Bussi, Javier
Marí, Gonzalo
Méndez, Fernanda
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

BOOTSTRAP ROBUSTO EN REGRESIÓN LINEAL: EL CASO DE TRES PREDICTORES

Introducción

La inferencia realizada a través de los estimadores robustos debe considerar la precisión de las estimaciones de los parámetros y reflejar la fuerza de los efectos observados. Es importante considerar la posibilidad de que los datos recolectados contengan observaciones atípicas (outliers) y por lo tanto se pretenda que la inferencia sea robusta ante esa situación. La inferencia paramétrica se basa en la distribución asintótica de los estimadores con sus respectivas propiedades pero su derivación es obtenida bajo escenarios ideales que involucran ciertos supuestos y la ausencia de outliers. Estos escenarios no garantizan la robustez de la inferencia y se basan en aproximaciones basadas en tamaños muestrales grandes. La inferencia que utiliza métodos clásicos Bootstrap requiere menos supuestos acerca de los datos pero implican un alto costo computacional y una pérdida de robustez de las estimaciones ante la presencia de observaciones atípicas. Una alternativa que cuenta con la ventaja de ser computacionalmente más sencilla y resistente a la presencia de outliers en la muestra es el Boostrap Rápido y Robusto (FRB: Fast and Robust Bootstrap). Este método en principio puede ser utilizado para cualquier estimador que pueda ser escrito como solución de un sistema de ecuaciones suaves de punto fijo como por ejemplo los estimadores MM.

Este trabajo describe este método en el caso de regresión lineal con más de un predictor. Se presentan simulaciones para el caso particular de tres regresores aleatorios, para distintos escenarios basados en suponer diferentes distribuciones del error, en donde se compara el rendimiento del FRB con la distribución asintótica empírica del estimador de regresión MM. Esta comparación se realiza a través de la cobertura y la amplitud de cada uno de los intervalos de confianza generados. Con el fin de ilustrar la utilidad del método FRB, se presenta una aplicación a datos reales donde se lo compara con el método Bootsrap Clásico (BC), teniendo en cuenta la estimación en la muestra original obtenida a través del estimador de regresión MM.

Metodología

Se describe a continuación el método FRB en el caso de regresión lineal. Se considera el escenario donde se cuenta con variables explicativas aleatorias y donde se aplica el método a estimadores de regresión MM. Las observaciones x_i no están prefijadas como en diseño de experimentos sino que son variables aleatorias que se observan conjuntamente con la variable dependiente y_i. Se describe brevemente a continuación el modelo lineal. Las observaciones x_i tienen dimensión px1 donde se considerará el caso de p=4, es decir, una regresión lineal con tres variables explicativas y una ordenada al origen. Se cuenta con n obser-

vaciones independientes e idénticamente distribuidas (i.i.d) de vectores aleatorios (yi, $\mathbf{z_i}'$)′ con distribución común H donde xi =(1, $\mathbf{z_i}'$)′.

Se considera el modelo lineal:

$$\mathbf{y_i} = \mathbf{x_i'}\boldsymbol{\beta_0} + \boldsymbol{\sigma_0}\boldsymbol{\varepsilon_i} \qquad \mathbf{i} = \mathbf{1}, \ldots, \mathbf{n} \tag{2.1}$$

La situación ideal corresponde a suponer que y_i y $\mathbf{z_i}$ son independientes, donde $y_i \sim F_0$ siendo ésta alguna distribución simétrica (en general la normal estándar), $\mathbf{z_i} \sim G_0$, $(y_i, \mathbf{z_i}')' \sim H_0$. Los errores $\boldsymbol{\varepsilon_i}$ pueden seguir distintas distribuciones.

Los estimadores MM de regresión están basados en dos funciones de pérdida $\boldsymbol{\rho_0}$ y $\boldsymbol{\rho_1}$, las cuales determinan la eficiencia y el punto de quiebre de la estimación, respectivamente.

El estimador MM, $\boldsymbol{\beta_n}$, satisface la ecuación:

$$\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho_1}' \left(\frac{\mathbf{y_i} - \mathbf{x_i'}\boldsymbol{\beta_n}}{\boldsymbol{\sigma_n}} \right) = \mathbf{0} \tag{2.3}$$

donde $\boldsymbol{\sigma_n}$ es un estimador S de escala que minimiza el M estimador de escala $\boldsymbol{\sigma_n}$ $\boldsymbol{\beta}$ definido en la ecuación:

$$\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho_0} \left(\frac{\mathbf{y_i} - \mathbf{x_i'}\boldsymbol{\beta_n}}{\boldsymbol{\sigma_n}} \right) = \mathbf{0} \tag{2.4}$$

La distribución asintótica de los estimadores MM ya ha sido estudiada el caso del modelo central paramétrico. Sin embargo, esta situación no se presenta en el caso de distribucines de errores no normales donde se pretende utilizar estimadores MM altamente robustos. El método FRB descripto a continuación conduce a estimadores consistentes de la covariancia de $\boldsymbol{\beta_n}$ bajo condiciones generales que incluye al caso normal.

Sea $\boldsymbol{\beta_n}$ el estimador de regresión S asociado a la ecuación (2.4). Se desea realizar inferencias acerca del parámetro de regresión $\boldsymbol{\beta_0}$. Utilizando el mismo criterio de “plug-in” del método BC propuesto por Efron (1979), se propone el siguiente método que produce un gran número de estimadores de regresión $\boldsymbol{\beta_n^*}$ recalculados a partir de las muestras generadas. A partir de la función de distribución empírica de esta estadística se estima la distribución muestral del estimador de regresión $\boldsymbol{\beta_n}$. Para cada vector $(y_i, \mathbf{z_i}')'$ perteneciente a la muestra se definen los residuos asociados con $\boldsymbol{\beta_n}$ y $\boldsymbol{\beta_n}$:

$$\mathbf{r_i} = \mathbf{y_i} - \mathbf{x_i'}\boldsymbol{\beta_n} \tag{2.5}$$

$$\mathbf{r_i} = \mathbf{y_i} - \mathbf{x_i'}\boldsymbol{\beta_n} \tag{2.6}$$

Se debe notar que tanto $\boldsymbol{\beta_n}$como $\boldsymbol{\sigma_n}$ pueden ser expresados como el resultado de un ajuste

por mínimos cuadrados ponderados (MCP). Para i=1,…, n se definen los pesos $\boldsymbol{\omega_i}$ y $\boldsymbol{\upsilon_i}$ como:

$$\boldsymbol{\omega_i = \rho_1^{'}\ r_i\ \ \sigma_n\quad r_i} \tag{2.7}$$

$$\boldsymbol{\upsilon_i = \frac{\sigma_n}{nb}\frac{\rho_0\ r_i\ \sigma_n}{r_i}} \tag{2.8}$$

pudiendo entonces representarse los estimadores de la siguiente manera:

$$\boldsymbol{\beta_n = \ \ \sum_{i=1}^{n}\omega_i x_i x_i^{'}\ ^{-1}\ \ \sum_{i=1}^{n}\omega_i x_i y_i} \tag{2.9}$$

$$\boldsymbol{\sigma_n = \ \ \sum_{i=1}^{n}\upsilon_i\ \ y_i - x_i^{'}\beta_n} \tag{2.10}$$

Si se considera la muestra bootstrap de las observaciones:

$$\boldsymbol{y_i^*, x_i^{*\ '}, i = 1, \dots, n} \tag{2.11}$$

y se definen las variables aleatorias:

$$\boldsymbol{\beta_n^* = \ \ \sum_{i=1}^{n}\omega_i^* x_i^* x_i^{*'}\ ^{-1}\ \ \sum_{i=1}^{n}\omega_i^* x_i^* y_i^*} \tag{2.12}$$

$$\boldsymbol{\sigma_n = \ \ \sum_{i=1}^{n}\upsilon_i^*\ \ y_i^* - x_i^{*'}\beta_n} \tag{2.13}$$

donde:

$$\boldsymbol{\omega_i^* = \rho_1^{'}\ r_i^*\ \ \sigma_n\quad r_i^*} \tag{2.14}$$

$$\boldsymbol{\upsilon_i^* = \frac{\sigma_n}{nb}\frac{\rho_0\ r_i^*\ \sigma_n}{r_i^*}} \tag{2.15}$$

$$\boldsymbol{r_i^* = y_i^* - x_i^{*'}\beta_n} \tag{2.16}$$

$$\boldsymbol{r_i^* = y_i^* - x_i^{*'}\beta_n},\ \text{para i=1,…,n} \tag{2.17}$$

Es importante notar que los estimadores $\boldsymbol{\beta_n}$, $\boldsymbol{\beta_n}$ y $\boldsymbol{\sigma_n}$ no se recalculan para cada muestra bootstrap. Por lo tanto los estimadores recalculados en (2.12) y (2.13) pueden no reflejar la variabilidad del vector aleatorio $\boldsymbol{\beta_n^{'}, \sigma_n}^{'}$ debido a que los pesos $\boldsymbol{\omega^*}$ y $\boldsymbol{\upsilon^*}$ son calculados utilizando estimaciones fijas. Con el fin de ajustar los resultados a causa de esta situación, se aplica una corrección lineal a los estimadores $\boldsymbol{\beta_n^*}$ y $\boldsymbol{\sigma_n^*}$ recalculados y se los combina.

Sea entonces:

$$\mathbf{M_n} = \boldsymbol{\sigma_n} \quad \sum_{i=1}^{n} \rho_1'' \; r_i \; \sigma_n, x_i \; x_i x_i' \quad^{-1} \quad \sum_{i=1}^{n} \omega_i x_i x_i' \tag{2.18}$$

$$\mathbf{d_n} = \mathbf{a_n^{-1}} \quad \sum_{i=1}^{n} \rho_1'' \; r_i \; \sigma_n, x_i \; x_i x_i' \quad^{-1} \quad \sum_{i=1}^{n} \rho_1'' \; r_i \; \sigma_n, x_i \; r_i \, x_i \tag{2.19}$$

$$\mathbf{a_n} = \frac{1}{nb} \quad \sum_{i=1}^{n} \rho_0' \; r_i \; \sigma_n \; r_i \; \sigma_n \tag{2.20}$$

7

El valor recalculado de $\boldsymbol{\beta_n - \beta}$ a través del bootstrap rápido está dado por:

$$\boldsymbol{\beta_n^{R*} - \beta_n = M_n \; \beta_n^* - \beta_n \; + d_n \; \sigma_n^* - \sigma_n} \tag{2.21}$$

Es importante notar que al recalcular $\boldsymbol{\beta_n^{R*} - \beta_n}$ no se resuelven las ecuaciones (2.3) y (2.4). En cada muestra bootstrap se resuelve el sistema de ecuaciones (2.12) y se calcula el promedio ponderado (2.13). Los factores de corrección M_n, d_n y a_n surgen de resolver dos sistemas lineales y un promedio ponderado respectivamente, y se calculan solo una vez con la muestra completa. Para estimadores de regresión MM con $\boldsymbol{\rho_0'}$ redecendiente, los pesos $\boldsymbol{\omega_i}$ brindan al método estabilidad ante la presencia de outliers. Los puntos que estén alejados estarán asociados a pesos bajos en (2.9) y (2.10). Los puntos atípicos extremos, es decir aquellos con residuos asociados muy grandes, recibirán pesos nulos y por lo tanto no tendrán efecto alguno sobre los coeficientes recalculados. Los pesos $\boldsymbol{\upsilon_i}$ utilizados en el recálculo de la escala son también decrecientes a medida que aumenta el valor absoluto de los residuos y en consecuencia las observaciones atípicas son menos influyentes también en el recalculo de $\boldsymbol{\sigma_n^*}$.

Simulaciones

Se realizó un estudio de simulación para investigar el comportamiento de los intervalos de confianza para los coeficientes del modelo de regresión basado en el método bootstrap rápido y robusto. Se desea encontrar que los intervalos tengan un porcentaje de cobertura cercano al valor del porcentaje de confianza, y que sean relativamente angostos.

Se consideraron muestras de tamaño n= 30, 50 y 100 con p=4 variables explicativas. Estas variables explicativas incluyen un intercepto: $\mathbf{x_1 = 1,}$ y tres variables explicativas $\mathbf{x_2 \sim N\ 0,1}$, $\mathbf{x_3 \sim N\ 0,1}$, $\boldsymbol{y}\ \mathbf{x_4 \sim N\ 0,1}$, independientes entre sí. Se postulan distintas distribuciones para el término del error: Normal estándar, t de student con 3 grados de libertad y con 1 grado de libertad (Cauchy).

Se generaron 1000 conjuntos de datos a partir de las distribuciones anteriores y se construyeron intervalos de confianza del 95% para los parámetros del modelo de regresión lineal.

Los intervalos de confianza para los coeficientes del modelo de regresión con bootstrap rápido y robusto se construyeron usando el método de sesgo corregido y acelerado (BCa) y también se presentan los intervalos bootstrap básicos (Basic). Los intervalos bootstrap son

comparados con los intervalos de confianza basados en la normalidad asintótica de los MM estimadores.

La tabla 1 presenta el porcentaje de intervalos de confianza del 95% que se observó que contenían al verdadero valor del parámetro. Se consideraron los intervalos para los parámetros asociados a las variables $\mathbf{x_2}, \mathbf{x_3}, \ y \ \mathbf{x_4}$. Cuando el tamaño de muestra es de 30, el FRB es superior en todos los casos respecto al método AV, presentando una cobertura mayor y más cercana al valor nominal. En cuanto al ancho de los intervalos, los Bootstrap presentan una amplitud un poco mayor siendo la diferencia entre las mismas más acentuadas en el caso Normal. Los resultados observados respecto a la cobertura del método AV son los esperados debido al carácter asintótico del mismo.

Para los tamaños muestrales 50 y 100, se observa en general que la cobertura de los intervalos FRB son superiores a las correspondientes a los intervalos AV, aproximándose al valor nominal de cobertura. Se observa que en alguno casos la cobertura observada supera el 95%, haciéndolos un poco más conservadores. Respecto a las amplitudes, se observan resultados similares a los observados para n=30.

Tabla 1: Cobertura y amplitud promedio de los intervalos de confianza del 95% para el modelo de regresión lineal con p=3

n	Distribución	Parámetro	FRB (BCa)	FRB (Basic)	AV
30	Normal (0,1)	β_1	0.910 (0.976)	0.913 (0.969)	0.901 (0.780)
		β_2	0.937 (0.976)	0.933 (0.969)	0.913 (0.787)
		β_3	0.917 (0.990)	0.913 (0.982)	0.909 (0.786)
	Cauchy	β_1	0.943 (1.758)	0.931 (1.743)	0.904 (1.594)
		β_2	0.944 (1.803)	0.941 (1.785)	0.926 (1.601)
		β_3	0.944 (1.741)	0.931 (1.727)	0.927 (1.593)
	t-student (3)	β_1	0.923 (1.187)	0.922 (1.179)	0.903 (0.994)
		β_2	0.935 (1.251)	0.915 (1.244)	0.894 (0.987)
		β_3	0.930 (1.241)	0.930 (1.231)	0.914 (1.030)
50	Normal (0,1)	β_1	0.946 (0.650)	0.937 (0.645)	0.931 (0.591)
		β_2	0.949 (0.654)	0.946 (0.649)	0.935 (0.594)
		β_3	0.939 (0.644)	0.934 (0.639)	0.923 (0.585)
	Cauchy	β_1	0.960 (1.141)	0.963 (1.130)	0.950 (1.112)
		β_2	0.960 (1.143)	0.957 (1.133)	0.953 (1.106)
		β_3	0.948 (1.165)	0.949 (1.153)	0.940 (1.123)
	t-student (3)	β_1	0.937 (0.801)	0.939 (0.794)	0.929 (0.745)
		β_2	0.946 (0.813)	0.953 (0.806)	0.933 (0.753)
		β_3	0.946 (0.807)	0.950 (0.799)	0.942 (0.750)
100	Normal (0,1)	β_1	0.942 (0.420)	0.939 (0.416)	0.938 (0.404)
		β_2	0.954 (0.421)	0.947 (0.418)	0.947 (0.405)
		β_3	0.946 (0.422)	0.945 (0.418)	0.940 (0.406)
	Cauchy	β_1	0.945 (0.728)	0.952 (0.721)	0.947 (0.720)
		β_2	0.956 (0.728)	0.961 (0.721)	0.957 (0.721)
		β_3	0.950 (0.740)	0.955 (0.734)	0.954 (0.734)
	t-student (3)	β_1	0.957 (0.529)	0.955 (0.525)	0.954 (0.515)
		β_2	0.955 (0.523)	0.957 (0.518)	0.950 (0.508)
		β_3	0.942 (0.526)	0.943 (0.522)	0.940 (0.511)

Conclusiones

En este trabajo se presentó el método Bootstrap rápido y robusto que es una alternativa a los métodos bootstrap clásicos para estimar la distribución de los estimadores de regresión robustos. El método FRB es computacionalmente más sencillo y resistente a la presencia de outliers. Se describió este método en el caso de regresión lineal con tres predictores aleatorios. Se realizaron simulaciones para distintos escenarios con diferentes tamaños muestrales y distribuciones de los errores, en donde se comparó el rendimiento del FRB con la distribución asintótica empírica del estimador de regresión MM. Esta comparación se efectuó a través de la cobertura y la amplitud de cada uno de los intervalos de confianza generados. Una vez establecido el nivel de confianza y de acuerdo a lo observado para la cobertura y la amplitud se considera que el FRB provee mejores resultados que los correspondientes a la distribución asintótica empírica.

REFERENCIAS BIBLIOGRÁFICAS

Efron, B., Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Efron, B. (1979). Bootstrap Methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Hampel, F. (1968). *Contributions to the theory of robust estimation*. PhD. Thesis, University of California, Berkeley.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Hastie, T. T. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Second ed.). New York: Springer-Verlag.

Huber, P., Ronchetti, E. (2009). *Robust Statistics (Second ed.).* Hoboken, New Jersey: John Wiley & Sons, Inc.

Jureckova, J., Picek, J. (2006). *Robust Statistical Methods with R.* Boca Raton, Florida: Chapman & Hall/CRC. John Wiley and Sons, Ltd.

Maronna, R.A., Martin, R.D., Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*. John Wiley and Sons.

Salibian-Barrera, M. (2000). *Contributions to the theory of robust inference*. Ph.D. thesis, Dept. Statist., Univ. British Columbia,Vancouver.

Salibian-Barrera, M., Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics*, 30, 556-582.

Van Aelst, S., Willems, G. (2005). Multivariate regression S-estimators for robust estimation and inference. *Statistica Sinica*, 15, 981-1001.

Van Aelst, S., Willems, G. (2013). Fast and robust bootstrap for multivariate inference: The R package FRB. *Journal of Statistical Software*, 53 (3),1–32.URL:http://www.jstatsoft.org/v53/i03/.