# Tarea4 :: Bootstrapping Regression

*Marcos Olguín Martínez*

*Octubre de 2015*

## Contents



## Introducción

Considere el modelo de regresión múltiple ordinario, $Y_i = x_i^T \beta + \epsilon_i$, para $i = 1, \ldots, n$, donde las $\epsilon_i$ se asumen como variables aleatórias i.i.d. con media cero y varianza constante. Aquí, las $x_i$ y las $\beta$ son p-vectores de los predictores y parámetros, respectívamente. Un error ingénuo de bootstrapping sería volver a muestrear desde la colección de valores de respuesta para generar una nueva pseudo-respuesta, digamos una $Y_i^*$, para cada $x_i$ observada, y de este modo generar un nuevo conjunto de datos de la regresión. Entonces un vector de parámetros estimados mediante bootstrap, $\hat{\beta}^*$, serían calculados de esos pseudo-datos. Después de repetir el muestreo y estimación muchas veces, la distribución empírica de $\hat{\beta}^*$ sería usada para hacer inferencia sobre $\beta$. El **error** es que las $Y_i|x_i$ no son i.i.d. (ellos tienen diferentes medias condicionales). Por consiguiente no es apropiado generar una regresión bootstrap a los datos de la manera antes descrita.

Debemos preguntarnos que variables son i.i.d. para determinar una correcta aproximación bootstrap. Las $\epsilon_i$ son i.i.d. dado el modelo. Por lo tanto, una estartegia más apropiada sería hacer el bootstrap a los residuales.

## Regresión con Bootstrap

Empecemos por ajustar el modelo de regresión a los datos observados y obtener la respuesta ajustada $\hat{y}_i$ y los residuales $\hat{\epsilon}_i$. Muestremos un conjunto bootstrap de los residuales, $\{\hat{\epsilon}_i^*, \ldots, \hat{\epsilon}_n^*\}$, del conjunto de residuales ajustados, de manera aleatória y con reemplazo. (Notemos que las $\hat{\epsilon}_i$ en realidad no son independientes, aunque por lo general más o menos lo son.) Creamos un conjunto bootstrap de pseudo-respuestas, $Y_i^* = \hat{y}_i + \hat{\epsilon}_i^*$, para $i = 1, \ldots, n$.

Hacemos la regresión $Y^*$ sobre las $x$ para obtener un parámetro estimado bootstrap $\hat{\beta}^*$. Repetimos este proceso muchas veces para construir una distribución empírica para las $\hat{\beta}^*$ y que podemos usar para la inferencia.

Este enfoque es el más apropiado para los experimentos diseñados u otros datos donde los valores para las $x_i$ se fijan de antemano. La estrategia de realizar bootstrap a los residuos es el objetivo primario de los métodos simples de bootstrapping para otros modelos como los modelos autorregresivos, de regresión no paramétrica y los modelos lineales generalizados.

Realizar bootstrap sobre los residuales es relevante para la elección del modelo ofreciendo un ajuste apropiado a los datos observados, y en el supuesto de que los resíduos tienen varianza constante. Es importante destacar que si no se tiene la confianza de que éstas condiciones se cumplan, pensar en un método bootstrap distinto sea probablemente más apropiado.

Supongamos que los datos surgieron de un estudio observacional, donde tanto la respuesta y los predictores se miden a partir de un conjunto de individuos seleccionados al azar. En este caso, los pares $z_i = (x_i, y_i)$ pueden ser vistos como los valores observados para las variables aleatórias i.i.d. $Z_i = (X_i, Y_i)$ elaboradas a partir de una distribución conjunta respuesta-predictor. Para bootstrap, muestrear $Z_1^*, \ldots, Z_n^*$ completamente al azar con reemplazo del conjunto de pares observados, $\{z_i, \ldots, z_n\}$. Aplicar el modelo de regresión a los pseudo-datos resultantes para obtener un parámetro estimado bootstrap $\hat{\beta}^*$. Repetir estos pasos muchas veces, para entonces proceder a la inferencia como una primera aproximación. Esta aproximación de hacer bootstrap a los casos es a veces llamada *pares bootstrap*.

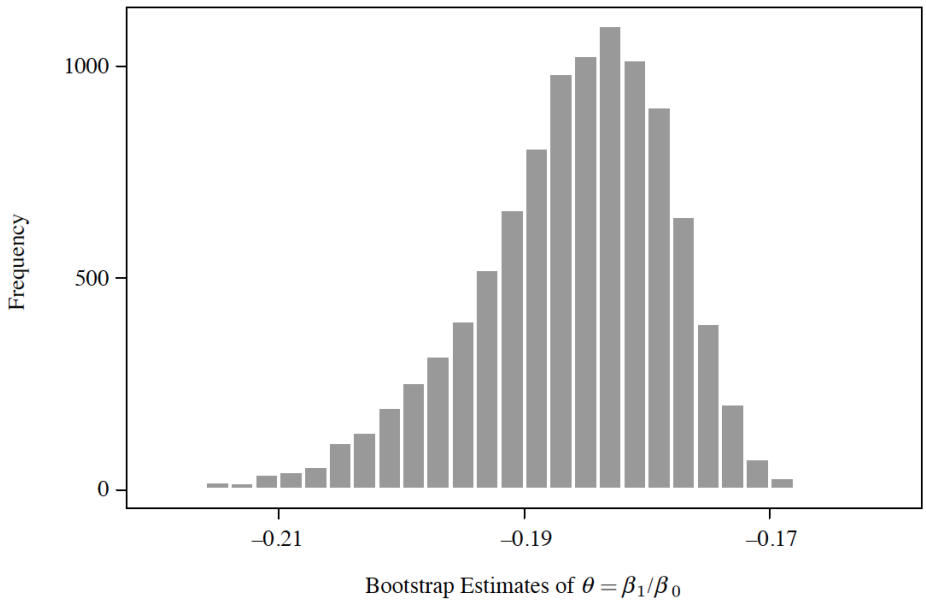
Si se tienen dudas acerca de la idoneidad del modelo de regresión, la constancia de la varianza residual, u otros supuestos de regresión, el emparejamiento bootstrap será menos sensible a las violaciónes de los supuestos que haciendo bootstrap a los residuos. El muestreo en emparejamiento bootstrap refleja más directamente el mecanismo original de la generación de datos en los casos donde los predictores no son considerados fijos.

## Ejemplo corrosión

En la siguiente base de datos se muestran 13 mediciones de la pérdida de corrosión ($y_i$) en aleaciones de cobre y níquel, cada una con un contenido específico de hierro ($x_i$). De interés es el cambio en la pérdida de la corrosión en las aleaciones a medida que aumenta el contenido de hierro, con relación a la pérdida de la corrosión cuando no hay hierro. Por lo tanto, considerar la estimación de $\theta = \beta_1/\beta_0$ en una regresión lineal simple.

```
##       x     y
## 1  0.01 127.6
## 2  0.48 124.0
## 3  0.71 110.8
## 4  0.95 103.9
## 5  1.19 101.5
## 6  0.01 130.1
## 7  0.48 122.0
## 8  1.44  92.3
## 9  0.71 113.1
## 10 1.96  83.7
## 11 0.01 128.0
## 12 1.44  91.4
## 13 1.96  86.2
```

Sean $z_i = (x_i, y_i)$ para $i = 1, \ldots, 13$ suponemos que adoptamos la aproximación mediante el emparejamiento bootstrap. Los datos observados producen la estimación $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_0 = -0.185$. Para $i = 2, \ldots, 10000$, trazamos un conjunto de datos de arranque $\{Z_1^*, \ldots, Z_{13}^*\}$ para remuestrear 13 pares de datos del conjunto $\{z_1, \ldots, z_{13}\}$ completamente aleatórios con reemplazo. La siguiente figura muestra el histograma de los estimadores obtenidos de las regresiones de los datos bootstrap. El histograma resume la variabilidad del muestreo de $\hat{\theta}$ como estimador de $\theta$.




# Más conceptos de Bootstrap y el paquete R

Como sabemos, el *Bootstrap* es un enfoque general de la inferencia estadística basada en la construcción de una distribución muestral de un estadístico por remuestreo de los datos a la mano. El término "Bootstrapping", debido a Efron (1979), es una alusión a la expresión "tirando a sí mismo por propio esfuerzo de uno", en este caso, usando la muestra de datos como una población de la cual repetidas muestras son tomadas. A primera vista, el enfoque parece circular, pero se ha demostrado que funciona.

Hay dos paquetes de R que nos pueden ayudar a realizar el bootstrap: **Paquete bootstrap** de Efron and Tibshirani (1993) y el **Paquete boot** de Davison and Hinkley (1997). De los dos, **boot**, es un tanto más capaz y es parte de la distribución estándar de R. El bootstrap es potencialmente muy flexible y se puede utilizar de muchas maneras, para usar el paquete **boot** se requiere algo de programación.

Hay varias formas de bootstrap, y, además, otros varios métodos de remuestreo que están relacionados con ella, como *jackknifing*, *validación cruzada*, *pruebas de aleatorización*, y *pruebas de permutación*. Aquí Vamos a insistir en el *bootstrap no paramétrico*.

El bootstrap no paramétrico permite estimar la distribución muestral de un estadístico empíricamente sin hacer suposiciones acerca de la forma de la población, y sin derivar la distribución de muestreo de forma explícita. La idea esencial del bootstrap no paramétrico es el siguiente: Se procede a extraer una muestra de tamaño $n$ de entre los elementos de la muestra **S**, muestreando con reemplazo. Llamemos al resultado de la muestra bootstrap $S_1^* = \{X_{11}^*, X_{12}^*, \ldots, X_{1n}^*\}$. Es necesario muestrear con reemplazo, ya que de otra manera simplemente se reproduce la muestra original de **S**. En efecto, estamos tratando la muestra **S** como una estimación de la población **P**; es decir, cada elemento $X_i$ de **S** es seleccionado para la muestra bootstrap con probabilidad $1/n$, imitando la selección original de la muestra **S** de la población **P**. Repetimos este procedimiento un gran número de veces, $R$, seleccionando muchas muestras bootstrap; la $bth$ muestra de bootstrap se denota $S^* = \{X_{b1}^*, X_{b2}^*, \ldots, X_{bn}^*\}$. Por consiguiente, la analogía clave de *bootstrap* es la siguiente:

**La población es a la muestra como la muestra es a las muestras bootstrap**

A continuación, se calcula la estadística de $T$ para cada una de las muestras bootstrap; es decir $T_b^* = t(S_b^*)$. Entonces la distribución de $T_b^*$ alrededor de la estimación original $T$ es análoga a la distribución muestral del estimador $T$ alrededor del parámetro de la población $\theta$. Por ejemplo, la media de las estadísticas bootstrapped,

$$\overline{T}^* = \hat{E}^*(T^*) = \frac{\sum_{b=1}^R T_b^*}{R}$$

Estima la esperanza del estidístico bootstrap; entonces $\hat{B}^* = \overline{T}^* - T$ es un estimador del sesgo de $T$, esto es, $T - \theta$. Similarmente, la varianza bootstrap estimada de $T^*$,

$$\widehat{Var}^*(T^*) = \frac{\sum_{b=1}^R (T_b^* - \overline{T}^*)^2}{R-1}$$

estima la varianza muestral de $T$. La raíz cuadrada de esta cantidad

$$\widehat{SE}^*(T^*) = \sqrt{\frac{\sum_{b=1}^R (T_b^* - \overline{T}^*)^2}{R-1}}$$

es el error estándar bootstrap estimado de $T$.

La selección aleatória de muestras bootstrap no es apecto escencial del bootstrap no paramétrico, y al menos en principio podríamos enumerar todas las muestras bootstrap de tamaño $n$. Entonces podríamos calcular $E^*(T^*)$ y la $Var^*(T^*)$ exactamente, en lugar de tener las estimaciones de ellos. El número de muestras bootstrap, sin embargo, es astronómicamente grande a menos que $n$ sea pequeña. Hay, por lo tanto, dos fuentes de error en la inferencia bootstrap: (1) el error inducido mediante el uso de una muestra **S** en particular para representar la población; y (2) el error de muestreo se produce al no enumerar todas las muestras bootstrap. La última fuente de error puede ser controlada haciendo el número de repeticiones bootstrap $R$ suficientemente grande.

Hay varias aproximaciones para construir los intervalos de confianza bootstrap. El intervalo con teoría normal asume que el estadístico $T$ es normalmente distribuido, que a menudo es aproximadamente el caso de las estadísticas en muestras sufcientemente grandes, y utiliza la estimación bootstrap de la varianza muestreal, y tal vez de sesgo, para construir un intervalo de confianza $100(1-\alpha)\%$ de la forma

$$\theta(T - \hat{B}^*) \pm z_{1-\alpha/2}\widehat{SE}^*(T^*)$$

donde $z_{1-\alpha/2}$ es el $1-\alpha/2$ cuantil de la distribución normal estándar (es decir, 1.96 para un 95% de intervalo de confianza, cuando $\alpha = 0.05$).

Una aproximación alterna, llamada el *intervalo percentil bootstrap*, es usado en los cuantiles empíricos de $T_b^*$ para formar un intervalo de confianza para $\theta$:

$$T^*_{(lower)} < \theta < T^*_{(upper)}$$

donde $T^*_{(1)}, T^*_{(2)}, \ldots, T^*_{(R)}$ son las replicaciones bootstrap ordenadas del estadístico; $lower = [(R+1)\alpha/2]$; $upper = [(R+1)(1-\alpha/2)]$; y los corchetes indican redondear al entero más cercano. Por ejemplo, si $\alpha = 0.05$ correspondiente a un intervalo de confianza del 95%, y R=999, entonces el lower = 25 y upper = 975.

El *sesgo-corregido*, intervalo percenti acelerado (o $BC_a$) realiza algo mejor que los intervalos de percentiles que se acaban de describir. Para encontrar el $BC_a$ intervalo para $\theta$ calcular:

$$z = \Phi^{-1}\left[\frac{\#_{b=1}^R (T_b^* \leq T)}{R+1}\right]$$

donde $\Phi^{-1}(\cdot)$ es la función cuantil de la normal estándar, y $\#(T_b^* \leq T)/(R+1)$ es la proporción (ajustada) de repeticiones bootstrap en o por debajo de la estimación T de $\theta$ con la muestra original.

Si la distribución muestral bootstrap es simétrica, y si $T$ es insesgada, entonces esta proporción será cercana a 0.5, y el factor de corrección z será cercano a 0.

- Sea $T_{(-i)}$ que representa el valor de $T$ que se produce cuando la $i$ th observación es borrada de la muestra; hay $n$ de esas cantidades. Sea $\overline{T}$ el promedio de las $T_{(-i)}$; tal que $\overline{T} = \sum_{i=1}^{n} T_{(-i)}/n$. Entonces calculmos

$$a = \frac{\sum_{i=1}^{n}(\overline{T} - T_{(-i)})^3}{6\left[\sum_{i=1}^{n}(T_{(-i)} - \overline{T})^2\right]^{\frac{3}{2}}}$$

- Con el factor de corrección $z$ y $a$ a la mano, calculamos

$$a_1 = \Phi\left[z + \frac{z - z_{1-\alpha/2}}{1 - a(z - z_{1-\alpha/2})}\right]$$

$$a_2 = \Phi\left[z + \frac{z + z_{1-\alpha/2}}{1 - a(z + z_{1-\alpha/2})}\right]$$

donde $\Phi(\cdot)$ es la función de distribución acumulativa de la normal estándar. Los valores $a_1$ y $a_2$ son usados para localizar los puntos finales del intervalo de confianza percentil corregido

$$T^*_{(lower^*)} < \theta < T^*_{(upper^*)}$$

donde $lower^* = [Ra_1]$ y $upper^* = [Ra_2]$. Cuando los factores de corrección $a$ y $z$ son cero, $a_1 = \Phi(-z_{1-\alpha/2}) = \Phi(z_{\alpha/2}) = \alpha/2$, y $a_2 = \Phi(z_{1-\alpha/2}) = 1 - \alpha/2$, el cual corresponde al intervalo percentil (sin corregir).

Para obtener suficiente precisión del 95% percentil bootstrap o el intervalo de confianza $BC_a$, el número de muestras bootstrap, $R$, debe ser del orden de 1000 o más; para los intervalos bootstrap de la teoría normal podemos salir con un menor valor de $R$, por ejemplo, del orden de 100 o más, ya que todo lo que necesitamos hacer es estimar el error estándar de la estadística.

**Para ver dos ejemplos revisar la aplicación en shiny de esta tarea.**

# Bibliografía

Fox, John and Weisberg, Sanford. Bootstrapping Regression Models in R. last revision: 5 June 2012. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Bootstrapping.pdf

Geof H. Givens and Jennifer A. Hoeting. Computational Statistics. Department of Statistics, Colorado State University, Fort Collins, CO.