UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2008, Vol. 1, No. 1

Un acercamiento a los datos censurados y el bootstrap

An Approach to Censored Data and the Bootstrap

Yesid Rodriguez[a]
heivarrodriguez@usantotomas.edu.co

Ronne Tamayo[b]
ronnetamayo@bancodecredito.com

Resumen

En este articulo de carácter divulgativo, se busca mostrar algunos resultados relacionados con las curvas de sobrevida bajo el método de Kaplan-Meier (K.M.), y su aplicación a los datos censurados por métodos de remuestreo, específicamente Jackknife y Boostrap.

***Palabras clave*:** Boostrap, Jackknife, método de Kaplan-Meier.

Abstract

The purpose of this paper is to communicate some recent results about survival analysis under the Kaplan-Meier curve and its aplication to censored data by resampling methods such as Jackknife y Boostrap.

***Key words*:** Boostrap, Kaplan-Meier, Jackknife.

1. Introducción

El estimador Jackknife de la varianza, se puede considerar como una aproximación al estimador Bootstrap, cuando el estadístico de estudio es suficientemente suave[1]; pero esto no implica, que el estimador Bootstrap, sea siempre mejor que el estimador Jackknife. Puesto que el método Jackknife, requiere menos cálculos que el método Bootstrap, realizar procedimientos de tipo Bootstrap puede resultar innecesario en estimación de varianzas, cuando se han utilizado los procedimientos de Jackknife.

[a]Profesor. Universidad Santo Tomas
[b]Analista de riesgo. Banco de Crédito. Helm Financial services

[1]Se dice que un estadístico es suave, cuando su función de distribución lo es. Una función de distribución F es suave, si es continua o tiene una función de densidad f. Una discusión mas profunda acerca de este concepto, puede encontrarse en Shao & Tu (1995, p.113)

El método Bootstrap (Efron 1979), es una técnica de estimación de la distribución muestral de un estadístico. El procedimiento Bootstrap, consiste en hacer una serie de replicaciones a cada miembro de una muestra, gran cantidad de veces. Una forma corta, se hace tomando una nueva muestra de la misma muestra con reemplazo, ya que cada vez que se selecciona una observación para la muestra, se permite a cada elemento de la muestra original tener la misma probabilidad de ser seleccionado. Por lo tanto, seleccionar una muestra es equivalente a reproducir cada elemento gran cantidad de veces y realizar un muestreo sin reemplazo.

En este artículo, se considera una situación llamada *datos censurados a derecha* y se hace uso de las técnicas Bootstrap para responder a cuestiones relacionadas con la estimación de curvas bajo el método de Kaplan-Meier (K.M.), como:

1. ¿Cuál es el error estándar de la curva K.M.?
2. ¿Cuál es el error estándar de un parámetro de localización, basado en K.M.?
3. ¿Cuál es el intervalo de confianza para la estimación?

Con respecto a la primera pregunta, se sabe que el método Bootstrap provee una nueva justificación para la formula de Greenwood y sugiere que el método Bootstrap, puede realizarse en situaciones de censura mas complejas.

La segunda cuestión ha sido abordada por Miller (1964), quien propuso soluciones mediante el método de Jackknife. Finalmente, para la tercera cuestión se tienen los intervalos de confianza para muestras no-paramétricas pequeñas que son bien conocidas para la mediana (Lehman 1975), pero no para otros estimadores. Otras exposiciones para esta situación incluyen los trabajos de Johnson (1978) acerca de intervalos de confianza basados en la distribución t-student.

2. Bootstrap en datos censurados

El Bootstrap para datos censurados es extremadamente simple y la teoría necesaria para su realización es mínima. Supongamos que se observa.

$$X_i = x_i, \text{para } i = 1, 2, 3, \ldots, n, \text{ donde } X_i \sim iid$$

Acorde a alguna función de distribución de probabilidad F desconocida. La realización x_i puede ser un valor real, bidimensional o tomar valores en R^n.

Entre los parámetros $\theta(F)$ están: la media, la mediana y la correlación. Este puede ser estimado, si se usa el estimador $\hat{\theta}(\hat{F})$, donde $\hat{F}$ es la función de distribución empírica agregando $1/n$ en cada valor observado x_i.

Sea $\sigma(F)$, alguna medida de precisión que se usa, si F es conocida. Por ejemplo: $\sigma(F) = SD_i = (\theta(\hat{F}))$, la desviación estándar de $\hat{\theta}$, cuando $X_1, \ldots, X_n \sim F$. El estimador Bootstrap de "precisión", es simplemente $\hat{\sigma}_{boot} = \hat{\theta}(\hat{F})$.

En otras palabras, $\hat{\sigma}_{boot}$ es una medida de *precisión* si el verdadero F es equivalente a $\hat{F}$. De la misma forma, $\hat{\sigma}_{boot} = \sigma(\hat{F})$ es el estimador máximo verosímil no-paramétrico de $(\hat{F})$. Efron (1979), muestra que el estimador de Jackknife de las desviaciones estándar es un aproximación lineal para $\hat{\sigma}_{boot}$ bajo las siguientes condiciones:

- Una muestra Bootstrap $X_1^*, \ldots, X_n^*$, es extraída de $\hat{F}$, en la cual cada X_i^*, independientemente toma los valores x_j, con probabilidad $1/n$, $j = 1, \ldots, n$. En otras palabras $X_1^*, \ldots, X_n^*$, es una muestra independiente de tamaño n, extraída sin reemplazo del conjunto de observaciones $x_1, \ldots, x_n$.
- Estos dan una función de distribución empírica Bootstrap $\hat{F^*}$, la distribución empírica de los n valores $X_1^*, \ldots, X_n^*$, y un correspondiente valor Bootstrap $\hat{\theta^*} = \theta(\hat{F^*})$
- Los dos anteriores pasos se repiten independientemente un gran número de veces N y sus resultados son los valores Bootstrap $\hat{\theta^{*1}}, \ldots, \hat{\theta^{*N}}$.
- El valor de $\hat{\sigma}_{boot}$ es aproximadamente la desviación estándar de los $\hat{\theta^*}$ valores,

$$\hat{\sigma}_{boot} = \sqrt{\frac{\sum(\theta^{*j})^2 - \left(\sum \theta^{*j}\right)^2 / N}{N-1}}$$

Los datos censurados a derecha, son de la forma $(x_1, d_1), \ldots, (x_n, d_n)$, donde x_j, es la $j - esima$ observación, censurada o no, y

$$d_j = \begin{cases} 1 & \text{si } x_j \text{ no censura} \\ 0 & \text{si } x_j \text{ censura} \end{cases}$$

Por conveniencia se asumirá que $x_1 < x_2 < \ldots < x_n$.

Si tenemos algún estimado funcional $\hat{\theta} = \theta(datos)$, basado en $\{(x_1, d_1), \ldots, (x_n, d_n)\}$, se argumenta que el estimador apropiado Bootstrap, $\hat{\sigma}_{boot}$ es el mismo para el caso en donde los datos no presentan censura, excepto que los datos individuales son ahora las parejas (x_j, d_j). En este caso, el procedimiento es el siguiente:

- Se extrae una muestra Bootstrap $(X_1^*, D_1^*), \ldots, (X_n^*, D_n^*)$, por muestras independientes n veces con reemplazo de $\hat{F}$.
- Sea $data^*$, que representa el conjunto de datos artificial, con el cual se calcula $\hat{\theta^*} = \theta(data^*)$
- Independientemente se repiten los pasos anteriores, N veces , obteniendo $\hat{\theta^{*1}}, \ldots, \hat{\theta_n^*}$
- Se calcula $\hat{\sigma}_{boot}$ como: $\hat{\sigma}_{boot} = \sqrt{\frac{\sum(\theta^{*j})^2 - (\sum \theta^{*j})^2/N}{N-1}}$

Ahora, si consideramos estadísticas de la forma $\hat{\theta} = \theta(\hat{S}^0)$, donde $\hat{S}^0(t)$, es la curva de K.M, es posible escribir $\hat{S}^0 = \Phi(\hat{F})$, donde Φ , es un cierto plano de distribuciones en $R \times \{0, 1\}$, para distribuciones en R, descrita por Pettereson (1977).

Por otro lado, considerando el mecanismo aleatorio de censura propuesto por Efron (1967) y Gilbert (n.d.), se tiene que

$$d_j = \text{mín}\{X_0^i, W_i\} \tag{1}$$

Donde X_0^i es la variable de interés y W_i, es alguna variable independiente censurada. La observación es la pareja (X_i, D_i), con $D_i = 1$ o 0 como $X_i = X_0^i$ o W_i, respectivamente.

La curva de K.M. $\hat{S}^0(t)$, es un estimador insesgado de la verdadera curva de sobrevida X^0, es decir $\hat{S}^0(t) \equiv Pr(X^0 > t)$ y está dada la formula:

$$\hat{S}^0(t) = \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j-1} \right)^{d_j} \tag{2}$$

Aquí k_t es el valor de k tal que $t[x_, x_{k+1})$; en otras palabras, el valor más grande observado censurado o no, es igual a o menor que t. (Si no hay censuras entonces todos los $d_j = 1$, y $\hat{S}^0(t) = \frac{n-k_t}{n}$, conocida como la función de distribución acumulada ordinaria (cdf)). Kaplan & Meier (1958), mostraron que $\hat{S}^0(t)$ es el estimador máximo-verosímil no-paramétrico para S_t^0. Si $d_n = 0$, entonces $\hat{S}^0(t) > 0$.

Una observación no censurada de X_0^i corresponde a una observación censurada de W_i y vice-versa, luego la verdadera curva de sobrevida para W se denota $R(t) \equiv Pr(W > t)$, teniendo un estimador máximo verosímil no-paramétrico dado por la siguiente expresión:

$$\hat{R}(t) = \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j+1} \right)^{1-d_j}$$

Ahora (1), implica que la verdadera curva de sobrevida para X, denotada por $S(t) \equiv Pr(W > t)$, es el producto de $S(t) = S^0(t)R(t)$. Luego el estimador máximo-verosímil no-paramétrico para $S(t)$ es:

$$\hat{S}(t) = \hat{S}^0(t)\hat{R}^0(t) = \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j+1} \right) = \frac{n-k_t}{n} \tag{3}$$

Esta expresión representa la distribución adicionando $1/n$ en cada x_j observado, censurado o no. (Nótese que (3) no es afectado en $\hat{S^0}(t)$, si $d_n = 0$; o correspondiente a $\hat{R}$ si $d_n = 1$).

Una versión obvia para datos censurados de l procedimiento Bootstrap para aleatorizar datos censurados, es obtener independientemente $X_i^{0*} \sim S^0$ y $W_i^* \sim R^0$ y definiendo $X_i^* = \text{mín}(X_i^{0*}, W_i^{0*})$, $D_i^* = 1$ o 0 como $X_i^* = X_i^{0*}$ o W_i^*respectivamente.

Nótese que $X_i^* = x_j$ con probabilidad $1/n$ para $j = 1, 2, \ldots, n$. Sin embargo, si $X_i^* = x_j$, entonces $D_i^* = d_j$ debido a:

1. $\hat{S^0}(t)$ agrega solo los x_j , teniendo $d_j = 1$
2. $\hat{R}$ agrega solamente estos x_j , teniendo $d_j = 0$
3. se asume que no hay empates

3. Discusión

Aleatorizar censurando es matemáticamente conveniente, debido a que esta situación puede ser completamente irreal en algunos casos. Por ejemplo, en una situación podemos encontrar que los tiempos de censura $w_1, \ldots, w_n$ toman valores fijos, de los cuales todos son conocidos por la estadística, siendo o no las x_i censuras. Un método Bootstrap obvio en este caso es calcular $\hat{S^0}(t)$ como (1), escogiendo $x_1^{0*}, \ldots, x_n^{0*} \sim iid\ \hat{S^0}(t)$ y definiendo $X_i^* = X_i^{0*}$ o w_i.

Esto no es lo mismo que muestrear n veces con reemplazo a partir de $\{(x_1, d_1), \ldots, (x_n, d_n)\}$, pero mediante simulaciones de Montecarlo se puede ver que los resultados pueden ser muy similares.

Agradecimientos

Agradecemos especialmente al árbitro anónimo por sus valiosos comentarios.

Referencias

Efron, B. (1967), 'The two sample problem with censored data', *Proccedings of the Fifth Berkley Symposium of Mathematical Statistic and Probability* **4**, 831 – 853.

Efron, B. (1979), 'Bootstrap methods: Another look at the jackknife', *Annals of Statistics* (7), 1 – 26.

Gilbert, J. P. (n.d.), 'Ramdon censorship', *Unpublished Ph.D Thesis* .

Johnson, N. J. (1978), 'Modifeied t tests and confidence intervals for asymmetric populations', *Journal of the American Statistical Association* pp. 536 – 597.

Kaplan, E. L. & Meier, P. (1958), 'Nonparametric estimation for incomplete observations', *Journal of the American Statistical Associaton* **53**, 457 – 481.

Lehman, E. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, Holden Day.

Miller, R. G. (1964), 'A thrustworthy jackknife', *Annals of Mathematical Statistics* **39**, 567 – 586.

Pettereson, A. V. (1977), 'Expressing the kaplan-meier estimator as a function of empirical subsurvival functions', *Journal of the American Statistical Associaton* **72**(360), 854 – 858.

Shao, J. & Tu, D. (1995), *The Jackknife and Bootstrap*, first edn, Springer-Verlag.