Regresión Bootstrap

Jorge Luis Ramos Zavaleta
jorge.ramos@cimat.mx
Centro de Investigación en Matemáticas. Unidad Monterrey.

***Resumen*—La regresión lineal por OLS es uno de los métodos más empleados en el modelado estadístico, sin embargo existen muchos casos en los que se pueden generar tendencias y terminar con un mal ajuste en los datos. Uno de dichos casos en el que se puede generar uno de estos problemas en la presencia de datos atípicos (outliers), los cuales por la naturaleza del cálculo por OLS hacen que los coeficientes de regresión les den un mayor peso a dichos datos atípicos. En la literatura pueden encontrarse algunos otros métodos que permiten lidiar con datos atípicos como la regresión robusta, la cual depende del uso de pesos. Una alternativa a este método es la regresión bootstrap, que no incorpora datos extra sino que usa los ya existentes para generar coeficientes más robustos a dichos datos atípicos.**

I. Regresión Lineal

La regresión lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes X_i y un término aleatorio ϵ. Este modelo puede ser expresado como:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon \tag{1}$$

donde:

Y: variable dependiente, explicada o regresando.
$X_1, X_2, \ldots, X_n$: Variables explicativas, independientes o regresores.
$\beta_0, \beta_1, \ldots, \beta_n$: parámetros que establecen la influencia de las variables independientes sobre el la variable dependiente.

El caso más simple de la regresión lineal se da cuando se tiene solo una variable explicativa, en ese caso el modelo toma la forma:

$$Y = \beta_0 + \beta_1 X_1 \tag{2}$$

y la solucón analítica para este caso es muy simple y tiene la forma

$$\hat{\beta_1} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \hat{\beta_1}\bar{x}$$

en ambas ecuaciones puede observarse la presencia de las medias muestrales para cada una de las variables, por lo que se establecen pesos iguales a una probabilidad uniforme para cada una de las observaciones en cada variable, por lo que le da igual peso a observaciones que son atípicas que a otras que no por lo que la presencia de datos atípicos tiende a influenciar ambos parámetros.

Una solución directa para resolver este problema es reajustar los pesos que cada observación provee para generar los valores de dichos parámetros lo cual es la base de la regresión robusta. Sin embargo, esto implica que se debe tener una medida correcta para establecer dichos pesos, lo que en muchos casos no siempre sucede debido a que en dimensiones mas grandes el problema de encontrar datos atípicos no es tan simple, por lo que establecer un peso adecuado para dichas observaciones se puede volver complicado.

II. Regresión Bootstrap

Una alternativa no paramétrica a la regresión robusta es la regresión bootstrap. El método de Bootstrap hace referencia al uso de la muestra original de los datos y varios remuestreo de ella para generar estadisticos calculados a partir del remuestreo. Para generar una regresión bootstrap existen al menos dos formas: haciendo bootstrap con respecto a los residuales o haciendo bootstrap generando nuevos coeficientes.

En el primer caso se requiere realizar una regresión para conseguir los residuales, E_i. A partir de dichos residuales se obtiene nuevos valores para la variable a predecir simplemente sumando los estimados a los residuales, esto es:

$$Y^*_{bi} = \hat{Y}_i + E^*_{bi}$$

donde $\hat{Y}_i$ es el valor ajustado en la regresión original, y E^*_{bi} los residuales del remuestreo en la b-ésima muestra bootstrap. En cada muestra bootstrap se genera una nueva regresión usando las variables explicativas originales y Y^*_{bi}. Una desventaja de este método es que el remuestreo deja X fijo, y el procedimiento asume implícitamente que la forma funcional del modelo de regresión que se ajusta a los datos es correcta y que los errores se distribuyen de manera idéntica.

El segundo enfoque consiste en realizar bootstrap directamente usando los coeficientes de la regresión. Consideremos el conjunto $z'_i \equiv [Y_i, X_{i1}, \ldots, X_{ik}]$, para cada valor de la variable de respuesta y de sus regresores asociados en cada observación. Entonces, las observaciones $z'_1, \ldots z'_n$ pueden ser remuestreadas, y calcular para cada muestra los estimadores de la Ecuación (1) para cada muestra bootstrap, produciendo r conjuntos de coeficientes de regresión por bootstrap, $\beta^*_r = [\beta^*_0, \ldots, \beta^*_n]$. Al terminar todos los remuestreos obtenemos el promedio de cada uno de los coeficientes, β^*_r, y serán esos nuestros nuevos coeficientes.

III. Ejemplo

Para ejemplificar el funcionamiento de la regresión Bootstrap, consideramos un caso muy simple. Tomamos los datos del número de muertes por año, y queremos ver si el número de pubs en Londres instalados por año tiene relación con las muertes.

Los resultados de ambas regresiones se presentan en la figura III, donde puede observarse que la recta obtenida usando la regresión bootstrap (en color rojo) se ve mucho menos afectada por la presencia del dato atípico, a diferencia de la recta obtenida usando regresión lineal (en color azul).

Se obtuvieron también los intervalos de confianza para el parámetro correspondiente al numero de pubs; en el caso de la regresión lineal se tiene que el intervalo de confianza al 95 % es $IC_{95\,\%} = (-0{,}26617, 9{,}39093)$ y para el caso de la regresión bootstrap es $IC_{95\,\%} = (1{,}74807, 16{,}13362)$. Por su parte, en la regresión boostrap se puede ver que el intervalo no contiene al cero, lo cual indica que la relación entre el número de pubs y el número de muertes por año es significativa. En este ejemplo se puede observar el efecto que tiene un dato atípico sobre la inferencia de la relación entre las variables estudiadas.

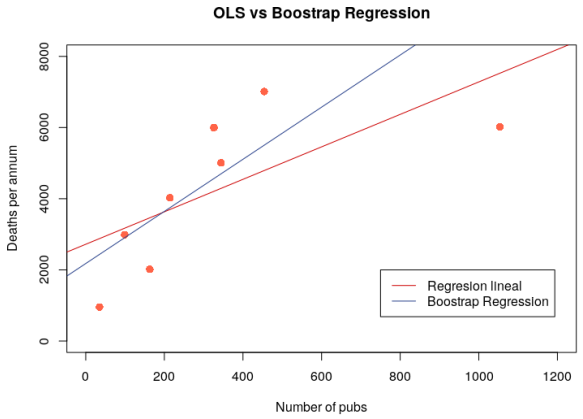

Figura 1. Resultados de Regresión por Boostrap vs OLS.

IV. Conclusiones

La mayoría de las veces, los residuales de la regresión no cumplen con los supuestos para OLS; problemas como la heterocedasticidad y otras violaciones como la colinealidad o la no inclusión de variables importantes se presentan comunmente. La presencia de datos que, aún cuando no son valores extremos, tienden a afectar en mayor medida a los coeficientes de regresión tambien es posible tener. En estos casos, el utilizar un enfoque de regresión Bootstrap ayuda bastante, ya que usando el Bootstrapping proporciona errores estándar más precisos, reduciendo la varianza del error, y genera intervalos de confianza más robustos. La desventaja que presenta este método es que es computacionalmente intensivo y, a medida que la cantidad de variables y observaciones aumenten el costo computacional tiende a aumentar al menos de manera lineal con ellos. Además, debe considerarse que el tener un elemento aleatorio en el remuestreo puede hacer que replicar un ejercicio no sea tan factible.

V. Apéndice

El código de R con el que se realizó el ejercicio se presenta a continuación:

```
# Introducimos los datos
pubs <- c(35,99,163,214,
          326,344,454,1054)
mortality<-c(957,2990,2017,4025,
             5997,5007,7012,6022)

#Generamos un dataframe con los datos
x<-as.data.frame(cbind(pubs,mortality))

# MODELO POR OLS
model1 <- lm(mortality ~., data=x)
summary(model1)

# Los coeficientes del modelo
betahat <- coef(model1)

# Fijamos una semilla para
# permitir reproducibilidad
set.seed(100)

# MODELO POR BOOTSTRAP
bstar <- NULL
n <- length(x$mortality)

#Numero de remuestreos a hacer
B=1000

for(draw in 1:B){
  Dstar<-x[sample(1:n,size=n,replace=T),]

  #Generamos el modelo con los datos
  #obtenidos por el remuestreo
  model=lm(mortality ~., data=Dstar)
  bstar=rbind(bstar, coef(model))
}

# Los nuevos coeficientes a
# partir de los obtenidos por bootstrap
boost.coef<-apply(bstar, 2, mean)

# INTERVALOS DE CONFIANZA
# (para el numero de pubs)

# Para el modelo de regresion lineal
confint(model1)
```

```
# Para el parametro obtenido por
# bootstrap

#Ordenamos los valores obtenidos por
# bootstrap para el parametro del numero
# de pubs

sorted.param<-sort(bstar[,2])

# Los limites inferior y superior
# para establecer los CI

bounds<-c(B*(0.05/2), B*(1-(0.05/2)))

#Calculamos los CI por percentiles

boost.ci<-c(sorted.param[bounds[1]],
        sorted.param[bounds[2]])
```