conocimiento de Latín. Si utilizamos la nota de Latín para determinar la nota en Griego mediante una recta de regresión, observando el diagrama de dispersión, ¿cuál de las opciones mostradas abajo podría ser la recta correcta?

a) Y=1.97+0.64X
b) Y=1.97-0.64X
c) Y=-1.97+0.64X
d) Y=-1.97-0.64X