Para hacer regresión lineal, ¡grafique!

Esquema para recordar algunos puntos expuestos por Florentino Menéndez en la mesa redonda sobre regresión lineal múltiple, junio 2002. Cátedra de Metodología de la Investigación III
Departamento de Sociología - Universidad de la República - Uruguay

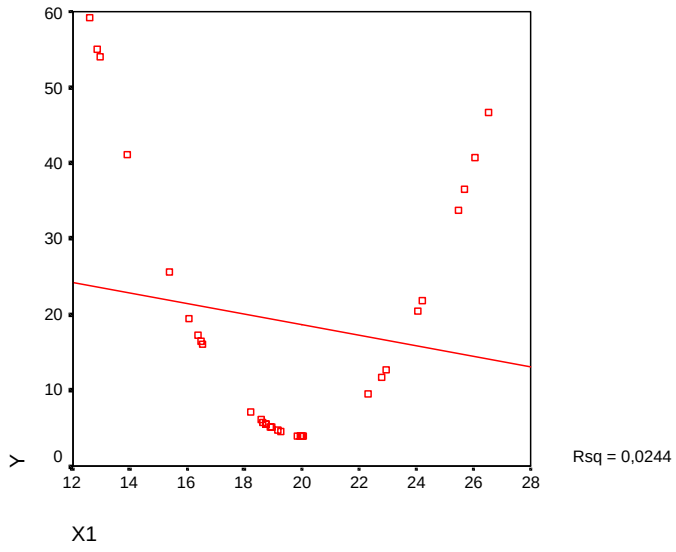
Hay razones de todo tipo que aconsejan graficar a la hora de hacer regresión lineal simple.

Relaciones curvilíneas.

La primer razón por la cual no se puede calcular la r de Pearson y los coeficientes de regresión sin graficar primero, es que la relación entre x e y puede ser curvilínea.

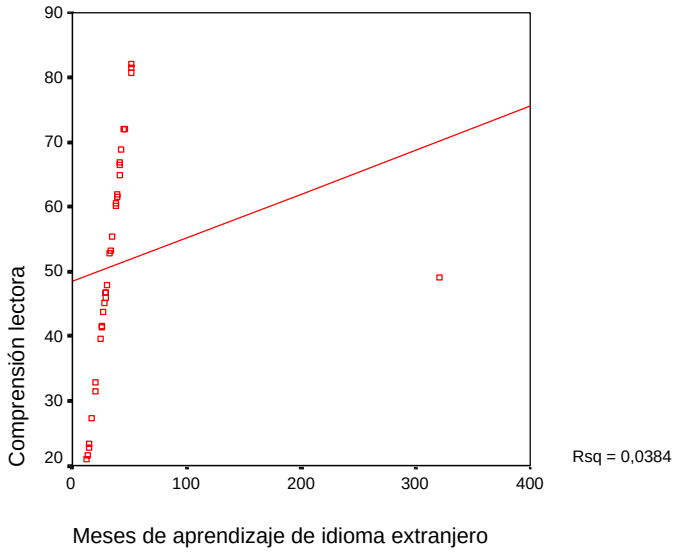
En la gráfica que sigue, se ve una relación perfecta entre x e y. Es perfecta porque conociendo x podemos saber el valor exacto que tomará y. Sin embargo, si hubiésemos calculado irreflexivamente r2, nos hubiese dado, tal cual se ve, 0.02, y quizás habríamos pensado que no había asociación.

Para evitarse este problema, **¡grafique!**



Datos erróneos.

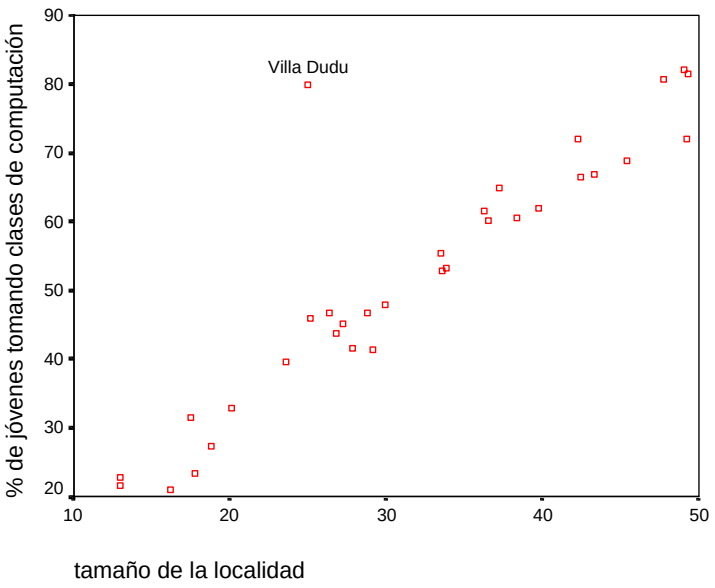
En el ejemplo que sigue se muestran los daños que puede causar un dato entrado erróneamente y no depurado. Aquí se entró un valor x sin la coma correspondiente. Como se verá, un r2 de Pearson que debía ser muy alto, quedó en 0.04 y la pendiente no ajusta bien con los datos correctamente entrados.



La graficación ayuda a detectar errores. Por tanto, para depurar, **¡grafique!**

Casos desviantes

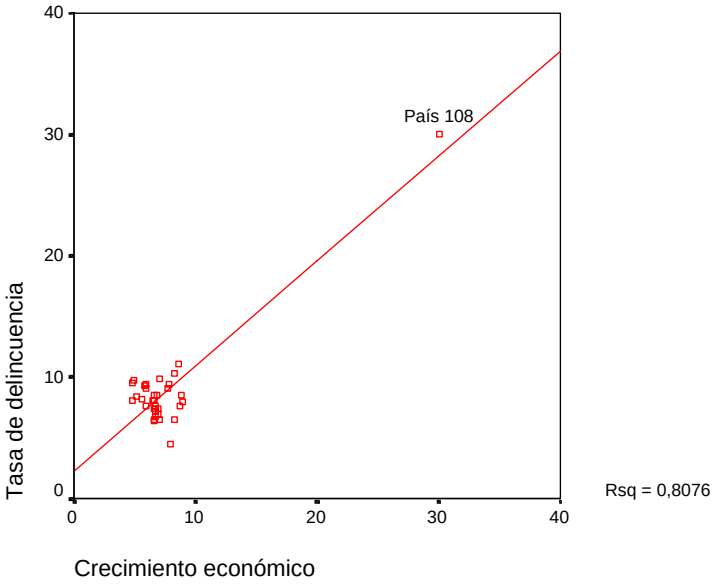
En ocasiones hay casos que se apartan fuertemente de lo esperable. En nuestro caso, la hipotética Villa Dudú muestra muchos más jóvenes aprendiendo computación de lo que esperaríamos. Esto nos lleva a preguntarnos ¿qué pasa allí? Estudiando Villa Dudú quizás localicemos nuevas variables que inciden sobre la variable dependiente. Quizás allí se haya instalado una empresa que exporta software.



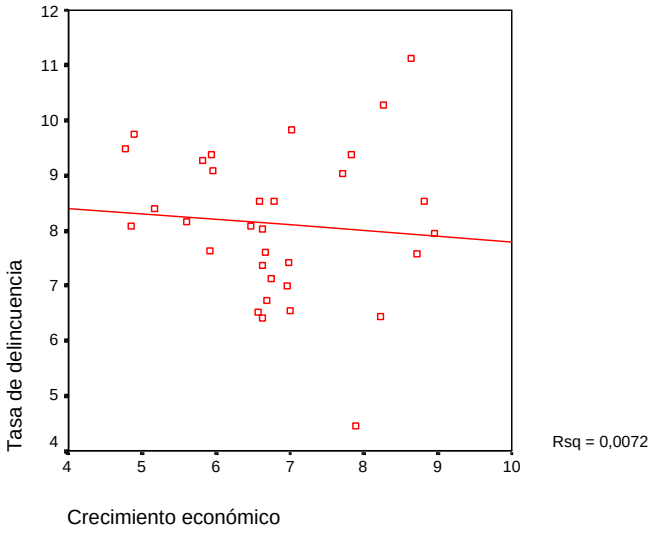
Para detectar casos desviantes, **¡vuelva a graficar!**

Puntos influyentes.

Los outliers en el eje de las x, tienen un muy fuerte impacto sobre la recta de regresión. Veamos el siguiente ejemplo:



Estamos estudiando la relación entre crecimiento económico y tasas de delincuencia. Sucede que un país hipotético, que hemos identificado como País 108, parece tener una fortísima influencia sobre r2 y sobre la recta de regresión. Para saber que tanto pesa, graficamos sin el punto influyente.



Según puede verse, al retirar del análisis el País 108, cambió radicalmente la pendiente de la recta de regresión y r2. El punto que parecía ser influyente, lo era!

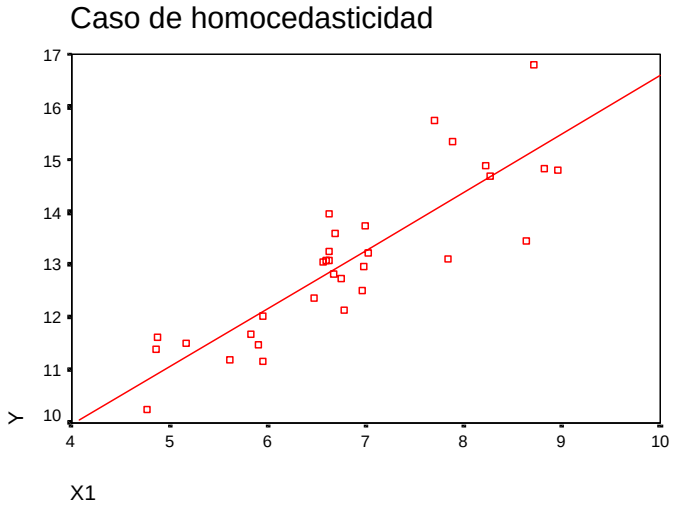
Dar por buenos sin mayor reflexión el r2 y los coeficientes de regresión de la primera gráfica sin advertir que dependen de un solo punto, es sacar conclusiones aventuradas. Enfrentados a una situación de este tipo, debemos ser concientes de ella, y elegir reflexivamente como proceder.

Para detectar los puntos influyentes... ¡ya sabe lo que tiene que hacer!

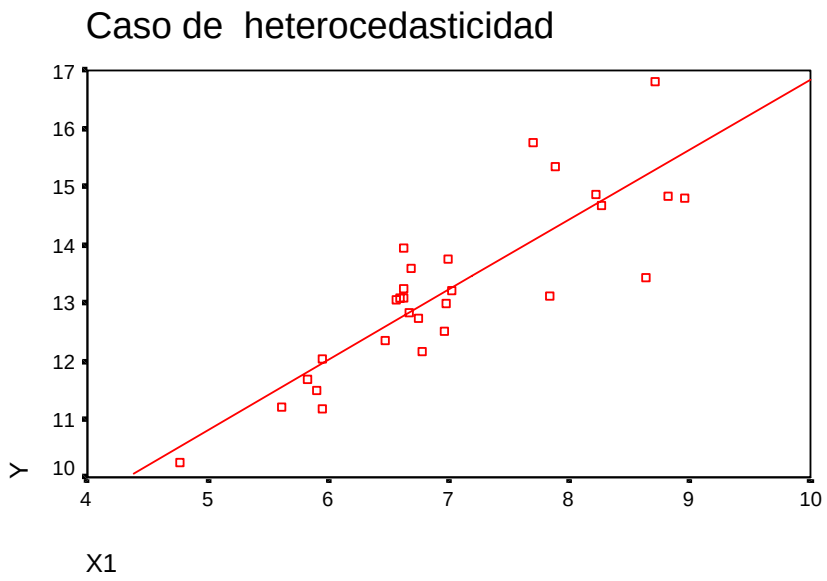
Homocedasticidad y heterocedasticidad.

Homocedasticidad significa que la dispersión alrededor de la recta de regresión es igual para los diversos valores de x. Los valores observados tienden a caer en una zona que podríamos definir por dos paralelas a la recta de regresión.

Heterocedasticidad implica no homocedasticidad. En el caso que veremos, la dispersión en la gráfica con heterocedasticiad, aumentará conforme aumentan los valores de x. Los puntos quedan con límites en forma de cono más estrecho abajo y más abierto arriba.



Los puntos tienen aproximadamente igual dispersión en todo el recorrido.



Nótese que los puntos se van abriendo, como en embudo.

Conforme aumenta x, aumenta la dispersión.

En resumen:

Si Ud. desea correr una regresión lineal entre x e y, antes de correrla, grafique. Ello le permitirá detectar:

Relaciones curvilíneas
Datos erróneos
Casos desviantes
Puntos influyentes
Heterocedasticidad.

Por tanto, **¡primero grafique y luego piense!**