Elkin Castaño Vélez
Centro de Investigaciones Económicas -CIE-
Universidad de Antioquia

Robustez estadística

Lecturas de Economía. No. 24. Medellín, septiembre-diciembre de 1987. pp. 85-99

• Resumen. El uso de modelos paramétricos estocásticos exactos tales como el normal, log-normal, exponencial, poisson, gama, etc. está hoy profundamente arraigado en la práctica estadística. La razón es que ellos permiten la representación aproximada de un conjunto de datos que puede ser facilmente descrita e interpretada. Sin embargo, es bien conocido que el mundo real no se comporta tan bien como lo describen estos modelos. Recientemente surge una técnica estadística la cual emplea también los modelos paramétricos pero la inferencia es realizada para un entorno del modelo asumido. Es decir, aunque emplea modelos paramétricos, los procedimientos que construye no dependen fundamentalmente de las hipótesis inherentes a ellos.

• Abstract. The use of precise stochastic parameter models, such as the normal, log-normal, exponential, poisson and gamma models, is nowadays deeply entrenched in statistical practice. The cause of this trend is their ability to approximately represent a series of data, that can be easily described and interpreted. Nevertheless, it is well known that the real world doesn't behave as these models describe it. Recently, a statistical technique, employing parametrical models, appeared. This method tests part of the assumed model that is, althorugh it employs parametrical models the procedure, it builds, doesn't fundamentally depend on the included hypothesis.

INTRODUCCION

La estadística es el arte y la ciencia de extraer información útil desde un conjunto de datos empíricos. Una manera efectiva de extraer dicha información es usar modelos estocásticos paramétricos, es decir modelos cuya forma exacta es conocida. Esta estrategia es referida con el nombre de aproximación "clásica".

El uso de modelos paramétricos estocásticos rigurosos tales como el normal, log-normal, exponencial, Poisson, etc., está hoy profundamente arraigado en la práctica estadística. La razón es que ellos permiten la descripción aproximada de un conjunto de datos por medio de un modelo estocástico fácil de ser descrito e interpretado y del cual se pueden generar las observaciones reales y otras posibles y/o futuras. Sin embargo es bien conocido que el mundo real no se comporta tan bien como lo describen estos modelos.

Ahora bien, debido a la belleza y elegancia de la teoría de los modelos paramétricos es fácil caer en la tentación de olvidar que ellos son solamente aproximaciones a la realidad. Con frecuencia se piensa que la realidad no es más que una aproximación a algunos modelos paramétricos y que las posibles discrepancias entre aquélla y éstos es debida a errores de observación. Así, por ejemplo, durante gran parte del siglo pasado todo el mundo creía

en la distribución normal; los matemáticos porque pensaban que era un hecho experimental y los experimentadores porque creían que era una consecuencia de un teorema de la matemática. Sin embargo, el Teorema Central del Límite solamente habla de un cierto límite imaginario a ser alcanzado bajo ciertas suposiciones. No dice nada de qué tan lejos se halla de ese límite en una situación concreta ni de cuál debe ser el tamaño de la muestra para que el límite sea válido, ni tampoco dice nada sobre si la situación real que estamos estudiando satisface las suposiciones. Lo que un chequeo empírico puede probar en el mejor de los casos es que esa situación real está en un entorno de la distribución señalada por el modelo, pero que jamás que coincide exactamente con ella.

Una esperanza tácita al ignorar los desvíos de los modelos ideales era la que los procedimientos estadísticos óptimos bajo el modelo estricto tendrían que seguir siendo aproximaciones óptimas bajo modelos aproximados. Pero la experiencia práctica mostró lo completamente errado de tal esperanza. Frecuentemente desvíos moderados conllevan efectos catastróficos mucho mayores que los anticipados por la mayoría de los estadísticos. Debido a los inconvenientes con el uso de modelos paramétricos, surgió la estadística no-paramétrica y algunos de sus métodos, tal es el test de Wilcoxon, han llegado a ser muy populares en sus aplicaciones.

El principio básico de la estadística no-paramétrica es el de hacer tan pocas hipótesis sobre los datos como sea posible y aún así poder responder a problemas específicos (tales como: ¿existe diferencia entre antes y después?").

Mientras algunos problemas de esta clase encontraron soluciones satisfactorias los modelos paramétricos siguieron jugando un papel sobresaliente debido a su capacidad de describir en forma más completa la información contenida en un conjunto de datos y debido a que son útiles en un amplio rango de aplicaciones, especialmente en situaciones complejas.

Las cualidades de las dos aproximaciones anteriores son la materia prima que usa la teoría de la estadística robusta. Ella emplea modelos paramétricos sobre los cuales construye procedimientos que no dependen fundamentalmente de las hipótesis inherentes a ellos, es decir, emplea modelos paramétricos pero la inferencia es realizada para un entorno del modelo asumido. En palabras de Hampel. et. al. (1986) "Estadística robusta, [. . .] es la estadística de los modelos aproximados".

Este artículo está dividido en cuatro secciones. En la sección I se presentan los desvíos más frecuentes de los modelos paramétricos; en la sección II se da un ejemplo en el cual se ilustra la pérdida de eficiencia de los procedimientos clásicos frente a desvíos en un modelo paramétrico básico. La sección III discute la teoría de la robustez frente a la teoría clásica y la no-paramétrica, así como también algunos conceptos de optimalidad en la estimación introducida por la teoría de la robustez. Por último, en la sección IV se muestran algunas ideas de robustez en la estimación empleando la "función de influencia", la cual es una de las herramientas más potentes de la teoría de la robustez.

I. DESVIOS MAS FRECUENTES DE LOS MODELOS PARAMETRICOS

Ahora bien, ¿cuáles son los desvíos más comúnmente presentes?; ¿cuál es el grado en el que ellos afectan el comportamiento de los procedimientos clásicamente óptimos?

Podemos distinguir cuatro causas principales para los desvíos de los modelos paramétricos estrictos:

1. La recurrencia de errores "burdos".

2. El redondeo y el agrupamiento de datos.

3. El carácter solamente aproximado del modelo que puede haber sido concebido nada más que por la aplicación de un resultado matemático (El caso del Teorema Central del Límite).

4. Dejando a un lado las suposiciones en torno a la verdadera distribución subyacente de los datos, está el problema de la independencia o cualquier otra estructura de correlación que se asuma entre los datos. Puede ser que esa estructura sea satisfecha sólo aproximadamente.

Con respecto a 1, las causas más comunes de la ocurrencia de errores "burdos" son las que provienen de errores en la copia, grabación o cálculos de los datos.

Con respecto a 2, la influencia de los errores que provienen de redondeo y/o agrupamiento de los datos puede ser minimizada pero nunca olvidada. Existen situaciones en las que juega un papel de importancia. Por ejemplo, en el estudio de magnitudes que van a ser determinadas localmente tales como en el caso de estimación de funciones de densidad, en la estimación de percentiles, etc.

Con respecto a 3, grandes conjuntos de mediciones de alta calidad, de los cuales se ha sabido que no tienen errores burdos, frecuentemente presentan pequeños pero no imperceptibles desvíos del modelo normal. Esto ya se conocía pero fue ignorado, consciente o inconscientemente, luego de la popularidad que ganó el método de los mínimos cuadrados. También debe ser tenido en cuenta que en modelos más complejos, ciertas características asumidas como la linealidad y/o la aditividad de los efectos, son buenas aproximaciones pero jamás son válidas exactamente.

Con respecto a 4, podemos decir que es muy poco lo que se conoce sobre ese tema, aunque es uno de los problemas más frecuentes y relevantes.

En cuanto a los efectos de los desvíos moderados de la normalidad o de otros modelos paramétricos, tales como los que ocurren en los datos de alta calidad, no causan efectos tan catastróficos sobre las técnicas clásicas como los producidos por los errores burdos. Desde este punto de vista podríamos concluir que el uso de métodos robustos en esta situación no es absolutamente necesario. Sin embargo, la inevitable pérdida de eficiencia de los métodos clásicos en esos casos puede ser mucho más de lo que ingenuamente se supone con frecuencia.

El método de máxima verosimilitud de Fisher es incierto en general a menos que el modelo sea conocido exactamente o que el mismo sea adecuado para describir esos desvíos moderados; por ejemplo, postulando una distribución subyacente a los datos que tenga colas más pesadas que la de la distribución normal. A pesar de todo, este método puede ser robustificado de una forma que puede ser confiable sobre una amplia gama de modelos potencialmente útiles con apenas pequeñas pérdidas de eficiencia cuando se asume el modelo inicial.

II. EL PELIGRO DE LOS METODOS TRADICIONALES

Para ver algunos ejemplos importantes sobre la pérdida de eficiencia de los procedimientos clásicos frente a desvíos de los modelos ver, por ejemplo, a Tukey (1960). El siguiente ejemplo es debido a Huber (1981), página 2:

Suponga una muestra aleatoria grande de tamaño n en la que están mezcladas observaciones X_i buenas y malas. Suponga, además, que cada observación tiene probabilidad $1-p$ de ser buena y p de ser mala. En el primer caso X_i es de una distribución $N(u, var)$, en el último es de una distribución $N(u, 9\,var)$. En otras palabras todas las observaciones tienen la misma media, pero los errores están incrementados en un factor de 3.

Equivalentemente, podríamos decir que las X_i son independientes, idénticamente distribuidas con distribución común

$$\bar{F}(x) = (1-p)F + pG \qquad (1)$$

donde F es la $N(u, var)$ y G es la $N(u, 9var)$.

Para estimar la dispersión, dos medidas reconocidas son la desviación absoluta media

$$dn = Suma(abs(X_i - \bar{X}))/n$$

y la desviación media cuadrática

$$Sn = (Suma\,(X_i - \bar{X})^2/n)^{1/2}$$

Entre Eddington (1914) y Fisher (1920) existió una disputa sobre los méritos relativos de dn y Sn. Eddington recomendaba el uso del primero mientras que Fisher señalaba que para observaciones normales Sn es alrededor del 12% más eficiente que dn (Tabla 1).

Por supuesto, los dos estadísticos miden diferentes características del error de la distribución. Por ejemplo si los errores son exactamente normales, Sn converge a desviación estándard, mientras que dn converge a la raíz cuadrada de $2/pi$ o aproximadamente a 0.8 veces la desviación estándard.

Para comparar los comportamientos de estas dos medidas de dispersión podemos usar la eficiencia relativa asintótica (ARE) de *dn* con relación a *Sn*, definida como sigue:

$$ARE(p) = \lim \frac{var(Sn)/(E(Sn))^2}{var(dn)/(E(dn))^2}$$

$$= \frac{\dfrac{3(1+80p)}{(1+8p)^2} - 1}{4\,\dfrac{P_i(1+8p)}{2(1+2p)^2} - 1}$$

donde p es la probabilidad de que la observación X_i sea mala.

Los resultados para varios valores de p están resumidos en la tabla siguiente:

Tabla 1 **El ejemplo de Huber**

p	*ARE(p)*
0	0.876
0.001	0.948
0.002	1.016
0.005	1.198
0.010	1.439
0.020	1.752
0.050	2.035
0.100	1.903
0.150	1.689
0.250	1.371
0.500	1.017
1.000	0.876

El resultado es inquietante: si $p = 0$ (es decir, no existe contaminación) la tabla muestra que *Sn* es más eficiente que *dn* (alrededor del 12%). Pero solamente 2 observaciones malas en 1000 son suficientes para quitar el 12% de ventaja de *Sn* sobre *dn* pasando a ser ese último un poco más eficiente. ARE *(p)* obtiene un valor máximo mayor que 2 alrededor de $p = 0.05$, es decir con el 5% de las observaciones malas.

Este hecho es particularmente desafortunado puesto que en ciencias físicas las muestras típicas "buenas" parecen ser bien modeladas por una ley de errores de la forma (1) con p en el rango entre 0.01 y 0.1. Así que las desviaciones de la forma anterior restan importancia a la teoría tradicional asintótica de optimalidad: en la práctica, ante esta situación, deberíamos ciertamente preferir *dn* a *Sn* puesto que es mejor para todo p entre 0.002 y 0.5.

III. ROBUSTEZ

Ahora bien, ¿qué es, entonces, la estadística robusta?

Realmente es una fusión de las virtudes de las dos aproximaciones anteriores. Los modelos paramétricos son usados como vehículo de información, y se implementan procedimientos que no dependen en forma crítica de las hipótesis inherentes a estos modelos. Así la estadística robusta en un sentido amplio y no técnico tiene en cuenta el hecho de que muchas de las suposiciones que son asumidas comúnmente en estadística son solamente aproximaciones a la realidad. En esta situación, el problema de la teoría clásica paramétrica consiste en que ella deduce procedimientos que son optimales solamente bajo modelos paramétricos exactos, pero no dicen nada con respecto al comportamiento de tales técnicas cuando los modelos son válidos sólo aproximadamente. La estadística no-paramétrica tampoco se ocupa específicamente de este problema. Puede ocurrir para ambas teorías que ligeras violaciones de las hipótesis produzcan un comportamiento bastante pobre de procedimientos clásicamente optimales.

A pesar de que el concepto de robustez es antiguo (ideas germinales se encuentran en Fisher (1922), sólo en las últimas décadas se han hecho esfuerzos conducentes a su formalización en el intento de construir una "Teoría de Robustez". Las causas probables se encuentran tal vez en el poco desarrollo de teorías matemáticas adecuadas, o a la difícil tarea de realizar tediosos

cálculos sin la ayuda de los computadores modernos o quizá a que la comunidad estadística era poco consciente del comportamiento catastrófico de las técnicas clásicas en situaciones aproximadamente "normales".

En el presente existe una gran variedad de formas de abordar la robustez. Sinembargo, es importante concentrarnos sobre las formas que intentan capturar las características esenciales de los problemas de la robustez en la vida real, dentro de un contexto matemático (Huber, 1964; Hampel, 1968). Dentro de este enfoque se han desarrollado nuevos conceptos estadísticos que describen el comportamiento de los procedimientos no sólo sobre el modelo paramétrico estricto, sino en una "vecindad de él". Estos conceptos conducen a nuevas cuestiones de optimalidad y plantean nuevas tareas a los estadísticos matemáticos. La teoría proporciona también elementos para juzgar y comparar procedimientos estadísticos sobre diferentes aspectos de la robustez.

En particular los analistas de datos encuentran en esta teoría un esquema formal para preguntas tales como:

- ¿Existen en la muestra subconjuntos de datos que dicen cosas diferentes a las que dicen la totalidad?
- ¿Qué sugiere la mayoría de los datos?
- ¿Cuál es la influencia en el resultado final de los diferentes subconjuntos de datos sacados de la muestra total?
- ¿Cuáles son los datos de mayor importancia en la elección del modelo y/o para los resultados finales?
- ¿Cuáles datos deben ser analizados con mayor cuidado?
- ¿Qué proporción de datos "malos" puede tolerar un procedimiento usado?
- ¿Cuáles son los procedimientos más seguros y eficientes?

Veamos algunos nuevos conceptos de optimalidad en estimación.

La eficiencia, un concepto central en la estadística clásica, es también empleado en la estadística robusta en las ideas de *protección* y *premio* (Anscombe, 1960a) y de estimación robusta *minimax* (Huber, 1981). Las ideas de protección y premio tienen como base un modelo paramétrico básico F y un estimador $\tilde{n}$ optimal del parámetro de interés, un modelo alternativo F' el cual es un entorno del modelo básico y un estimador T rival del estimador optimal el cual será más robusto que el optimal si cumple con dos condiciones de eficiencia:

(i) $\dfrac{var(T/F')}{var(\tilde{n}/F')}$ debe ser pequeña y,

(ii) $\dfrac{var(T/F)}{var(\tilde{n}/F)}$ debe estar cerca a uno.

Estas dos condiciones exigen al estimador robusto que sea mejor que el estimador optimal bajo el modelo alternativo y que sea casi tan bueno como el optimal bajo el modelo básico. Cantidades equivalentes a las dos medidas anteriores de eficiencia son las definidas por Anscombe como

$$\text{Protección} = \frac{var(\tilde{n}/F') - var(T/F')}{var(\tilde{n}/F')} \quad \text{y,}$$

$$\text{Premio} = \frac{var(T/F) - var(\tilde{n}/F)}{var(\tilde{n}/F)}$$

Algunas ideas de Anscombe son presentadas en Barnett y Lewis (1984). En el desarrollo del estimador robusto *minimax* de Huber se considera ya no un modelo alternativo sino una familia de ellos y se define el estimador robusto *minimax* aquel cuya varianza máxima sobre la familia de distribuciones es tan pequeña como sea posible.

Además de estos conceptos, la estadística robusta ha desarrollado otras

ideas de optimalidad, basadas en el efecto de la contaminación (es decir, de la presencia en la muestra de observaciones de otra distribución diferente a la asumida) sobre un estimador dado. Estos nuevos conceptos se relacionan con la respuesta a preguntas como:

- ¿Es el efecto proporcional al número de contaminantes presentes en la muestra?

- ¿Suponiendo que hay un sólo contaminante, cómo puede relacionarse su magnitud con el efecto que produce sobre el estimador?

- ¿Cuál es el peor efecto que un contaminante puede tener?. En particular , ¿está acotado o no?

- ¿Cuál es la cantidad máxima de contaminantes que un estimador puede absorber antes de llegar a ser no confiable?

Aspectos como los anteriores se basan en una potente herramienta de la estadística robusta llamada la Función de Influencia o Curva de Influencia de un estimador debida a Hampel (1974).

La idea puede resumirse con algunos ejemplos sencillos. Suponga que $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una población con función de distribución F la cual depende de un parámetro desconocido y que $T = T(X_1, X_2, \ldots, X_n)$ es un estimador del parámetro. Sea X_c un contaminante fijo (es decir que la distribución de X_c es tal que $P(X = X_c) = 1$). Queremos medir el efecto sobre T al agregar X_c a la muestra de n observaciones buenas.

Suponga, por ejemplo que T es la media muestral $\overline{X}$. Entonces el efecto de X_c es el de cambiar el estimador en una cantidad

$$\overline{X}_c - \overline{X} = (n\overline{X} + X_c)/(n+1) - \overline{X} = (X_c - \overline{X})/(n+1)$$

donde $\overline{X}_c$ es la media muestral sobre las $n+1$ observaciones.

El efecto estandarizado por la cantidad de contaminación es

$$(n+1)(\overline{X}_c - \overline{X}) = X_c - \overline{X}$$

el cual excede cualquier cota cuando Xc es suficientemente grande.

Análogamente puede verse que si T es S^2 entonces,

$$(n+1)(Sc^2 - S^2) = (Xc - X)^2 - (n+1)S^2/n$$

Vemos que el efecto estandarizado excede cualquier cota cuando Xc es suficientemente grande. En estos ejemplos observamos entonces que el efecto de un contaminante sobre dichos estimadores es impredecible.

Los efectos por unidad de contaminación, como los descritos anteriormente, son llamados las Funciones de Influencia de Muestras Finitas (o Curvas de Influencia de Muestra Finita) y se acostumbra denotarlos como $IC_{T,\ F;\ n}(Xc)$ donde T denota el estimador usado, F la distribución, n el tamaño muestral y Xc el contaminante.

Ahora, $Lim\ IC_{T,\ F;\ n}(Xc) = IC_{T,\ F}(Xc)$ es la función de influencia asintótica o simplemente función de influencia y es una herramienta realmente útil.

En los ejemplos anteriores puede probarse que:

$$(i)\ IC_{T,\ F}(Xc) = Xc - u$$
$$(ii) IC_{T,\ F}(Xc) = (Xc - u)^2 - \text{varianza}$$

Observamos que se mantiene la misma información sobre el efecto de un contaminante que en el caso de la versión para muestras finitas.

Consideremos ahora a T como la mediana muestral Md. Puede probarse que

$IC_{T,\ F}(Xc) = sign(Xc - m)/2f(m)$ donde m es tal que $n = 2m + 1$ y f es la función de densidad de probabilidad (f.d.p.) de X.

De aquí vemos una diferencia cualitativa esencial entre la mediana muestral y la media muestral: el efecto de un contaminante sobre la mediana muestral es acotado mientras que en la media muestral no.

IV. ALGUNAS MEDIDAS DE ROBUSTEZ DE ESTIMADORES

El empleo de la función de influencia puede arrojar luz sobre la robustez de un estimador. Las siguientes son algunas medidas sobre el comportamiento de un estimador obtenidas usándola como materia prima:

(i). Sensibilidad a errores burdos, $R_{T,F}$

Mide la peor influencia aproximada que una cantidad fija de contaminación puede tener sobre el valor de un estimador. Se calcula como

$$R_{T,F} = \underset{Xc}{Sup}\ ABS\ IC_{T,F}(Xc)$$

(ii). Sensibilidad al cambio local, $B_{T,F}$

Mide el peor efecto posible de ajustar un contaminante modificando su valor, por ejemplo winsorizando. Se calcula como

$$B_{T,F} = Sup\ ABS\ (IC_{T,F}(Xc) - IC_{T,F}(W)/(Xc - W))$$

(iii). Punto de rechazo.

Suponga que la función de influencia desaparece en todos los puntos Xc fuera de un intervalo finito $Xc - u < p$, centrado sobre u (u otro punto de localización adecuado) de F. Esto implica que las observaciones fuera del intervalo $(u - P, u + P)$ no tienen influencia sobre el estimador T. (Es decir, el procedimiento de estimación rechaza tales observaciones).

(iv). Punto de ruptura

Es la más pequeña proporción de contaminación la cual puede producir que el estimador sobrepase cualquier cota. En otras palabras es la máxima cantidad de contaminación que un estimador puede absorber antes de ser completamente no confiable.

CONCLUSION

Han pasado dos décadas desde que los conceptos fundamentales de la estadística robusta fueron formulados. Es el momento para que los nuevos

procedimientos encuentren una amplia aplicación práctica. En este momento existen programas de computación sobre los procedimientos más ampliamente recomendados en estimación y *tests* de hipótesis en problemas de localización, escala y análisis de regresión. Algunos de ellos son el sistema *ROBSYS*, el cual está basado en el conjunto de subrutinas *ROBETH* (A. Randriamihrisoa y otros, 1985) y el programa de regresión robusta *PROGRESS* (Leroy y Rosseeaw, 1985).

BIBLIOGRAFIA

Ancombe, F.J. (1960a). "Rejections of Outliers". *Technometrics*, 2. pp. 123-147.

Barnett, V. y Lewis, T. *Outliers in Statistical Data.* New York, John Wiley & Sons, 1984.

Fisher, R. "On the Mathematical Foundations of Theoretical Statistics". *Philosophical Transactions of the Royal Society*, Serie A, No. 222, 1922. pp. 309-368.

Hampel, F. "The Influence Curve and its Rôle in Robust Estimation". *JASA*, Vol. 69. pp. 383-393.

Hampel, F. "Contributions to the Theory of Robust Estimation". Tesis de Ph.D. Universidad de California, Berkeley, 1968.

Hampel, F., Ronchetti, E., Rosseeaw, P. y Stahel, W. *Robust Statistics.* New York, John Wiley & Sons, 1986.

Huber, P. *Robust Statistics.* New York, John Wiley & Sons, 1981.

Leroy, A. y Rosseeaw, P. *PROGRESS*, Center for Statistics y O.R., Universiteit Brussel, 1985.

Randriamiaharisoa, A., Marazzi, A. y Braoudakis, G. *ROBSYS.* Lausane, Institut Universitaire de Medecine Sociale et Préventive, 1985.