MATE3523-MATE4523. Estadística No Paramétrica y Remuestreo. Proyecto II, 2017-I

Comparación de métodos de remuestreo en el contexto de regresión.

El objetivo del proyecto es comparar distintos esquemas de remuestreo en cuanto a la eficiencia de estimación y a la probabilidad de contención del parámetro verdadero en el contexto de regresión cuando se estiman dos parámetros a la vez.
El archivo "bootstrap regresion.R" anexo contiene dos esquemas de generación de datos para regresión, uno homocedástico y el otro no. Estos esquemas incluyen la generación de la matriz de diseño, el vector de errores (normal en un caso y en el otro no) y el vector de variables respuesta, para $n = 90$ datos. También incluye el código para implementar el remuestreo por residuos y el remuestreo por "pares" cuando se estima un solo parámetro.

Para entender mejor y realizar el proyecto Ud. debe investigar los siguientes conceptos: *distancia de Mahalanobis* y *elipsoide de mínimo volumen*.

Elipses de confianza para estimar el par (β_2, β_3)

Para el par de parámetros (β_2, β_3) de los modelos considerados, se quiere comparar los procedimientos de remuestreo en la estimación de un conjunto de confianza que será una elipse. Una medida de calidad es la cobertura: ¿en que porcentaje de los casos la elipse producida contiene al par de parámetros verdadero? El otro aspecto a considerar es la eficiencia: ¿cual método produce elipses de menor área (es decir, una estimación más precisa)?

Como se usa la estimación por elipse de contención

1. Se toma el conjunto de pares

$$\mathcal{D} = \{(\hat{\beta}_2^{*,b}, \hat{\beta}_3^{*,b}), b \in 1 : B\}$$

producidos por el remuestreo y se resta del conjunto el estimador original $(\hat{\beta}_2, \hat{\beta}_3)$: $\mathcal{D}_c = \mathcal{D} - (\hat{\beta}_2, \hat{\beta}_3)$.

2. Se estima un centro y una elipse de contención de la nube $\mathcal{D}_c$. La estimación de la elipse de contención implica una estimación de matriz de covarianza (mediante la covarianza del conjunto de puntos que quedan dentro de la elipse, covarianza a la cual se aplica un factor de corrección). Sean $\tilde{\gamma}^*$ y C^*, el centro y la matriz de covarianza estimados.

3. Denotemos $\hat{\beta} = (\hat{\beta}_2, \hat{\beta}_3)$ y $\hat{\beta}^* = (\hat{\beta}_2^*, \hat{\beta}_3^*)$. Suponiendo simetría elíptica de los estimadores se tiene, aproximadamente

$$\hat{\beta}^* - \hat{\beta} \sim \mathcal{E}(\tilde{\gamma}^*, C^*)$$

donde $\mathcal{E}(\mu, A)$ denota una distribución elipticamente simétrica con centro μ y matriz de covarianza A. Por el principio de remuestreo se concluye que

$$\hat{\beta}^- \beta \sim \mathcal{E}(\tilde{\gamma}^*, C^*)$$

donde β es el vector (β_2, β_3). Se deduce entonces que el conjunto de confianza para β puede definirse a través de la distancia de Mahalanobis:

$$\Pr(\beta \in \{u \in \mathbb{R}^2 : (u - (\hat{\beta} - \tilde{\gamma}^*))^t (C^*)^{-1} (u - (\hat{\beta} - \tilde{\gamma}^*)) \leq r\}) \approx 0.95$$

siendo r el cuantil de 95% de las normas de Mahalanobis de los vectores en $\mathcal{D}_c$, es decir, de los valores

$$(\hat{\beta}^{*,b} - \hat{\beta} - \tilde{\gamma}^*)^t (C^*)^{-1} (\hat{\beta}^{*,b} - \hat{\beta} - \tilde{\gamma}^*)$$

El cálculo del elipsoide de volumen mínimo puede hacerse con el comando CovMde del paquete rrcov del lenguaje R. Por ejemplo
c1=CovMve(datos,alpha=.90), donde alpha especifica la fracción de datos a incluir en la elipse. De la estructura de datos que produce el comando CovMde pueden extraerse el centro y la matriz de covarianza con los comandos getCenter y getCov. Por ejemplo,
c1=CovMve(datos,alpha=.90)
getCov(c1).
El comando getDistance permite obtener las distancias Mahalanobis de la muestra respecto al centro y covarianza estimados por CovMve.

Esquemas de remuestreo

Se pide considerar en total cuatro esquemas de remuestreo: remuestreo de residuos y remuestreo de pares, y dentro de cada uno de estos, remuestreo no-paramétrico clásico y remuestreo suavizado. Para la elección del h en remuestreo suavizado en dimensión mayor que 1, ver el artículo de Bowman y Foster anexo.

Otros parámetros a variar son los siguientes: generar el vector de errores con la desviación estándar especificada en el archivo "bootstrap regresion.R" anexo y generar el vector de errores con cuatro veces esa desviación estándar. Usar muestras de 90 datos o muestras de 270 datos (para esta última opción modifique apropiadamente el código suministrado).

Dependiendo de su capacidad de cómputo, para ambos esquemas de generación de los datos (errores homocedásticos normales / errores heterodásticos no-normales) y cada tamaño de muestra, genere 200 o 500 conjuntos de datos, producir en cada caso el conjunto de estimación para cada método de remuestreo y determinar si el verdadero par (β_2, β_3) cae en el conjunto. Determinar también el volumen de la elipse de estimación (ver comando getDet). Extraer conclusiones en base a los 200 o 500 resultados obtenidos en cada caso.

Su informe debe incluir (i) Una introducción a los problemas considerados, (ii) Los detalles de implementación de las simulaciones realizadas, (iii) Discusión de resultados y (iv) Conclusiones. Procure que su código en R sea vectorizado, en la medida de lo posible.

Valor del proyecto: 20 pts. Fecha de entrega de su informe, martes 23-03-2017. Debe ser recibido antes de medianoche en la dirección electrónica ajquiroz@gmail.com.