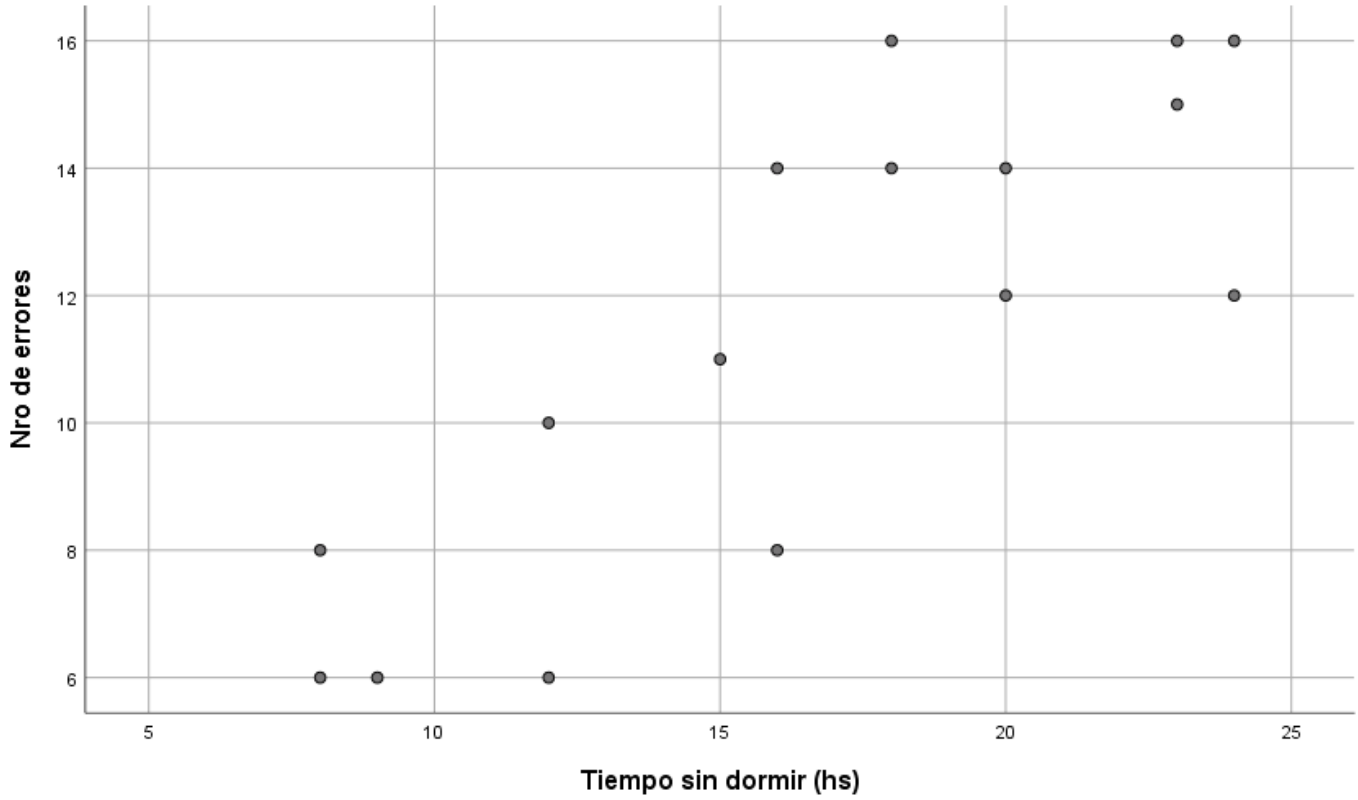
a) Grafica de dispersión



A simple vista no parece haber una relación lineal entre las variables, los puntos están muy dispersos entre sí, aunque se puede apreciar una cierta tendencia a aumentar los errores a medida que aumentan las horas sin dormir.

b) Modelo de regresión lineal:

Se busca la función de x más simple, lineal, que permita aproximar el valor de Y, mediante la fórmula:

$$\hat{Y} = a + bx$$

X: Variable independiente, predictora o explicatora, generalmente un dato en base al cual se quiere encontrar su valor correspondiente de la variable y.
Y: Variable dependiente, predicha o explicada, la cual varía en función de x.

Generalmente el valor de Y no coincide con el valor de $\hat{Y}$, ya que se encuentra afectada por un error el cual se denomina residuo o error residual:

$$e = Y - \hat{Y}$$

La recta muestral que estima a la población será:

$$Y = \alpha + \beta x + \varepsilon$$

Donde α es la ordenada al origen, β la pendiente, y ε es una variable aleatoria con E(ε)=0 y Var(ε)= σ^2.

El supuesto del modelo es que los residuos son independientes y se distribuyen en forma normal con media 0 y varianza constante σ^2, de no cumplirse esto, el modelo será invalido.

c) Siendo las variables:

X: Tiempo sin dormir(hs)
Y: Nro de errores

Se obtiene la siguiente tabla de Excel:

	Tiempo sin dormir(hs)	Nro de errores	x^2	y^2	x*y					
	8	8	64	64	64					
	8	6	64	36	48					
	9	6	81	36	54		Sxx=	469,75		
	12	6	144	36	72		Syy=	210		
	12	10	144	100	120		Sxy=	261		
	15	11	225	121	165		S^2=	4,64175473		
	16	8	256	64	128		S=	2,15447319		
	16	14	256	196	224					
	18	16	324	256	288		Error α	-5,17424304		
	18	14	324	196	252		Error β	-0,29591269		
	20	14	400	196	280					
	20	12	400	144	240		Confianza	99		
	23	15	529	225	345		t:	-2,97684273		
	23	16	529	256	368					
	24	12	576	144	288		b=	4176	=	0,55561469
	24	16	576	256	384			7516		
							a=	2,2629058		
SUMAS	266	184	4892	2326	3320					
Promedio	16,625	11,5								
N	16									

$$b = \frac{S_{xy}}{S_{xx}} = \frac{261}{469{,}75} = 0{,}556$$

$$a = \bar{y} - b\bar{x} = 11{,}5 - 16{,}625 * 0{,}556 = 2{,}226;$$
$$Utilizando\ todos\ los\ decimales\ el\ resultado\ es\ a = 2{,}263$$

La ecuación de la recta finalmente será:

$$\hat{y} = 2{,}263 + 0{,}556\ x$$

d) Intervalo de 99% de confianza para β.

$$b - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{S_{xx}}} < \beta < b + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{S_{xx}}}$$

$$t_{\frac{\alpha}{2}} \frac{S}{\sqrt{S_{xx}}} = 0{,}2959$$

$$0{,}556 - 0{,}2959 < \beta < 0{,}556 + 0{,}2959$$

$$0{,}2601 < \beta < 0{,}8519$$

e) Para poder estimar la cantidad de errores en un sujeto con 18,5 horas sin dormir, solo basta con reemplazar x=18,5 en la ecuación de la recta

$$\hat{y} = 2{,}263 + 0{,}556 * 18{,}5$$
$$\hat{y} = 12{,}549$$

Como el número de errores es un numero entero, se estima que cometerá 13 errores.

f) No, el modelo es utilizable solo para valores que se encuentren entre el valor máximo y mínimo de los datos usados para construirlo.

g) El coeficiente de correlación muestral es el denominado r el cual se calcula de la siguiente manera:

$$\hat{\rho} = r = b * \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} * S_{yy}}}$$

Este determina la relación lineal entre las variables de estudio.

h)

$$r = \frac{261}{\sqrt{469{,}75 * 210}} = 0{,}831$$

i) El coeficiente de determinación muestral es simplemente r^2=0,69, este explica el porcentaje de la dependencia de la variable dependiente en función de la independiente, para este caso, 69%.

j) Test:

H_0: ρ=0; H_1: ρ>0

$$t = \frac{b}{s/\sqrt{S_{xx}}} = \frac{0{,}556}{2{,}15/\sqrt{469{,}75}} = 5{,}605$$

El p-value de ese valor t es un número muy pequeño, menor a 0,01

Como el p-value obtenido es menor al nivel de significancia se rechaza la hipótesis nula y se puede afirmar que ρ es mayor a 0.

k) Test:

H_0: Los residuos se distribuyen de manera normal
H_1: Los residuos NO se distribuyen de manera normal

Salida del software:

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Unstandardized Residual	,139	16	,200*	,965	16	,758

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Como n<50 utilizamos el test de Shapiro-Wilk, se observa que el p-value calculado es 0,758, un número muy elevado, por el cual no rechazamos H_0 y podemos decir nuestro modelo es válido.