
SOBRE ALGUNAS TECNICAS DE REMUESTREO: El Metodo de "JACKKNIFE"

Roberto Behar Gutiérrez, Profesor Titular
Mario Yepes Arango, Profesor Asociado
Dep. de Información y Sistemas
Universidad del Valle
Cali - Colombia

RESUMEN

En el presente trabajo se hace una presentación y revisión de la tecnica de estimación "jackknife" introducida por QUENOUILLE. Se examina el papel que juega esta técnica en la reducción de sesgos, en la estimación de varianzas, en la estimación por intervalos y en la estimación de funciones no lineales de los parámetros de un modelo lineal.

Se presenta además un estimador tipo "jackknife" para el sesgo de un amplio conjunto de estimadores y se compara el desempeño del "jackknife ordinario" y el "jackknife agrupado" en la reducción del sesgo.

La bibliografía mencionada representa una buena parte de los trabajos publicados sobre la metodología Jackknife.

1 INTRODUCCION

En algunas ocasiones se presenta dificultad para encontrar estimadores de un parámetro θ que posean propiedades deseables como insesgamiento y precisión; además, cuando se dispone de un estimador sesgado, generalmente no se dispone de metodología para la estimación del sesgo. Otras veces, se dispone de un estimador insesgado pero no se conoce una expresión cerrada para su varianza o para un estimador de la misma, lo cual obliga a realizar aproximaciones de distinta naturaleza, las cuales no producen los resultados más deseables.

Para el modelo de regresión lineal, en lo que respecta a la estimación de parámetros, existe un gran desarrollo, basado en la teoria general de los modelos lineales, lo cual proporciona muy valiosos recursos para resolver problemas de estimación y pruebas de hipótesis, sobre funciones lineales de los parámetros; sinembargo en no pocas oportunidades y en variadas disciplinas es de interes corriente la estimación de funciones no lineales de los parámetros de un modelo lineal, lo cual introduce una complicación que es necesario estudiar; una solución simple que algunos adoptan es usar como estimador de la función paramétrica:

$$\theta = h(\beta)$$

el estimador natural:

$$\hat{\theta} = h(\hat{\beta})$$

donde $\hat{\beta}$ es el estimador de mínimos cuadrados del modelo lineal original. Sin embargo, el estimador $h(\hat{\beta})$ no siempre tiene propiedades deseables importantes y su bondad depende mucho de la naturaleza de la función $h(\cdot)$; en particular: $\hat{\theta} = h(\hat{\beta})$

no es, en general, insesgado para el parámetro de interés: $\theta = h(\beta)$

Estos problemas y otros que podran deducirse en la descripción de la tecnica, son resueltos o por lo menos atenuados al usar el metodo de estimación "jackknife", cuando se cumplen algunas condiciones, que al parecer no son muy restrictivas.

Un ingenioso dispositivo para reducir el sesgo de un estimador fué propuesto por Quenouille (1949); su idea fué explorar las propiedades de un estimador construido a partir de un estimador dado, t , de un parámetro θ . Quenouille sugirió "cortar" la muestra en dos submuestras de igual tamaño y evaluar en cada una de ellas el estimador inicial t , obteniéndose así los valores t_1 y t_2 , los cuales se combinan en una sola expresión, generando, de esta manera, un nuevo estimador al que le estudio sus propiedades, observando que para cierto tipo de estimadores t , sesgados, se obtenian grandes mejoras en cuanto a reducción del sesgo. Posteriormente, Quenouille (1956), generalizó su idea y propuso la subdivisión de la muestra en g grupos de tamaño h , $n=gh$, y exploró su aplicabilidad general. TUKEY (1958) avanzó en el hallazgo de propiedades distribucionales aproximadas para el estimador de Quenouille, lo cual permite realizar a partir de este, estimaciones por medio de intervalos de confianza. Además, el propio Tukey popularizó una valiosa aplicación para este tipo de estimador, la cual se basa en propiedades de la varianza del estimador de Quenouille con respecto a la varianza del estimador original t , las que puedan usarse para la estimación de esta última, con base en la primera. Esta aplicación ha permitido resolver muchos problemas de muestreo.

Esta tecnica para generar estimaciones, debido a que el procedimiento de construcción requiere del corte de la muestra en submuestras se le ha llamado el método "jackknife" (cuchillo de cortar carne) y con este nombre se ha hecho popular en la literatura científica. A pesar de que esta tecnica surge en los años cincuenta, solo tiene popularidad en las últimas dos decadas, debido a que su cálculo requiere de una adecuada infraestructura de computo. El desarrollo de la informática ha permitido su mayor aplicabilidad y sobre todo, ha permitido la intensificación de la investigación empirica de sus propiedades a traves de la simulación.

Poco a poco el método de estimación "jackknife" ha ido penetrando en todas las areas de la estadística, convirtiéndose así, en un tema de actualidad en la mayoría de las revistas científicas de estadística y otras ciencias que hacen uso intensivo de estimaciones, basta leer el libro de SEBER (1982), para notar las múltiples referencias sobre este tema en aplicaciones relacionadas con la estimación de la abundancia animal.

En lo que respecta a sus aplicaciones en lo relacionado a funciones no lineales de parámetros del modelo lineal, debe mencionarse a HINKLEY (1977) y varios más. Sin pretender ser exhaustivos, se mencionarán algunos nombres de investigadores que adicionalmente a Quenouille y Tukey, han estado ligados al desarrollo teórico de esta tecnica de estimación; ellos son J.N.K. RAO (1965), RAO Y WEBSTER (1966), TIN (1965), P.S.R.S. RAO (1974), P.S.R.S. RAO Y J.N.K RAO (1971), MILLER (1964 , 1968 , 1974 b), SCHUCANY - GRAY - OWEN (1971), DUNCAN (1978), D. KREWSKI Y J.N.K RAO (1981). Más recientemente estan los trabajos de J. SHAO Y C.F.J. WU (1986), C.F.J. WU (1986, 1987) y los de EFRON (1981, 1982, 1985).

2 DESCRIPCION DEL METODO

Sean $Y_1, Y_2, \dots, Y_n$ una muestra aleatoria y t un estimador del parámetro θ , basado en la muestra de tamaño n .

Denotemos t_i al estimador t evaluado para los $(n - 1)$ elementos que quedan despues de separar el i -ésimo elemento de la muestra:

$$t_i = t(Y_1, Y_2, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n) \quad (1)$$

A

$$s_i = nt - (n - 1)t_i, \quad i = 1, \dots, n \quad (2)$$

se le llama i -ésimo pseudovalor.

El estimador "jackknife" de θ asociado al estimador inicial t y a la muestra $Y_1, Y_2, \dots, Y_n$, es el promedio de estos pseudovalores:

$$t_J = \frac{1}{n} \sum_{i=1}^n s_i \quad (3)$$

Si

$$t . = \frac{1}{n} \sum_{i=1}^n t_i \quad (4)$$

entonces:

$$t_J = nt - (n - 1)t . \quad (5)$$

que es, tal vez, la forma más usual de expresar el estimador "jackknife".

Conviene destacar, con base en lo anterior, que para obtener estimaciones "jackknife" se requiere un procedimiento de estimación inicial t y realizaciones de la muestra aleatoria:

$$t ; Y_1, Y_2, \dots, Y_n \Rightarrow \boxed{\text{Procedimiento "jackknife"}} \Rightarrow t_J$$

Segun Miller (1974), los terminos: "pseudovalores" para referirse a los s_i 's y estimación "jackknife" para referirse al metodo de estimación ideado por Quenouille, fueron acuñados por Tukey (1958).

Ya se habia mencionado que Quenouille (1956) sugirió una generalización de su procedimiento de estimación, hoy conocido como "jackknife agrupado", para el cual se supone que:

$$n = g.h \quad (6)$$

es decir las n observaciones se organizan en g grupos, cada uno de tamaño h .

Sea t_i la estadística t , calculada con las observaciones que quedan después de remover el i -ésimo grupo ($i = 1, \dots, g$), de la muestra.

Ahora los pseudovalores son:

$$s_i = gt - (g - 1)t_i \quad i = 1, \dots, g \quad (7)$$

y el estimador de "jackknife" agrupado:

$$t_J = \frac{1}{g} \sum_{i=1}^g s_i \quad (8)$$

ó

$$t_J = gt - (g - 1)t. \quad (9)$$

donde:

$$t. = \frac{1}{g} \sum_{i=1}^g t_i$$

Notese que el procedimiento de estimación "jackknife" es un caso particular del procedimiento de "jackknife agrupado", cuando $g = n$ y $h = 1$.

Este método requerirá menos esfuerzos de computo a medida que g se hace pequeño (h grande). Por tanto, se debe tratar de adoptar el estimador de "jackknife agrupado" que tenga las mayores propiedades deseables con el g más pequeño posible.

3 APLICACIONES

En la introducción, se hizo referencia a varias aplicaciones de la técnica de "Jackknife". A continuación se presentan, sin pretender ser exhaustivos, las que se refieren a reducción y estimación del sesgo, estimación de la varianza de un estimador, construcción de intervalos de confianza, estimación de funciones no lineales de parámetros en un modelo lineal.

3.1 REDUCCION Y ESTIMACION DEL SESGO

Supongamos que el estimador t para el parámetro θ , tiene un sesgo de la forma:

$$\begin{aligned} B(t) = \text{sesgo}(t) &= \sum_{k=1}^{\infty} \frac{a_k}{n^k} \\ &= \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \dots \end{aligned} \quad (10)$$

$$= O \left(\frac{1}{n} \right)$$

donde $a_1, a_2, \dots$, son constantes independientes del tamaño de la muestra, n .

Sea t_J el estimador "jackknife agrupado" de θ correspondiente al estimador t y a la muestra aleatoria $Y_1, Y_2, \dots, Y_n$, con base en g grupos de tamaño h , ($n = gh$), y considerese su valor esperado

$$\begin{aligned} E(t_J) &= E [gt - (g-1)t .] \\ &= g E(t) - (g-1) E(t .) \end{aligned}$$

$$E(t) = \theta + B(t) = \theta + \sum_{k=1}^{\infty} \frac{a_k}{n^k} \quad (11)$$

Y:

$$E(t .) = E \left[\frac{1}{g} \sum_{i=1}^g t_i \right] = \frac{1}{g} \sum_{i=1}^g E(t_i)$$

como:

$$E(t_i) = \theta + \sum_{k=1}^{\infty} \frac{a_k}{h^k (g-1)^k}, \quad i = 1, 2, \dots, g \quad (12)$$

entonces:

$$\begin{aligned} E(t .) &= \frac{1}{g} \sum_{i=1}^g \left[\theta + \sum_{k=1}^{\infty} \frac{a_k}{h^k (g-1)^k} \right] \\ E(t .) &= \theta + \sum_{k=1}^{\infty} \frac{a_k}{h^k (g-1)^k} \end{aligned} \quad (13)$$

Luego:

$$\begin{aligned} E(t_J) &= g\theta + g \sum_{i=1}^{\infty} \frac{a_k}{h^k g^k} - (g-1)\theta - (g-1) \sum_{k=1}^{\infty} \frac{a_k}{h^k (g-1)^k} \\ &= \theta + \sum_{k=1}^{\infty} \left[\frac{a_k}{h^k g^{k-1}} - \frac{a_k}{h^k (g-1)^{k-1}} \right] \\ &= \theta + \sum_{k=2}^{\infty} \frac{a_k}{h^k} \left[\frac{1}{g^{k-1}} - \frac{1}{(g-1)^{k-1}} \right] \end{aligned}$$

$$= \theta + \sum_{k=2}^{\infty} \frac{a_k}{n^k} \left[g - \frac{g^k}{(g-1)^{k-1}} \right]$$

si hacemos:

$$b_k = a_k \left[g - \frac{g^k}{(g-1)^{k-1}} \right] \quad (14)$$

Obtenemos:

$$E(t_J) = \theta + \sum_{k=2}^{\infty} \frac{b_k}{n^k} \quad (15)$$

$$= \theta + O\left(\frac{1}{n^2}\right)$$

Por lo tanto, el sesgo de t_J ,

$$\begin{aligned} B(t_J) &= \sum_{k=2}^{\infty} \frac{b_k}{n^k} \\ &= O\left(\frac{1}{n^2}\right) \end{aligned} \quad (16)$$

es de orden $\frac{1}{n^2}$.

En consecuencia el procedimiento de estimación "jackknife agrupado", t_J , conlleva una mejora substancial, sobre el procedimiento de estimación t , en lo que al sesgo se refiere, ya que t tiene un sesgo de orden $\frac{1}{n}$.

Resumiendo: Para cualquier estimador t de un parámetro θ con un sesgo de la forma (10) de orden $\frac{1}{n}$, el estimador "jackknife agrupado" correspondiente, tiene un sesgo de orden $\frac{1}{n^2}$, es decir, elimina el término $\left(\frac{a_1}{n}\right)$ del sesgo de t .

Esta propiedad de la estimación "jackknife" es de suma importancia, ya que existe una amplia colección de estimadores con sesgo de la forma (10), (Quenouille 1956).

En situaciones en las que el sesgo de t no es de la forma (10) o en las que no se tiene idea de su estructura, bien vale la pena plantear estudios amplios para observar el desempeño del estimador "jackknife".

OBSERVACIONES:

(a) ESTIMACION DEL SESGO:

La expresión (9) del estimador t se puede reescribir en la forma:

$$t_J = t - (g-1)(t. - t) \quad (17)$$

Ahora, se desea evaluar el valor esperado de $(g-1)(t. - t)$:

$$E[(g-1)(t. - t)] = (g-1)[E(t.) - E(t)]$$

$$= \sum_{k=2}^{\infty} \frac{a_k}{n^k} \left[\frac{g^k}{(g-1)^{k-1}} - (g-1) \right]$$

$$= \frac{a_1}{n} + \sum_{k=2}^{\infty} \frac{c_k}{n^k}$$

(haciendo, $c_k = a_k \left[\frac{g^k}{(g-1)^{k-1}} - (g-1) \right]$)

como se puede observar:

$$E[(g-1)(t - t)] = \frac{a_1}{n} + O\left(\frac{1}{n^2}\right)$$

que es de la misma forma del sesgo de t .

Este hecho llevó a algunos autores como Efron y Stein en 1981, a llamar a:

$$\hat{B}(t) = (g-1)(t - t)$$

estimador Quenouille o "jackknife" del sesgo de t .

En (17) es interesante notar que: t_J lleva incorporado el estimador del sesgo de t ; en forma más precisa:

$$t_J = t - \hat{B}(t)$$

es decir, el estimador "jackknife" correspondiente a t , es igual al estimador original t menos el estimador "jackknife" del sesgo de t .

(b) "JACKKNIFE ORDINARIO" vs "JACKKNIFE AGRUPADO", EN LA REDUCCION DEL SESGO:

Es evidente que los resultados obtenidos en la presente sección para el "jackknife agrupado" son validos en el caso particular $g = n$ y ($h = 1$), que corresponde al "jackknife" ordinario.

Por otra parte, se habia mencionado que los "esfuerzos" computacionales para el cálculo de las estimaciones "jackknife", son mayores a medida que g tiende a ser n , (es decir, a medida que el "jackknife agrupado" se "parece" más al ordinario), y ahora se ha verificado que, sin importar el valor de g , el sesgo del estimador t_J es de orden $\frac{1}{n^2}$. Por qué entonces no tomar siempre $g = 2$, o el menor valor posible de acuerdo con n ?

Para aclarar esto, vale la pena re-examinar el término genérico del sesgo de t_J :

$$\frac{b_k}{n^k} = \frac{a_k}{n^k} \left[g - \frac{g^k}{(g-1)^{k-1}} \right]$$

Como se puede notar este término no solamente depende de g , sino que disminuye en valor absoluto, cuando g crece arbitrariamente.

Esto significa que a pesar de que todos los t_J s tienen sesgo de orden $\frac{1}{n^2}$, el valor absoluto de dichos sesgos disminuye a medida que g aumenta.

La decisión entonces es adoptar un valor g "de equilibrio" con el cual las mejoras desde el punto de vista del sesgo, no compensen los costos computacionales adicionales, al aumentar g .

3.2 ESTIMACION DE LA VARIANZA DE UN ESTIMADOR

Supóngase que el estimador t del parámetro θ tiene propiedades deseables, sin embargo, debido a la relativa complejidad del modelo que involucra a θ , o por la estructura del método de muestreo, se dificulta encontrar una expresión para la varianza, $V(t)$, del estimador. Se presenta ahora una manera de aproximarse al conocimiento de la varianza $V(t)$.

Sea $t_J = nt - (n-1)t$. el estimador "jackknife" del parámetro θ . Donde los simbolos tienen la misma interpretación dada en la descripción del método.

Quenouille (1956) demostró que bajo ciertas condiciones no muy restrictivas para la función que define el estimador t , se cumple que:

$$\sigma_{t_J} = \sigma_t \left[1 + O\left(\frac{1}{n}\right) \right] \quad (18)$$

Lo cual muestra la coincidencia asintótica de $V(t_J)$ con $V(t)$.

Este hecho se utiliza para usar estimadores de $V(t_J)$ como estimadores para $V(t)$, siendo mayor la aproximación a medida que la muestras se hacen grandes; así pues, puede escribirse que:

$$\hat{V}(t) = \hat{V}(t_J) = \frac{1}{n(n-1)} \sum_{i=1}^n (s_i - t_J)^2$$

o en la forma más general, cuando el "jackknife" se calcula con base en el corte de la muestra en g grupos:

$$\hat{V}(t) = \frac{1}{g(g-1)} \sum_{i=1}^g (s_i - t_J)^2$$

al reemplazar a s_i y t_J en función de t_i y t , puede probarse que:

$$\hat{V}(t) = \frac{(g-1)}{g} \sum_{i=1}^g [t_i - t]^2$$

de esta manera se presenta una solución al problema planteado.

3.3 CONSTRUCCION DE INTERVALOS DE CONFIANZA

Otra utilidad de la tecnica de Jackknife es la posibilidad de construir intervalos de confianza aproximados para el parámetro θ . Tukey (1958) mostró que los psendovalores:

$$s_i = gt - (g-1)t_i$$

si se supone que son independientes, generan la cantidad pivotal:

$$T = \frac{t_J - \theta}{\left[\frac{1}{g(g-1)} \sum (s_i - t_J)^2 \right]^{\frac{1}{2}}}$$

la cual tendria aproximadamente distribución t-student con $(g-1)$ grados de libertad y por lo tanto se distribuye aproximadamente normal estandar cuando g se hace grande, caso en el cual tiene más validez el supuesto de independencia de los psendovalores s_i . Simembargo cuando este supuesto no se cumple la distribución no tiene buena aproximación a la t-student con $(g-1)$ grados de libertad, aunque autores como Jones (1965) han propuesto algunos correctivos para estos casos. Este punto es todavia hoy, motivo de estudio.

3.4 ESTIMACION DE FUNCIONES NO LINEALES DE LOS PARAMETROS DE UN MODELO LINEAL

Se tiene el caso de la estimación de una función: $\theta = h(\beta)$ no lineal, donde β es tal que:

$$y = x\beta + \epsilon; \quad E[\epsilon] = 0; \quad V(\epsilon) = \sigma^2 I$$

en estos casos se toma como estimador inicial de θ , el estimador natural:

$$t(y) = \hat{\theta} = h(\hat{\beta})$$

el cual generalmente es sesgado y a partir de este estimador t , se construye un estimador usando la técnica de "jackknife", esperando de esta manera una reducción del sesgo e inclusive una posible reducción de la varianza, como se ha mostrado empíricamente en no pocos casos.

References

- [1] ARUESEN, J.N. 1969 *Jackknifing U-statistics*. Ann Math. Statist 40, 2076-2100.
- [2] AGUILAR, G.R. 1984 *Study on interpenetrating subsampling ratio and regression estimation with or without jackknifing*. Thesis P.H.D. University of California. Riverside, U.S.A.
- [3] DUNCAN, G.T. 1978 *An empirical study of jackknife-constructed confidence regions in non-linear regression*. Technometrics 20, 123-129.
- [4] DURBIN, J. 1959 *A note on the application of quenouille's method of bias reduction to the estimation of ratios*. Biometria 46, 477-480.
- [5] EFRON, B.; STEIN, C. 1981 *The Jackknife estimate of variance*. Annals of statistical 9, 586-596.
- [6] EFRON, B. 1981 *Nonparametric estimates of standar error. The Jackknife, the bootstrap and other methods*. Biometrika 68, 389-399.
- [7] EFRON, B. 1985 *Bootstrap confidence intervals for a class or parametric problems*. Biometrika 72, 45-58.
- [8] EFRON, B. 1982 *The Jackknife, the bootstrap and other resampling plans*. Society for industrial and applied mathematics. Philadelphia, U.S.A.
- [9] FAY, R.E. 1985 *A jackknifed chi-square test for complex samples*. JASA, 80, 148-157.
- [10] FOX, T., HINKLEY, D.; LARNTZ, K. 1980 *Jackknifing in nonlinear regression*. Technometrics 22, 29-33.
- [11] HINKLEY, D.V. 1977 *Jackknifing in unbalanced situations*. Technometrics 19, 285-292.
- [12] HINKLEY, D. 1982 *Jackknife methods*. Encyclopedia of statiscal sciencie, 280-286.
- [13] KREWSKI, D.; RAO, J.N.K. 1981 *Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods*. The annals of statistics 9, 1010-1019.
- [14] MILLER, R.G. 1964 *A trustworthy jackknife*. Annals of mathematical statistics 35, 1594-1605.
- [15] MILLER, R.G. 1968 *Jackknifing variances*. Annals of mathematical statistics 39, 567-582.

-
- [16] MILLER, R.G. 1974 a. *The Jackknife a review*. Biometrika, 61, 1-15
- [17] MILLER, R.G. 1974 b. *An unbalanced Jackknife*. Ann Statist 2, 880-891.
- [18] PARR, W.C.; SHUCANY, W.R. 1982 *Jackknifing L-statistics with smooth weight functions*. JASA, 77, 629-638.
- [19] PARR, W.S. 1983 *Small sample results for ratio estimators*. Biometrika 58, 625-630.
- [20] QUENOUILLE, M.H. 1949 *Approximate test of correlation in time series*. J.R. Statist, Soc. B 11, 68-69.
- [21] QUENOUILLE, M.H. 1956 *Notes on bias in estimation*. Biometrika, 43, 353-360.
- [22] RAO, J.N.K. 1965 *A note on the estimation of ratios by Quenouille's Method*. Biometrika, 52, 647-649.
- [23] RAO, J.N.K.; WEBSTER, J.T. 1966 *On two methods of bias reduction in the estimation of ratios*. Biometrika 53, 571-577.
- [24] RAO, P.S.R.S.; RAO, J.N.K. 1971 *Small sample results for ratio estimators*. Biometrika 58, 625-630.
- [25] RAO, P.S.R.S. 1974 *Jackknifing the ratio estimator* Sankhya: The indian journal of statistics 36, 84-97.
- [26] SHUCANY, W.R. - GRAY, H.L. - OWEN, D.B. 1971 *On bias reduction in estimation*. JASA. 66, 524-533.
- [27] SHAO, J. ; WU, C.F.J. 1986 *Some general theory for the jackknife* Technical report, university of wisconsin, madison.
- [28] SEBER, G.A.F. 1982 *The estimation of animal abundance and related parameters*. Charles Griffin y Company Ltda, London.
- [29] TIN, M. *Comparison of some ratio estimators*. Jasa 60, 294-307.
- [30] TUKEY, J.M. 1958 *Bias and confidence in not-quite large samples*. Annals of mathematical statistics, (abstract), 29, 614.
- [31] WU, C.F.J. 1986 *Jackknife, bootstrap and other resampling methods in regression analysis*. The annals of statistics 14, 1261-1295.
- [32] WU, C.F.J. 1987 *On the asymptotic properties of the jackknife histograms* Technical report, university of wisconsin, madison.