Residuos, Residuos studentizados y valores DFFIT

Su uso en Regresión Lineal Simple y Múltiple.
Florentino Menéndez
Junio 2002 - Cátedra de Mitología de la Investigación IIII
Departamento de Sociología - Universidad de la República

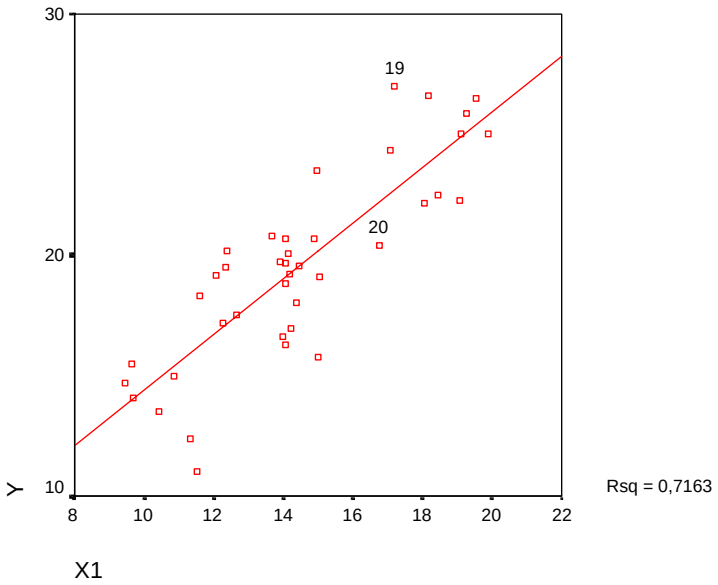
Regresión Lineal Simple

El estudio de residuos es una herramienta formidable en el estudio de las regresiones lineales. Nos sirve para saber si se están cumpliendo las premisas de linealidad de las relaciones, homocedasticidad y normalidad de los residuos. Los residuos studentizados y los valores DFFIT nos ayudan a encontrar casos desviantes y puntos influyentes.

Es más fácil de comprender su lógica en el caso de las regresiones lineales simples. Por ello, se comenzaremos viendo su aplicación a éstas.

Definición de residuo

En el contexto de la regresión lineal, llamamos residuos a las diferencias entre los valores de la variable dependiente observados y los valores que predecimos a partir de nuestra recta de regresión.

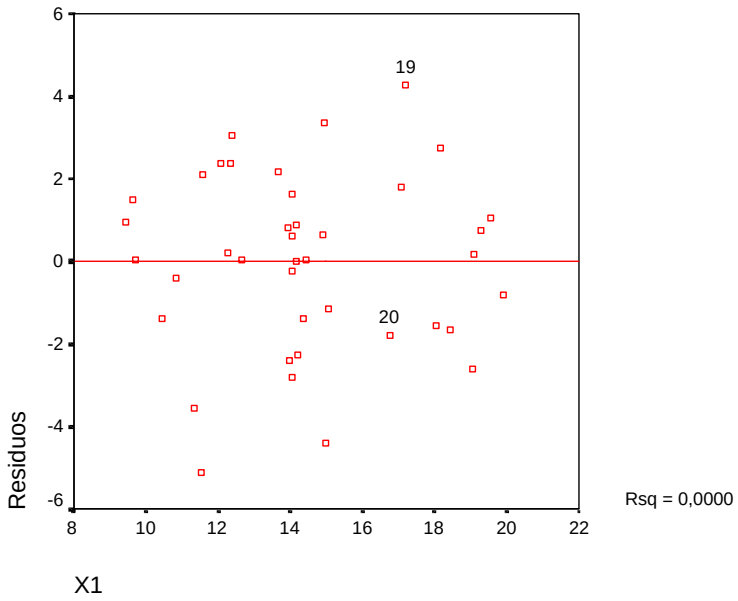


En el diagrama de dispersión mostrado han sido identificados dos casos, el 19 y el 20. El caso 20 tiene un pequeño residuo negativo: su valor en y observado es menor que el valor en y predicho, y la diferencia entre uno y otro no es muy grande en valor absoluto. El caso 19 tiene un residuo positivo: el valor observado es mayor que el predicho.

Gráfica de residuos contra variable independiente.

Los residuos pueden ser graficados contra distintas variables. Es muy habitual hacerlo contra las variables independientes.

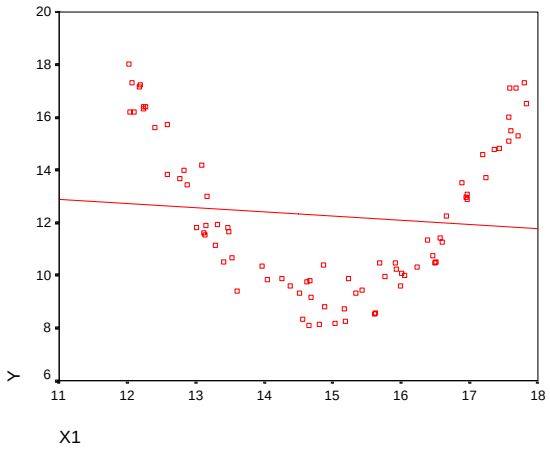
Veamos como luce el gráfico correspondiente a los mismos datos.



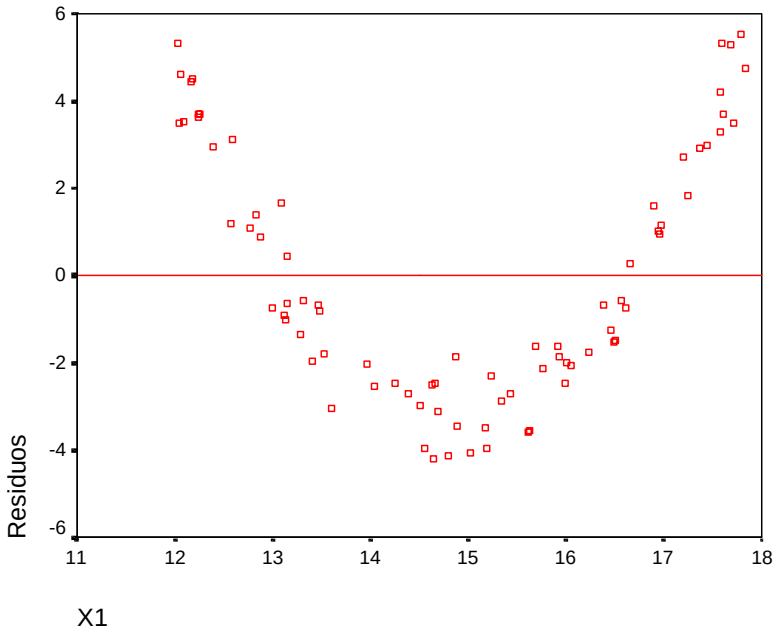
Este gráfico tiene esencialmente la misma información que el scatterplot de x contra y, sólo que la información aparece en una escala mayor, es más fácilmente visible. Nótese que el residuo del caso 20 sigue siendo negativo, el del caso 19 positivo, y que la magnitud absoluta del caso 19 es mayor que la del 20, tal como se podía observar en el primer scatterplot.

¿Qué debemos buscar en los scatterplots de residuos?

Una de las cosas que buscamos es detectar si existe curvilinealidad en las relaciones. Si el diagrama xy luce como el de abajo,



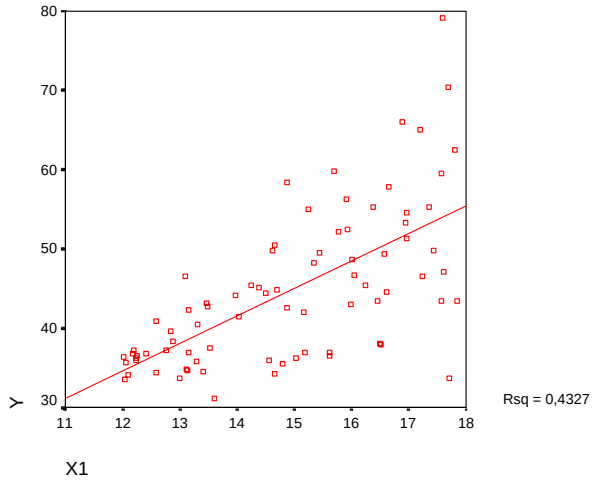
el diagrama de residuos lucirá así:



La curvilinealidad que advertíamos en el primer diagrama de x contra y, la vemos también en el ploteo de residuos.

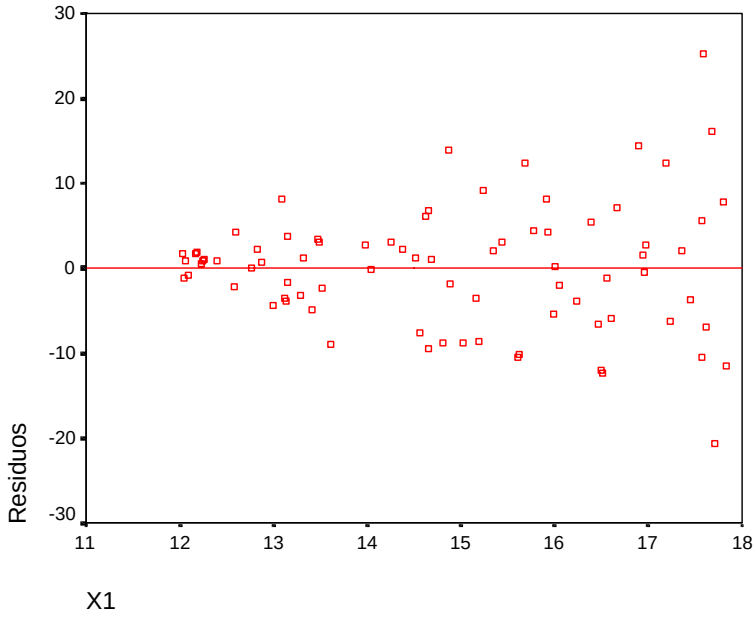
Lo que más nos importa retener de aquí:
<u>El scatterplot de residuos nos permite detectar curvilinealidad.</u>

La investigación de heterocedasticidad.



Notesé que la dispersión de los valores de y aumentan conforme aumenta x. Aquí hay heterocedasticidad. No hay igualdad de varianza para todos los puntos de x.

Veamos ahora el mismo caso en un gráfico de residuos contra x.



Se ve de forma magnificada el mismo fenómeno de heterocedasticidad.

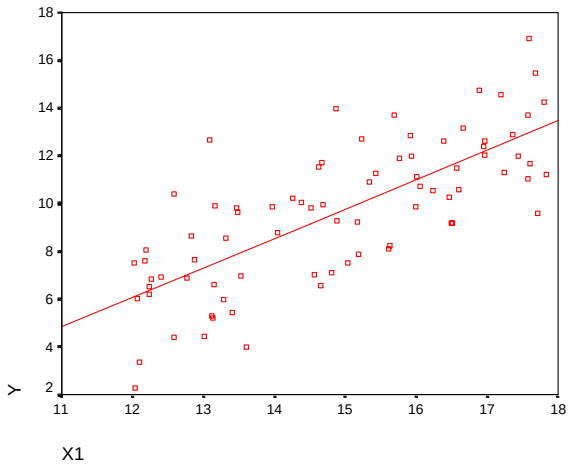
Lo que nos importa de aquí es:
<u>El scatterplot de residuos nos permite detectar heterocedasticidad.</u>

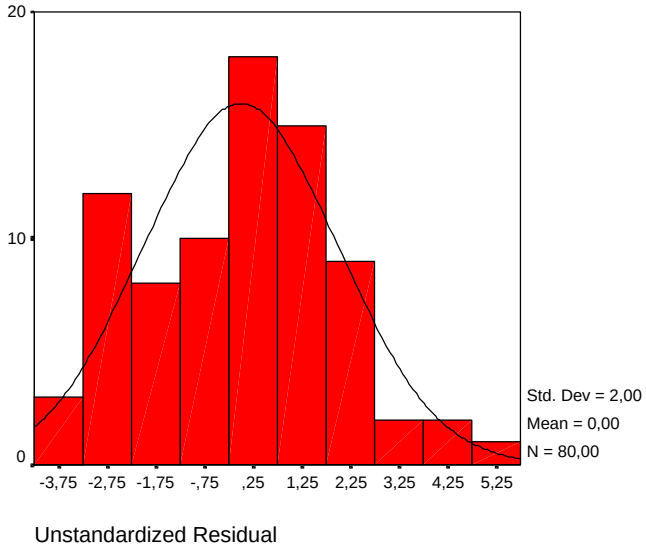
La normalidad de los residuos

Uno de los supuestos del modelo generalmente aceptado es que los errores o residuos, se distribuyen alrededor de la recta de regresión, en forma normal, con una media de cero.

Para verificar ese supuesto de normalidad, podemos recurrir a alguno de los varios gráficos que nos ofrece el SPSS: el histograma, el diagrama de tallos y hojas o un Q-Q plot. Aquí mostraremos solamente uno del ellos: el histograma.

Primero veremos el gráfico xy, y luego mostraremos el histograma de sus residuos.





El histograma con los residuos tiene un aspecto general normal. Por supuesto no reproduce exactamente la curva normal, pero podemos atribuir esas desviaciones, con

razonabilidad, al número no demasiado grande de residuos: 80. Si la muestra fuera mayor, tendríamos derecho a esperar mejor ajuste.

De la última parte se debe recordar que:
Un histograma de residuos nos permite estudiar su normalidad.

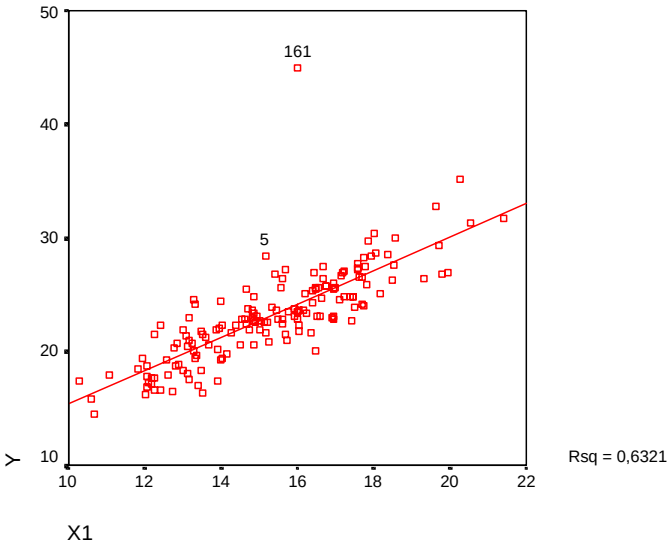
Residuos studentizados

Los residuos studentizados procuran dar una respuesta a la pregunta ¿qué observaciones están muy alejadas del valor previsto? ¿En qué valores es probable que esté teniendo casos desviados o medidas erróneas?

Estos residuos son estandarizados de una manera singular: cada residuo i-ésimo es dividido por la desviación estándar de todos los residuos salvo el i-ésimo.

Studentizacion es una vieja palabra que significa estandarización. Es una forma de estandarizar no muy lejana al cálculo de los valores z. En la estandarización z, a cada valor se le resta la media y se divide por una única desviación estándar del conjunto de residuos. En la studentización de los residuos no es necesario restar la media, ya que la media de los residuos es cero. Se divide por una desviación estándar distinta para cada elemento. La desviación utilizada, se calcula utilizando todos los residuos, salvo el que está siendo considerado.

El elemento con número de identificación 161 está muy alejado de la recta de regresión



en comparación con los demás. El segundo elemento más alejado con residuo positivo es el identificado como 5. Por tanto es de esperar que tengan valores más altos que el resto en su residuo studentizado. Veamos la salida ordenada de mayor a menor, del SPSS.

Case Summaries

	IDENT	Residuos studentizados
	161	8,36871
	5	2,19550
	125	1,91089
	24	1,76059
	17	1,60238
	123	1,41111
	14	1,40458
	153	1,38442
	82	1,33519
	4	1,29569

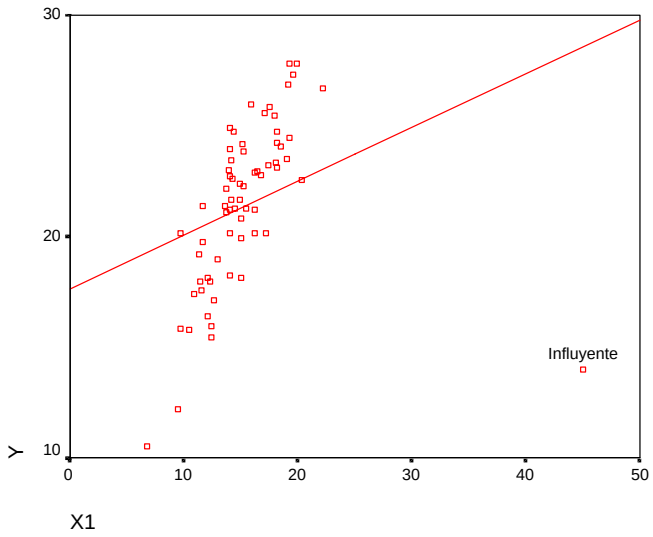
Tal como lo esperábamos, el elemento identificado como 161 tiene un residuo studentizado que se separa notablemente del resto. El elemento número 5, tiene un residuo de 2,2. Es un residuo alto pero no muy alejado del pelotón.

En resumen:
Los residuos studentizados nos permiten localizar los outliers de la relación.

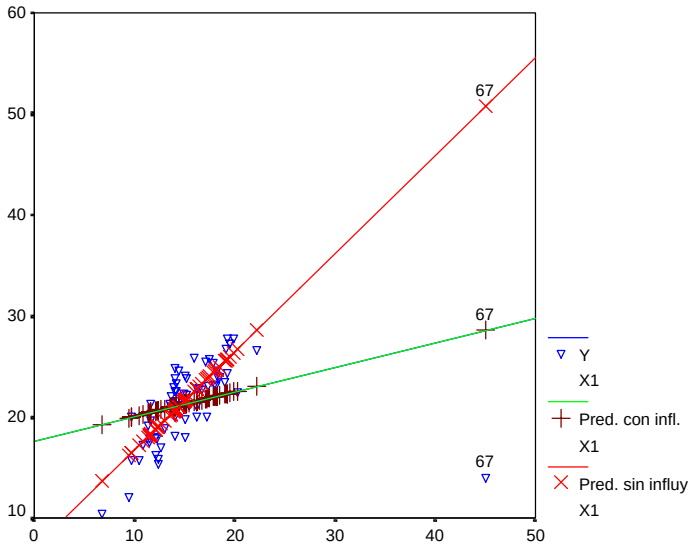
Valores DFFIT

DFFITs quiere decir “difference between fitted values”.

Estos valores nos ayudan a detectar los puntos influyentes en una regresión.



Véase en la gráfica de arriba el punto influyente. Ese punto atrae fuertemente hacia sí la recta de regresión. Si ese punto no estuviera, la recta sería mucho más empinada. En el gráfico que sigue haremos notar que tanto se desvía la línea de regresión por la presencia de dicho punto.



En la gráfica de arriba se plotean:

- los valores observados dibujados como pequeños triángulos que apuntan hacia abajo;

- los valores predichos según la recta de predicción trazada con todos los puntos (marcados con un signo +),

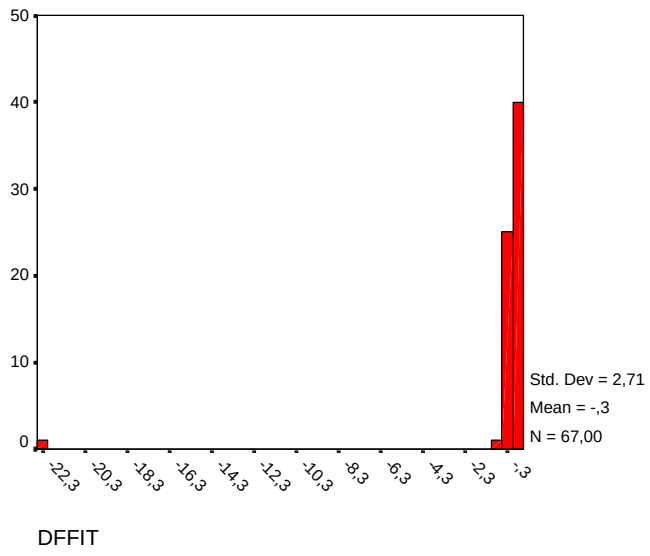
- los valores predichos según la recta de regresión trazada sin el punto influyente (marcados con el signo X).

Véase la gran distancia entre las distintas predicciones generadas para el punto influyente, esto es, para el punto aquí identificado como número 67. La distancia entre ambos valores predichos es el DFFIT para el caso 67.

En nuestro caso particular, el valor predicho por la recta de regresión calculada con todos los puntos para el x1 del valor influyente, es de 28.56. El previsto por la recta de regresión excluido el punto influyente, para el mismo x1, es de 50.71.

Por tanto, el valor DFFIT para dicho punto es de 28,56-50.71= -22.16

Para hallar los puntos influyentes, localizamos los valores de mayor valor absoluto entre los DDFIT.



En el histograma se advierte claramente el carácter de outlier del valor DFFIT= –22.16 No hay ninguno que ni se le acerque en su grado de influencia.

No olvidar que los puntos influyentes deben ser analizados con mucho cuidado. Un solo caso, o unos pocos, pueden llevarnos a sacar conclusiones apresuradas sobre el universo.

Lo más básico a recordar es:
Los valores DFFIT nos permiten detectar fácilmente los puntos influyentes.

Análisis de residuos, residuos studentizados y DFFITs en la regresión lineal múltiple

Habiendo explicado el uso de las herramientas diagnósticas, podremos rápidamente comprender cómo se chequean ordenadamente los supuestos de la regresión lineal múltiple. Seguiremos a David Moore y George McCabe (2000).

Refiriéndose a la etapa final del análisis de regresión lineal múltiple, ellos dicen:

"…debemos siempre examinar los residuos como una ayuda para determinar si el modelo de regresión múltiple es apropiado a los datos. Como hay varias variables explicativas, debermos correr varios gráficos de residuos; es usual graficar los residuos contra los valores predichos de y, y también contra cada una de las variables explicativas."

Siguiendo su criterio, correremos entonces los siguientes gráficos:

- *Residuos contra x1*
- *Residuos contra x2*
- *Residuos contra xn*
- *Residuos contra y predicha.*

Prosiguen Moore y McCabe:

"Busque outliers, observaciones influyentes, evidencia de una relación curva (en oposición a lineal), y todo aquello que sea inusual"

Para ello haremos:

- *Residuos studentizados*
- *Valores DFFITs*

Los outliers de la relación los podemos encontrar como valores elevados dentro de los residuos studentizados. (También podríamos buscarlos como outliers dentro del grupo de residuos: en realidad los residuos studentizados mayores de 3,0 o menores de –3.0 señalan los outliers –ese puede ser un buen punto de corte, aunque podría tomarse otro).

Los valores influyentes serán detectados con ayuda de los valores DFFITs. Un caso para ser fuertemente influyente, debe tener un valor apartado del común, ya sea positivo o negativo.

La búsqueda de evidencia de relaciones curvas se hará mediante el análisis de las gráficas de residuos. Nubes de puntos de claro contenido curvilíneo nos exigirán que modifiquemos nuestro modelo.

Moore y McCabe dicen finalmente:

“Si las desviaciones en el modelo están normalmente distribuidas, los residuos deberán estar normalmente distribuidos.”

Para ello graficaremos:

- Histograma con los residuos, para chequear su normalidad.

También son aptos para chequear normalidad un Q-Q plot o un diagrama de tallos y hojas, pero por simplicidad proseguiremos con el histograma.

Si hemos dado los pasos arriba indicados, hemos hecho un buen chequeo de la adecuación del modelo de regresión lineal múltiple a sus premisas básicas.

En resumen:

Remate la adecuación de sus variables a un modelo de regresión lineal múltiple efectuando:

- *Residuos contra x1*
- *Residuos contra x2*
- *Residuos contra xn*
- Residuos contra y predicha.

- *Residuos studentizados*
- Valores DFFITs

- *Histograma de residuos.*

Si no se advierten anormalidades severas, sus variables pueden ser modeladas con una regresión lineal múltiple.

Bibliografía citada:

Moore, D. y G. McCabe (2000) Introduction to the Practice of Statistics, 3ra. Edición, New York, W. H. Freeman and Company: 724.