Aplicación de Boostrapping en Regresión I

Un modelo de regresión lineal basado en n observaciones $(\mathbf{x}_i, y_i)$ es de la forma

$$y_i = \mathbf{x_i}\boldsymbol{\beta} + e_i \qquad (i=1,2,\ldots..n)$$

donde y_i son los valores observados de la variable de respuesta y, y los $\mathbf{x_i}$ son vectores filas de dimensión p y contienen los valores observados de la variables predictoras, $\boldsymbol{\beta}$ representa un vector columna p–dimensional de parámetros desconocidos, y e_i representan un error aleatorio y se suponen que son una muestra aleatoria de una distribución F que tiene valor esperado 0 y varianza σ^2. Si las variables predictoras no son aleatorias esto implica que $E(y)=x_i\boldsymbol{\beta}$.
El problema en regresión es tratar de estimar el vector de parámetros $\boldsymbol{\beta}$ usando los datos observados $(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2)\ldots\ldots..(\mathbf{x}_n,y_n)$. El método más popular de estimación es el de cuadrados mínimos que consiste en minimizar la suma de cuadrados de los errores e_i. Es decir, que el estimador mínimo cuadrático $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ minimiza

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x_i}\boldsymbol{\beta})^2 = \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2$$

donde $\mathbf{y}=(y_1,\ldots.y_2)'$ y $\mathbf{X}$ es una matriz n x p con filas $\mathbf{x_i}$. Como se ha visto en clases anteriores, esto produce

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Se puede mostrar que la matriz de covarianzas de $\hat{\boldsymbol{\beta}}$ está dado por $\mathbf{(X'X)^{-1}}\sigma^2$, donde σ^2 representa la varianza poblacional de los errores, por supuesto por lo general desconocida. En consecuencia, el error estándar (estimado) de los coeficientes $\hat{\beta}_j$ estará dado por

$$se(\hat{\beta}_j) = \sqrt{c_{jj}}\, s$$

donde c_{jj} es el j-ésimo elemento de la diagonal de $\mathbf{(X'X)^{-1}}$ y *s* es la estimación de σ.

La aplicación de Bootstrapping en regresión está centrada mayormente en inferencias acerca del vector de parámetros $\boldsymbol{\beta}$, en la estimación del error de predicción y en selección de variables.
La estimación por bootstrapping del error estándar de los $\hat{\beta}_j$ estará dada por la desviación estándar de la distribución de sus valores bootstraps.

Hay dos maneras de aplicar Bootstrapping en un modelo de regresión: Usando Bootstrapping en las observaciones y usando Bootstrapping en los errores o residuales.

Aplicando Bootstrapping en las observaciones.

En este caso estamos se considera que las observaciones $(\mathbf{x}_i,y_i)$ son una muestra aleatoria de una distribución multivariada F de dimension p+1. Lo que se trata entonces es aproximar F mediante el bootstrapping de su distribución Empirica $\hat{F}$. Para esto se toman muestras con reemplazamiento de tamaño n de la muestra original $(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2)……..(\mathbf{x}_n,y_n)$. Esto es equivalente a primero seleccionar al azar n indices distribuidos uniformemente en el conjunto {1,2,….n}. Sea J el conjunto de indices seleccionados entonces $(\mathbf{x}^*,y^*)=\{(x_j,y_j): j \in J\}$ constituye la muestra bootstrap. Luego se calcula el modelo de regresión obteniendo una secuencia de B estimados bootstraps β^* . Lo anterior se puede resumir en el siguiente algoritmo

For k=1:nboot

1. Seleccionar al azar los números $i_1,…..i_n$ al azar y con reemplazo de {1,…n}
2. Para j=1,…..n . hacer $\mathbf{x}_j^* = \mathbf{x}_{i_j}$ y $y_j^* = y_{i_j}$
3. Ajustar el modelo de regresión con la muestra bootstrap $(\mathbf{x}_1^*, y_1^*)$ … $(\mathbf{x}_n^*, y_n^*)$ para obtener β^* y un estimado del error estándar del estimador minimo cuadrático en cada muestra bootstrap.

El error estandar estimado por bootstrap será la desviación estándar de los β^*.

La siguiente función en R encuentra el error estándar de los coeficientes de regresión lineal múltiple según el método de las observaciones.

```
function(X,nboot)
{#Calcula el error estandar de los coeficientes de
# regresion por bootstrapping de las observaciones
nobs=dim(X)[1]
p=dim(X)[2]
bboot=matrix(0,nboot,p)
for (k in 1:nboot)
{indexb=sample(1:nobs,nobs,replace=T)
y=X[indexb,p]
xp=X[indexb,-p]
bboot[k,]=lm(y~xp)$coef}
seboot=apply(bboot,2,sd)
}
```

Ejemplo: Estimar usando bootstrapping el error estándar de los coeficientes de la línea de regresión que relaciona log(peso del cerebro(g)) versus log(peso del cuerpo(kg)) de 62 mamiferos (los datos son tomados del libro Applied Regresión de Weisberg, pagina 144).

Aplicando la función anterior se obtiene los siguientes resultados.

```
> bootreg1(mama1,100)
 [1] 0.09033087 0.02245737
> bootreg1(mama1,500)
 [1] 0.08847689 0.02140932
> bootreg1(mama1,1000)
 [1] 0.08737012 0.02340214
```

Usando la función **lm** hallamos los siguientes valores para los estimados de los coeficientes de regresión y su desviación estándar.

```
summary(lm(mama1[,2]~mama1[,1]))

Call:
lm(formula = mama1[, 2] ~ mama1[, 1])

Residuals:
    Min      1Q  Median      3Q     Max
-1.71549 -0.49228 -0.06162  0.43597  1.94829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
mama1[, 1]   0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-Squared: 0.9208,     Adjusted R-squared: 0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

Ejemplo: Usando los datos de millaje estimar el error estándar del vector de coeficientes de regresión β de la regresión de mpg versus hp, sp, wt, y vol

```
> bootreg1(as.matrix(milla1),100)
 [1] 33.24549989  0.11090936  0.33037822  0.30821502  0.02831608
> bootreg1(as.matrix(milla1),500)
 [1] 36.39938799  0.12023476  0.36293610  0.31419272  0.02718504
> bootreg1(as.matrix(milla1),1000)
 [1] 35.05440743  0.11421464  0.34904788  0.30001664  0.02697104
>
```

Usando la función **lm** hallamos los siguientes valores para los estimados de los coeficientes de regresión y su desviación estándar.

```
>  summary(lm(mpg~.,data=milla1))

Call:
lm(formula = mpg ~ ., data = milla1)

Residuals:
   Min     1Q Median     3Q    Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
hp            0.39221    0.08141   4.818 7.13e-06 ***
sp           -1.29482    0.24477  -5.290 1.11e-06 ***
wt           -1.85980    0.21336  -8.717 4.22e-13 ***
vol          -0.01565    0.02283  -0.685    0.495
```

Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-Squared: 0.8733, Adjusted R-squared: 0.8667
F-statistic: 132.7 on 4 and 77 DF, p-value: < 2.2e-16

Notar que hay bastante diferencia entre los errores estándar que da el Bootstrapping y los valores exactos. La razón de la discrepancia es que hay multicolinealidad (dependencia lineal) entre las variables predictoras y ello afecta el rendimiento del bootstrapping.

Aplicando Bootstrapping en los errores

Si consideramos que las variables predictoras son fijas, no aleatorias, entonces se puede aplicar remuestreo solo a los errores del modelo de regresión mínimo cuadrático. En cada muestra bootstrap la matriz X es la misma. Lo anterior se puede resumir en el siguiente algoritmo

1. Ajustar el modelo mínimo cuadrático $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
2. For k=1:nboot
   i) Para j=1,…..n .
   a. Hacer $\mathbf{x}_\mathbf{j}^* = \mathbf{x}_\mathbf{j}$
   b. Seleccionar una muestra con reemplazo ε* de los residuales del modelo mínimo cuadrático.
   c. Hacer $y_i^* = \mathbf{x}_i\hat{\boldsymbol{\beta}} + e_i^*$
   ii) Ajustar el modelo de regresión con la muestra bootstrap $(\mathbf{x}_1^*, y_1^*) \ldots (\mathbf{x}_n^*, y_n^*)$ para obtener $\boldsymbol{\beta}^*$ y un estimado del error estándar del estimador mínimo cuadrático.

La siguiente función en R ejecuta el algoritmo.

```
bootreg2= function(X,nboot)
{#Calcula los coeficientes de regresion y sus errores
# estandar por bootstrapping los errores
nobs=dim(X)[1]
p=dim(X)[2]
y=X[,p]
xp=X[,-p]
dim(xp)
unos=rep(1,nobs)
xp1=cbind(unos,xp)
reg1=lm(y~xp)
bboot=matrix(0,nboot,p)
for(i in 1:nboot)
{indb=sample(1:nobs,nobs,replace=T)
eb=reg1$resid
ebo=eb[indb]
tempo=reg1$coef
yb=xp1%*%tempo+ebo
bboot[i,]=lm(yb~xp)$coeff
}
```

```
seboot=apply(bboot,2,sd)
}
```

Aplicando la función a los datos del primer ejemplo se obtiene

```
> bootreg2(mama1,100)
 [1] 0.09009297 0.02650733
> bootreg2(mama1,500)
 [1] 0.09816633 0.02554436
> bootreg2(mama1,1000)
 [1] 0.09452829 0.02728547
```

Nuevamente los valores estimados de los errores estándar aparecen algo alejados de los valores exactos.

Aplicando la función al conjunto de datos **millaje** se obtiene

```
> bootreg2(as.matrix(milla1),100)
 [1] 22.75545076  0.07679813  0.23482847  0.19751453  0.02156899
> bootreg2(as.matrix(milla1),500)
[1] 22.68274981  0.07889179  0.23683260  0.20565917  0.02179406
> bootreg2(as.matrix(milla1),1000)

[1] 22.85122925  0.07929919  0.23640728  0.20968180  0.02241869
>
```

Es bien evidente que aqui este método hace un mejor trabajo que el método de bootstrapping las observaciones.

Bootstrapping de las observaciones es menos sensitivo a suposiciones (normalidad, varianza constante e independencia) que bootstrapping de los errores.
Bootstrapping de las observaciones tiende a producir distribuciones con mayor variabilidad.

La colinealidad de las variables predictoras afecta el rendimiento de ambos métodos.

Ejemplo: Aplicar bootstrapping residuales a los datos **hormone** en donde se relaciona amount=cantidad de hormona que queda en un envase después de un numero dado de horas de uso (ver Efron y Tibshirani página 107).

```
>bootreg2(hormone,100)
 [1] 0.764487292 0.004132574
> bootreg2(hormone,500)
 [1] 0.818505867 0.004191695
> bootreg2(hormone,1000)
 [1] 0.800653016 0.004333684
>.
```