INTRODUCCIÓN AL ANÁLISIS CUANTITATIVO EN SOCIOLOGÍA:

Diagnóstico de los residuos en el modelo de regresión lineal

Tabaré Fernández Aguerre
COLMEX - CES

Tercera versión - junio 2004

Diagnóstico de los residuos o errores (ϵ)

✎ Los objetivos de este capítulo de la guía "introducción al análisis cuantitativo en sociología: el modelo de regresión" son dos. Primero, exponer qué significado tiene el término de error incluido en los modelos, utilizando para ello un muy pequeño desarrollo estadístico, y por otro, inscribiendo ese sentido en el contexto de las discusiones epistemológicas contemporáneas sobre el determinismo. El segundo objetivo del capítulo es presentar algunas técnicas gráficas y estadísticas que permiten realizar un diagnóstico de los residuos a los efectos de identificar casos desviados y puntos influyentes en la regresión. Un estudio más avanzado de los residuos se completará en el análisis de los supuestos de homocedasticidad y de ausencia de auto-correlación.

✎ Es necesario partir enfatizando que nada es más crítico para la comprensión sustantiva de los supuestos de un modelo de regresión que la valoración del significado del término de error, así como también considerar las pruebas formales que se pueden realizar.

✎ Denotado por ε_i ha recibido distintos nombres e interpretaciones: "término de error", "residuo", "perturbación".

1. El sentido "teórico" del error

✎ Más importante que estos cambios de nombre son los cambios conceptuales ligados a los cambios epistemológicos sobre el papel de este término en una ecuación. A grandes rasgos, hay dos interpretaciones:

- La más antigua y también más extendida es la interpretación *determinista.* Esta afirma que ε representa todas las otras causas verdaderas ("Z") de Y que no han sido medidas porque la ciencia *no las conoce aún* o no *las ha podido medir*. En la medida en que la investigación avance, esas otras causas se irán progresivamente incorporando al modelo hasta que el fenómeno Y quede completamente determinado. Mientras tanto, los investigadores sólo podrán ajustar modelos muestrales que son incompletos y parciales;

modelos que entre sí pueden ser *distintos e igualmente válidos*. Pero que será progresivamente convergentes hacia el modelo poblacional "verdadero"y "real" .

- La explicación contemporánea interpreta *probabilística y constructivamente* al término ε. "Los trabajos muestran que lo que se presenta como dato de en la percepción tiene ya el carácter de construcción, en tanto que la parte del dato que procede del objeto está siempre incorporada a esquemas más o menos organizados que testimonian la actividad del sujeto" (Piaget 1958:*Etudes de Epistémologie Génétique V: La lecture de l'expérience*. PUF). Nuestros "datos" están cargados de teorías. Estas nos proporcionan esquemas de selección que re-organizan los "observables" estableciendo relaciones conceptuales entre ellos. Estos procesos cognitivos (constructivos) dependen por tanto de las teorías, de los observables y de los marcos epistémicos proporcionados por cada época histórica (García 2000: cap.6). Por tanto, la variación aleatoria existe tanto en la naturaleza como en el mundo social y no podrá eliminarse. [...] Un investigador puede dividir el mundo en componentes aparentemente sistemáticos y no sistemáticos y mejorar con frecuencia sus predicciones, pero nada de lo que haga para analizar sus datos podrá reducir de manera significativa el grado de variación no sistemática que existe en el mundo empírico. (King, Keohane & Verba 2000:7073)

✎ Desde la primera perspectiva, sería posible determinar perfectamente el valor de cualquier *i-ésimo* caso, a través de *K* variables X conocidas e incluidas en el modelo y de *L* variables Z que no son conocidas o que no han sido incluidas en el modelo. Por tanto, el término de error es una sumatoria de todas las variables Z desconocidas en el modelo:

$$\varepsilon_i = \sum c_l Z_{li}$$

✎ Por lo que una vez que se pudieran conocer, medir e incluir en el modelo todas las variables explicativas, se tendría una ecuación determinista del siguiente tipo:

$$y_i = b_0 + b_1 X_1 + b_2 X_2 + \ldots\ldots + b_k X_{ki} + c_1 Z_1 + c_2 Z_2 + c_3 Z_3 + \ldots + c_l Z_{li}$$

✎ En una perspectiva probabilística, el término ε representa el componente aleatorio del mundo "r", "ruido" o "caos sensorial no ordenado" y por ende

presente inevitablemente en todas nuestras observaciones:

$$\varepsilon_i = R$$

$$y_i = b_0 + b_1 X_1 + b_2 X_2 + ... + b_j X_{ij} + r$$

✎ Berry (1993:10-11) propone una formulación ecléctica de ambas perspectivas, rescatando la idea de variables no incluidas en el modelo y del componente aleatorio.

$$\varepsilon_i = \sum c_l Z_{li} + r$$

$$y_i = b_0 + b_1 X_1 + b_2 X_2 + ... + b_j X_{ij} + \varepsilon_i$$

- La distinción concreta de uno y otro componente en el error es una consideración particular de cada tema de investigación y de la extensión que haya tenido la revisión teórica antecedente.

- Será requisito por tanto, explicitar las razones por las que se han excluido variables conocidas de un modelo de regresión. Estas podrán ser muy diversas: se esperan efectos reducidos; puede sobredeterminarse el modelo; no todas las variables son de igual interés teórico, etc.

2. El análisis empírico de los errores o residuos del modelo

✎ Recordemos que el error para la *i-ésima* unidad se define como:

$$e_i = Y_i - \hat{Y}_i$$

✎ El método de los mínimos cuadrados ordinarios asegura que la sumatoria de los cuadrados del error es mínima. Para lograr esto, la recta (para regresiones simples) o el hiperplano de regresión (para multivariados) se ubica equidistante de las observaciones.

✎ Puede ser que en un análisis preliminar se identifiquen algunas observaciones cuyos valores sea en la variable dependiente o en alguna de las variables

independientes están fuera de la tendencia que caracterizan a esa distribución.

- Estos casos, denominados "out - layers", se identifican en primer lugar en función de la distribución de una o varias variables. Son casos fuera para una *tendencia determinada.*

- Merecen una atención especial porque pueden tener una influencia indebida en el ajuste de la regresión, dado que "fuerzan" a que el ajuste los tome en cuenta.

- En el caso de que se pruebe a través de gráficos o de estadísticos que en el modelo ajustado esto está sucediendo, tales observaciones o casos ya no son simplemente out-layers sino que deben tratarse como "**puntos influyentes**".
- Un punto será definido como influyente si al ser eliminado del análisis, genera cambios drásticos en las estimaciones que se realizan con el modelo (bondad de ajuste y parámetros).

3.Estadísticos para "out-layers"

✎ El análisis formal de los errores comienza con un estudio de los valores de los **residuos estandarizados** para cada uno de las *i-ésimas* observaciones.

- Designaremos a este estadístico como z_resid_i que se obtiene si se utiliza el SPSS ® en: analyze/regression/save/residuals/standarized . El paquete lo salvará como una nueva variable en el archivo de datos.

- Aquella observación que tiene un valor alto (en términos absolutos) de residuos estandarizados, se denominan **"out-layers" en el espacio de la "variable dependiente".**

✎ El investigador en función del en estudio definirá qué magnitud considerará como "fuera de la tendencia".

- Dado que la nueva variable, z_resid_i, tiene por construcción una media igual a cero y un desvío estándar igual a uno, el investigador determinará

el valor crítico, z_resid * , en términos de desvíos estándares.

- La convención más general es tomar como valor crítico en términos absolutos los 2 desvíos estándares. Esto es:

$$z\_resid^{*} = /2\ sd/$$

- Sin embargo, el investigador deberá determinar según el problema de investigación, qué definirá como valor crítico.

- La regla de decisión es que si el valor del residuo estandarizado observado es mayor en términos absolutos que el valor crítico, entonces el caso será definido como "out-layer".

✎ Debe puntualizarse que el análisis de los residuos permite determinar out-layers pero no permite conocer si ese punto ha ejercido una influencia desmedida en la estimación de los parámetros del modelo de regresión.

- Esta razón se deriva del método de MCO: si el punto es un punto influyente, ha forzado el ajuste "atrayendo" hacia sí el hiperplano de la regresión. En consecuencia, los residuos para este punto pueden ser iguales a cero.

✎ Los out-layers pueden deberse a valores fuera de tendencia en las variables independientes. Como se expresó antes, los residuos estandarizados no sirven para el propósito de detectar estos casos. Uno estadístico utilizado para este fin es el **estadístico de Leverage**

- en términos simples, el estadístico se define como una medida ponderada de las dispersiones de los valores observados en X para los promedios de cada una de las X variables (Chaterjee et al 2000: 88-90).

- En consecuencia, aquellas observaciones que tienen valores fuera de la tendencia en el espacio de los K - predictores incluidos en el modelo de regresión, tendrán valores altos en el estadístico de Leverage. En tal caso, serán designados como observaciones con alto valor de Leverage.

- El estadístico de Leverage tiene un recorrido cerrado entre 0, que es su valor mínimo, y 1 que es el valor máximo que puede tomar.

✎ La regla de decisión para identificar casos con alto valor de Leverage es la siguiente:

- Obténgase el valor de Leverage para cada observación. En el caso del SPSS ®, esto se obtiene en: analyze/regression/save/leverage.

- Calcúlese el valor promedio (p con una línea horizontal) del estadístico de Leverage para la base de datos con la que está trabajando, donde "p" es el número de regresores en el modelo y N es el número de casos total en la muestra:

$$\overline{p} = (p+1)/N$$

- Calcúlese el valor crítico del Leverage, tal que:

$$p^* = 2\overline{p}$$
$$p^* = 2(p+1)/N$$

- Identifiquese como observaciones con alto Leverage a todas aquellas cuyo valor de leverage sea superior al valor crítico.

- Estas observaciones deben ser tomadas como primeros candidatos en el análisis de los puntos influyentes.

4. Estadísticos para "puntos influyentes"

✎ la influencia de una observacion en el ajuste del modelo de regresión se mide por los efectos que produce en el ajuste cuando aquella observación es excluida del análisis. El borrado de cada observación identificada como potencial punto influyente, debe realizarse de uno a la vez.

✎ Existen diversos estadísticos para el análisis de los puntos influyentes:

- La distancia de Cook
- La medida de influencia de Hadi
- La medida DFFIT propuesta por Welsch & Kuh

✎ El DFITS o DFFIT (difference in fit) es una medida para identificar puntos influyentes propuesta por Welsch & Kuh fue propuesta en 1977. Es una medida escalada de la diferencia entre el valor estimado para la *i-ésima* observación utilizando toda la base de datos y el valor estimado para la *i-ésima* observación cuando ha sido eliminada de la base.

$$DFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sigma^2 \sqrt{p_{ii}}}$$

- En el SPSS ® , se hallan dos estadísticos: el DFFIT y el DFFIT estandarizado. Se aconseja utilizar este último, el cual se grabará como una nueva variable en el archivo de datos.
- Para el paquete SPSS ®, el valor crítico del DFFIT se establece como dos veces la raíz cuadrada del cociente de (p+1) / N, donde p es el número de parámetros incluidos en el modelo (p variables independientes más la constante) y N es el número de casos. En el caso de que la muestra sea pequeña, es conveniente utilizar como denominador (N-p-1) en lugar de N, tal como lo aconseja Chaterjee et al (2000: 105)

$$DFFIT^* = 2\sqrt{(p+1)/n}$$

- La regla de decisión establece que cuando el valor observado supera el valor crítico, la observación deberá ser considerada como punto influyente.