Simulación Estocástica

Bootstrap

Vanda Inácio de Carvalho

Primer Semestre 2015

Bootstrap

Introducción

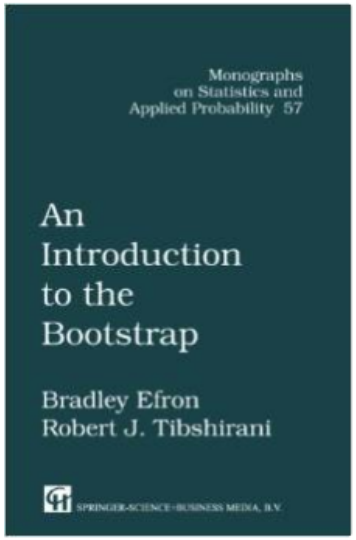
- El bootstrap fue introducido por Bradley Efron na década de 1970 y es una de las técnicas estadísticas más utilizadas.

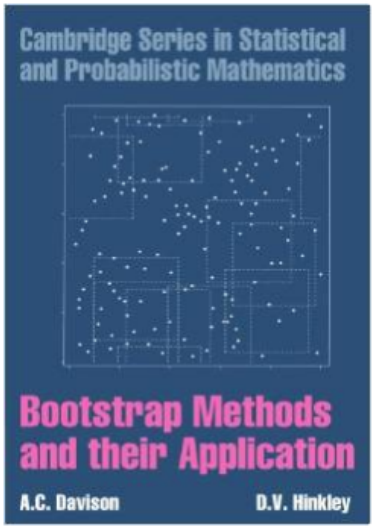
- **Objetivo**: evaluación fiable de la incertidumbre asociada a un estimador cuando:

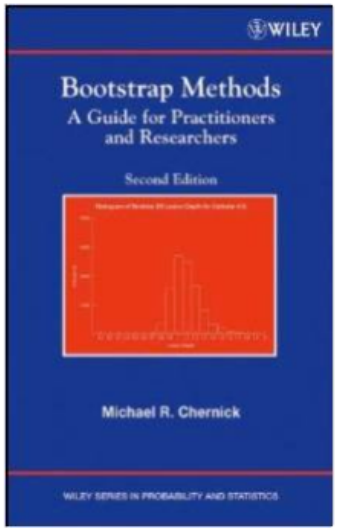
  1. el estimador es complejo;
  2. el tamaño de la muestra es pequeño;
  3. es modelo es 'non-standard'.

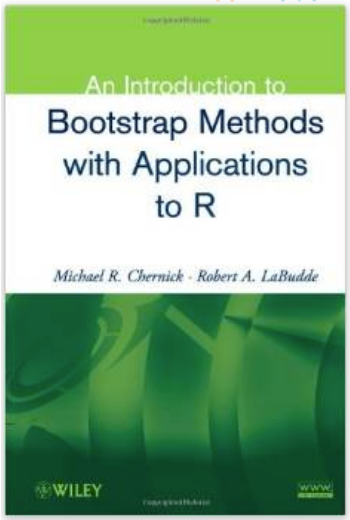
Bootstrap

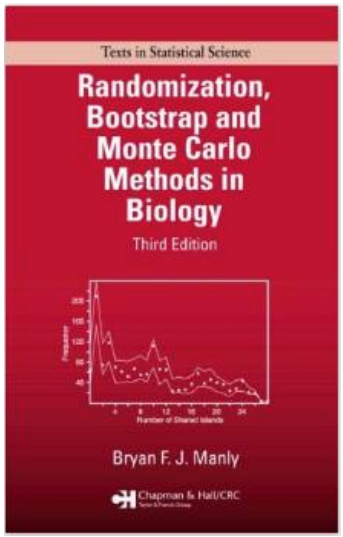
Referencias (portadas retiradas de Amazon.com)

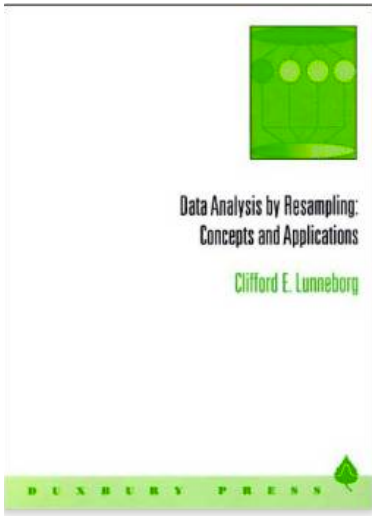












Bootstrap

Método

- Supongamos que tenemos una muestra $\mathbf{x} = (x_1, \ldots, x_n)$ de una distribución desconocida F.
- Supongamos que tenemos un estimador $\hat{\theta}(\mathbf{x})$ de un parámetro de interés θ (puede ser un vector).
- Vamos, además a suponer que las observaciones son independientes e identicamente distribuídas (iid), pero la metodología bootstrap puede ser adaptada a situaciones más generales.

Bootstrap

Método

- Idealmente, nos gustaría de tener más muestras de F.
- Así, podríamos evaluar el estimador también para estas muestras.
- Podríamos entonces obtener, por ejemplo, la varianza muestral del estimador y usar este valor como una estimación de la verdadera varianza del estimador.
- Desafortunadamente, frecuentemente, estamos limitados por cuestiones prácticas (tiempo, costos, etc) a tener simplemente una muestra.
- En este caso, usamos la única muestra que tenemos y hacemos bootstrap para estimar la varianza del estimador (y/o el sesgo, intervalos de confianza, etc).

Bootstrap

Método

- Hay dos tipos de bootstrap: el paramétrico y el no paramétrico.
- En ambos casos, reemplazamos la función de distribución que es desconocida, F, por una estimación $\widehat{F}$ obtenida.
- En el caso paramétrico, estimamos θ (máxima verosimilitud, método de los momentos, etc) y hacemos $F(x) = F(x; \widehat{\theta})$.
- En el caso no paramétrico, reemplazamos F por la función de distribución empirica de la muestra original $x_1, \ldots, x_n$

$$\hat{F}_{\text{emp}}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x).$$

Bootstrap

Bootstrap paramétrico versus bootstrap no paramétrico

- Es el dilema clásico y la respuesta es también la clásica.
- Si el modelo paramétrico es una muy buena descripción de los datos, entonces el bootstrap paramétrico debería proporcionar estimaciones más precisas que el bootstrap no paramétrico.
- Particularmente, si n es pequeño, las estimación empírica de F puede ser problemática.
- Por otra parte, el bootstrap no paramétrico es menos sensible a la especificación del modelo.

Bootstrap

Algoritmo

- Para $b = 1, \ldots, B$:
  1. generar una muestra bootstrap $x_1^{(b)}, \ldots, x_n^{(b)} \sim \widehat{F}$;
  2. calcular $\widehat{\theta}^{(b)}$ usando $x_1^{(b)}, \ldots, x_n^{(b)}$.
- Output después de B iteraciones

$$\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(B)}.$$

Bootstrap

Estimación del sesgo

- Las replicas bootstrap $\hat{\theta}^{(b)}$ son usadas para estimar propiedades de $\hat{\theta}$.
- El sesgo de un estimador $\hat{\theta}$ para θ está dado por

$$\text{sesgo}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta) = \mathbb{E}(\hat{\theta}) - \theta.$$

- Para una determinada muestra $\mathbf{x} = (x_1, \ldots, x_n)$, el estimador es $\hat{\theta}(\mathbf{x})$ y tenemos B iid estimadores, $\hat{\theta}^{(b)}$.
- La media muestral de las replicas $\{\hat{\theta}^b\}$ es no sesgada para su valor esperado $\mathbb{E}(\hat{\theta}^*)$, y así la estimación bootstrap del sesgo es

$$\text{sesgo}_{\text{boot}}(\hat{\theta}) = \overline{\widehat{\theta^*}} - \widehat{\theta},$$

con $\overline{\widehat{\theta^*}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{(b)}$ y $\hat{\theta} = \hat{\theta}(\mathbf{x})$ es la estimación basada en la muestra original.

- Note que en la metodología bootstrap, F es reemplazada por una estimación $\widehat{F}$, y así sustituimos θ por $\hat{\theta}$ para estimar el sesgo.

Bootstrap

Estimación de la varianza/ intervalos de confianza

- La varianza bootstrap del estimador $\hat{\theta}$ es

$$v_{\text{boot}} = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \overline{\hat{\theta}^{*}})^2.$$

- Una manera simple de calcular intervalos de confianza para $\hat{\theta}$ se basa en los quantiles empiricos de

$$\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}.$$

Bootstrap

Puntos Clave

- El estimador es algoritmico
  - aplicado a la muestra original $x_1, \ldots, x_n$ origina el estimador original $\hat{\theta}$;
  - aplicado a las muestras bootstrap $x_1^{(b)}, \ldots, x_n^{(b)}$ proporciona $\widehat{\theta}^{(b)}$;
  - $\widehat{\theta}$ puede ser (casi) de cualquier complejidad.
- La muestra es usada para estimar F
  - $\widehat{F} \approx F$ — suposición fuerte!
- Simulación reemplaza cálculos teóricos
  - elimina necesidad de habilidad matemática,
  - no elimina necesidad de pensamiento;
  - verificar el código cuidadosamente.
- Dos fuentes de error
  - estadística ($\widehat{F} \neq F$);
  - simulación ($B \neq \infty$).

Bootstrap

Ejemplos

Ejemplo

Suponga que los siguientes datos son una muestra aleatoria de una distribución Poisson con media λ

$$7 \quad 5 \quad 3 \quad 1 \quad 3 \quad 5 \quad 9 \quad 2 \quad 8 \quad 4 \quad 8 \quad 6$$

(a) *Determine la verdadera media y varianza del estimador de máxima verosimilitud de λ, $\hat{\lambda}$, y así determine el sesgo y el error estándar asociados a dicho estimador.*

(b) *Use un procedimiento bootstrap, parametrico y no parametrico, para estimar el sesgo y el error estándar de $\hat{\lambda}$. Además, construya también intervalos de confianza bootstrap.*

Bootstrap

Ejemplos

Ejemplo

- *La función de masa de probabilidad de la distribución Poisson está dada por*

$$f(x \mid \lambda) = \frac{e^{-x}\lambda^{x}}{x!}.$$

- *Así, la verosmilitud es*

$$L(\lambda) = \prod_{i=1}^{n} \left\{ \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \right\} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} xi}}{\prod_{i=1}^{n} x_i!},$$

mientras la log verosimilitud está dada por

$$\ell(\lambda) = \log L(\lambda) = -n\lambda + \log(\lambda)\sum_{i=1}^{n} x_i - \log\left(\prod_{i=1}^{n} x_i!\right)$$

Bootstrap

Ejemplos

Ejemplo

- *La función score (derivada da la log verosimilitud) es*

$$s(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i.$$

- *Resolvendo,* $s(\lambda) = 0$*, obtenemos* $\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$.

- *Comprobemos que la derivada de la función score (segunda derivada de la log verosimilitud) es menor que cero cuando evaluada en* $\bar{x}$

$$\frac{d}{d\lambda}s(\lambda) = -\frac{1}{\lambda^2}\sum_{x_i} \Rightarrow \frac{d}{d\lambda}s(\bar{x}) = -\frac{n}{\bar{x}} < 0 \Rightarrow \hat{\lambda} = \bar{X} \text{ es máximo.}$$

Bootstrap

Ejemplos

Ejemplo

- *Sesgo de* $\hat{\lambda}$

$$\begin{aligned} sesgo(\hat{\lambda}) = sesgo(\bar{X}) = \mathbb{E}(\bar{X}) - \lambda = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \lambda \\ = \mathbb{E}(X) - \lambda \quad (los\ X_i\ son\ iid) \\ = 0 \quad (X \sim Poisson(\lambda) \Rightarrow \mathbb{E}(X) = \lambda). \end{aligned}$$

$\hat{\lambda}$ *es un estimador no sesgado de* μ.

- *Varianza de* $\hat{\lambda}$

$$\begin{aligned} var(\hat{\lambda}) = var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{var(X)}{n} \quad (los\ X_i\ son\ iid) \\ = \frac{\lambda}{n} \quad (X \sim Poisson(\lambda) \Rightarrow var(X) = \lambda) \end{aligned}$$

Bootstrap

Ejemplos

Ejemplo

- *Así, el error estándar de* $\hat{\mu}$ *es*

$$se(\hat{\lambda}) = \sqrt{\frac{\lambda}{n}} \Rightarrow \widehat{se}(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{n}} = \sqrt{\frac{\bar{x}}{n}}.$$

- *Para los datos anteriores, tenemos* $\hat{\lambda} = 5.083$ *y* $\widehat{se}(\hat{\lambda}) = 0.651$.

- *Aproximando la distribuición de Poisson por la distribución normal y usando el teorema del limite central, tenemos que un intervalo de confianza (aproximado) a* 95% *para* λ *está dado por*

$$\left(\hat{\lambda} - 1.96\widehat{se}(\hat{\lambda}), \hat{\lambda} + 1.96\widehat{se}(\hat{\lambda})\right),$$

  que reemplazando por lo que fue obtenido anteriormente, resulta en $\lambda \in (3.808, 6.359)$.

- *Sin embargo, hace poco sentido usar el teorema del limite central para una muestra de tamaño* $n = 12$.

Bootstrap

Ejemplos

Ejemplo

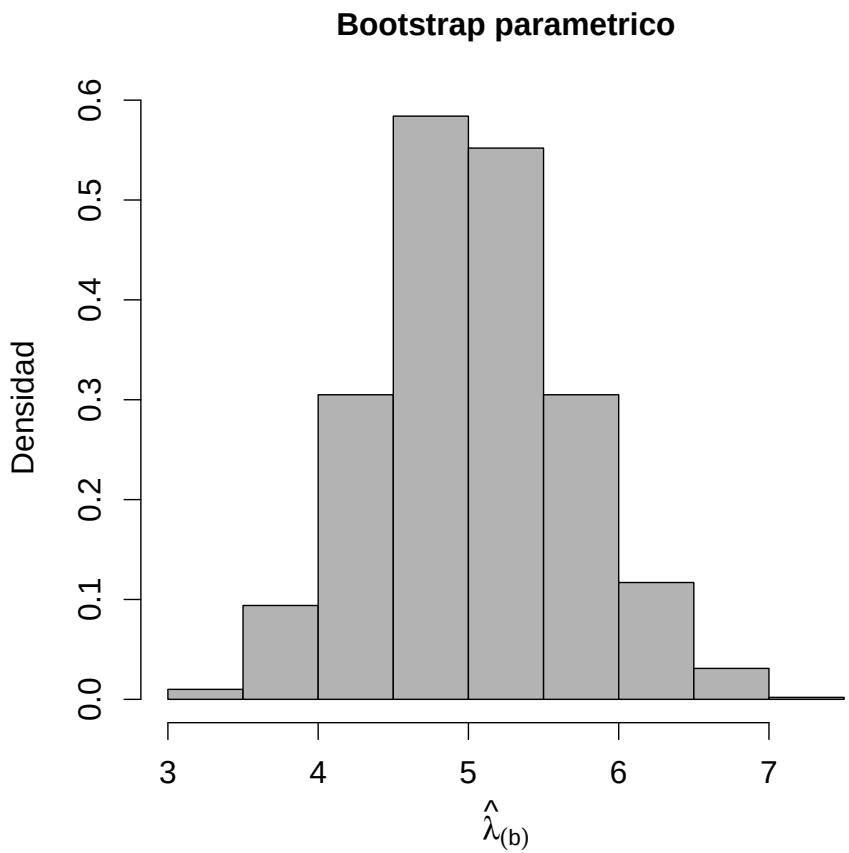
- *Vamos ahora emplear la metodología bootstrap. Empezemos con el bootstrap paramétrico.*
- *Para este ejemplo, el bootstrap parametrico consiste en simular muestras de una distribución Poisson de parámetro* $\hat{\lambda}$.

```
B=2000; lambdabp=numeric(B); lambdaest=mean(x); set.seed(123)
for(i in 1:B){
xbp=rpois(n,lambdaest)
lambdabp[i]=mean(xbp)
}
sesgop=mean(lambdabp)-lambdaest; sebp=sd(lambdabp);
icp=quantile(lambdabp,c(0.025,0.975))
```

- *Obtenemos: sesgo* $= -0.01408$, *se* $= 0.6545$ *y* $\lambda \in (3.8333, 6.4167)$.

Bootstrap

Ejemplos



Bootstrap

Ejemplos

Ejemplo

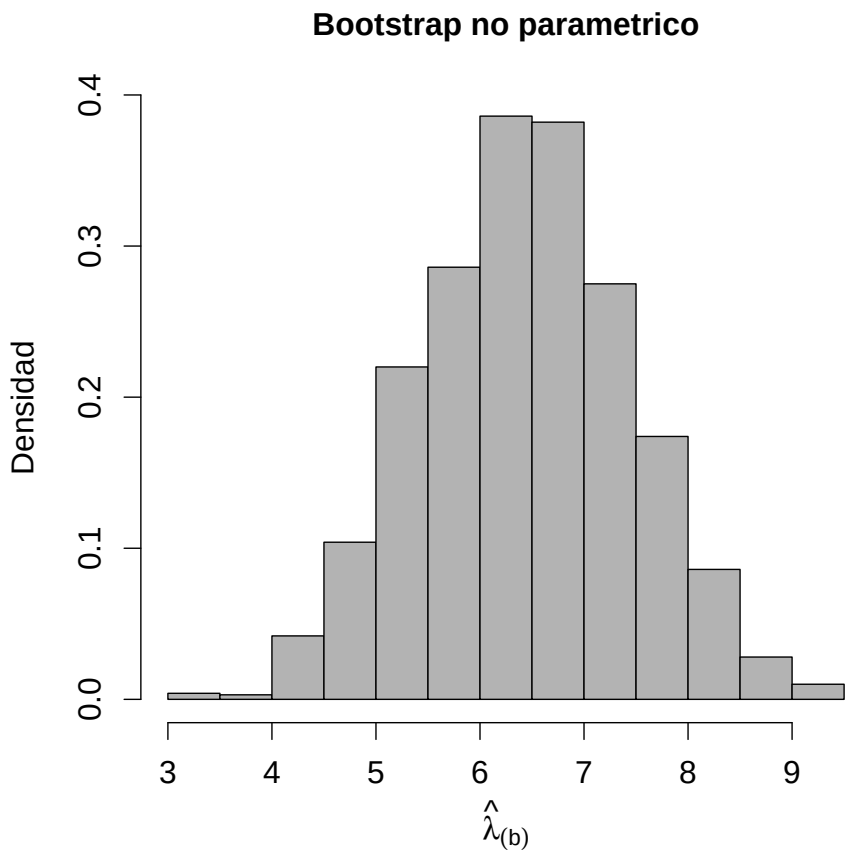
- *Para el bootstrap no paramétrico, tenemos que muestrear desde la función de distribución empirica de los datos.*
- *Remuestreando desde los datos presentados anteriormente, elegimos* $3, 5,$, *y* 8 *con probabilidad* $2/12$ *y* $1, 2, 4, 6, 7,$ *y* 9 *con probabilidad* $1/12$.
- *En* `R`, *el comando* `sample` *nos permite muestrear desde la función de distribución empirica de los datos observados.*

```
B=2000; lambdabnp=numeric(B); lambdaest=mean(x); set.seed(123)
for(i in 1:B){ xbnp=sample(1:n,size=n,replace=TRUE)
lambdabnp[i]=mean(xbnp)
}
sesgonp=mean(lambdabp)-lambdaest; sebnp=sd(lambdabp);
icnp=quantile(lambdabp,c(0.025,0.975))
```

- *Obtenemos: sesgo* $= 1.3967$, *se* $= 0.9979$ *y* $\lambda \in (4.5833, 8.41667)$.
- *Las estimaciones obtenidas son distintas a las obtenidas en el caso paramétrico, lo que se debe posiblemente al reducido tamaño de la muestra.*

Bootstrap

Ejemplos

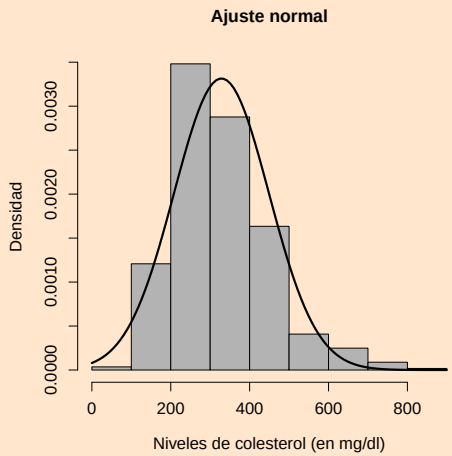


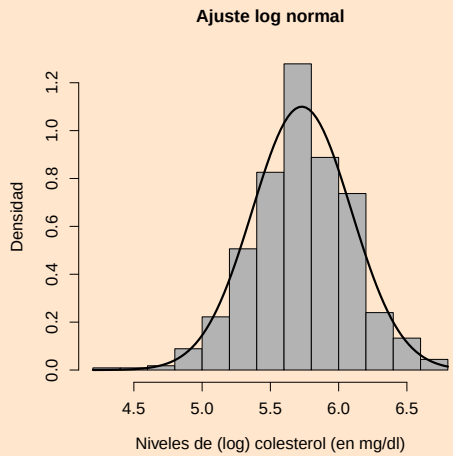
Bootstrap

Ejemplos

Ejemplo

- *Vamos a analisar datos de níveles de colesterol de 536 empleados bancarios en Belgica (datos usados en otro contexto en el libro Bayesian Biostatistics, Wiley, 2012).*
- *Supongamos que el objetivo es estimar y evaluar la incertidumbre asociada al tercer cuartil (percentil 0.75) de los niveles de colesterol.*
- *Usemos el bootstrap paramétrico y el bootstrap no paramétrico.*
- *Empezemos con el bootstrap paramétrico. Necesitamos de elegir una distribución para los datos.*





Bootstrap

Ejemplos

Ejemplo

- *El ajuste de un modelo normal a los datos transformados para la escala logartmica, aún que no sea perfecto, parece razonable.*
- *Trabajaremos entonces con los datos en la escala logarítmica.*

```
data=read.table(file="cholesterol.txt",dec=",",header=TRUE)
chol=data[,1]; n=length(chol); lchol=log(chol)

q75=quantile(lchol,0.75); B=2000; q75bp=numeric(B); set.seed(123)
for(i in 1:B){
cholbp=rnorm(n,mean(lchol),sd(lchol))
q75bp[i]=quantile(cholbp,0.75)
}

sep=sd(q75bp); sesgop=mean(q75bp)-q75
icp=quantile(q75bp,c(0.025,0.975))
```

- *Obtenemos que el error estándar estimado del tercer cuartil es igual a 0.0209, el sesgo estimado es -0.0044 y un intervalo de confianza basado en los percentiles empiricos es* $(5.9339, 6.0158)$.

Bootstrap

Ejemplos

Ejemplo

- *Apliquemos ahora la metodología bootstrap no paramétrica.*

```
q75=quantile(lchol,0.75); B=2000; q75bnp=numeric(B); set.seed(123)
for(i in 1:B){ cholbnp=sample(lchol,size=n,replace=TRUE)
q75bnp[i]=quantile(cholbnp,0.75)
}

senp=sd(q75bnp); sesgonp=mean(q75bnp)-q75
icnp=quantile(q75bnp,c(0.025,0.975))
```

- *Obtenemos que el error estándar estimado del tercer cuartil es igual a 0.0190, el sesgo estimado es -0.0004 y un intervalo de confianza basado en los percentiles empiricos es* $(5.9432, 6.0117)$.

- *En este ejemplo, debido al tamaño de la muestra y al hecho de que el ajuste normal es razonable, ambos enfoques dan resultados muy similares.*

Bootstrap

Ejemplos

Ejemplo

En una instalación de bombillas eléctricas, todas las bombillas están planeadas de ser reemplazadas regularmente después de 1200 horas. En orden de formar una opinión acerca de esta estrategia, la probabilidad de que una bombilla sobreviva esa cantidad de horas es de interés. Un test limitado dio los siguientes 20 valores de tiempo de vida

1354	*1552*	*1766*	*1325*	*2183*
1354	*1299*	*627*	*695*	*2586*
2420	*71*	*2195*	*1825*	*159*
1577	*3725*	*884*	*1014*	*965*

Construya un intervalo de confianza bootstrap para el problema de sobrevivencia de bombillas eléctricas.

Bootstrap

Ejemplos

Ejemplo

- *La probabilidad referida es*

$$\hat{p} = \frac{1}{20} \sum_{i=1}^{20} I(x_i > 1200) = 0.65,$$

donde x_i representa el tiempo de vida de la i-ésima bombilla.

- *Para la construcción del intervalo de confianza iremos adoptar un procedimiento bootstrap no paramétrico.*

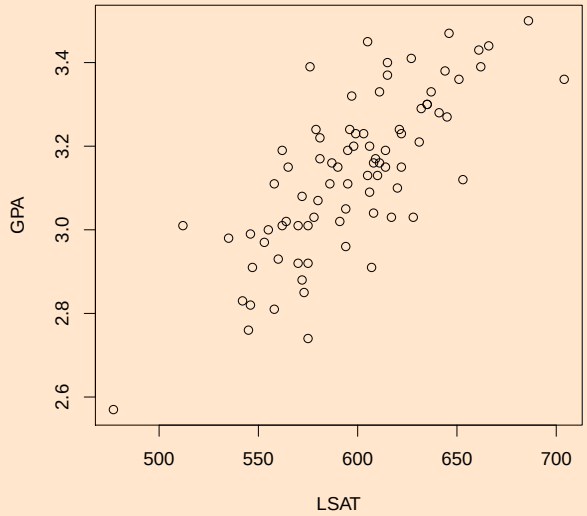
```
phat=mean(x>1200)

B=2000; phatb=numeric(B); set.seed(123)
for(i in 1:B){
xb=sample(x,size=n,replace=TRUE)
phatb[i]=mean(xb>1200)
}

quantile(phatb,c(0.025,0.975))
```

Bootstrap

Ejemplos

Ejemplo

- *Vamos a analisar el conjunto de datos* `law82` *presente en el paquete de* `R bootstrap`*.*
- *El conjunto de datos contiene las variables LSAT (promedio de las calificaciones en las pruebas de admisi´on a una facultad de derecho) y GPA (promedio de calificaciones de pregrado) para un universo de 82 escuelas.*
- *El objetivo es evaluar la correlación (y su incertidumbre) entre las variables LSAT y GPA.*



Bootstrap

Ejemplos

Ejemplo

- *Iremos utilizar un procedimiento bootstrap no paramétrico.*
- *Una idea que parece natural es muestrear, independientemente, las muestras correspondientes a las variables LSAT y GPA.*
- *Sin embargo, esta estrategia iria destruir la dependencia entre las variables LSAT y GPA.*
- *Así, una posible estrategia es muestrear un índice, que en este caso varía de 1 a 82, y muestrear el par (LSAT,GPA) correspondiente al índice muestreado.*
- *La correlación estimada es aproximadamente 0.76.*

Bootstrap

Ejemplos

```
require(bootstrap)
lsat=law82$LSAT; gpa=law82$GPA; n=length(lsat)
corr=cor(lsat,gpa); B=2000; corb=numeric(B); set.seed(123)
for(i in 1:B){
ind=sample(1:n,size=n,replace=TRUE)
lsatb=lsat[ind]; gpab=gpa[ind]
corb[i]=cor(lsatb,gpab)
}

sesgo=mean(corb)-corr; varb=var(corb); quantile(corb,c(0.025,0.975))
```

Bootstrap

Bootstrap en modelos de regresión

- Considere el modelo de regresión lineal simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

- Una posibilidad para hacer bootstrap en el modelo de regresión presentado y así obtener estimaciones del sesgo/ varianza/ intervalos de confianza de los parámetros es considerar $z_i = (y_i, x_i)$, el vector de la variable respuesta y regresor de la i-ésima observación.

- Las observaciones $(z_1, \ldots, z_n)$ pueden ser remuestreadas y β_0 y β_1 pueden ser calculados para cada una de las muestras bootstrap $(z_1^{(b)}, \ldots, z_n^{(b)})$ produciendo B conjuntos de parámetros bootstrap $(\beta_0^{(b)}, \beta_1^{(b)})$.

- Con base en estos B conjuntos de parámetros se puede calcular el sesgo, varianza, intervalos de confianza, etc, de cada uno de los parámetros.

Bootstrap

Bootstrap en modelos de regresión

```
# simular dados desde un modelo de regresion
# el verdadero valor de beta0 es 1.5, beta1 es 2, y la varianza de los
residuos es 1

n=500; x=runif(n,0,1); eps=rnorm(n,0,1)
y=1.5+2*x+eps
fit=lm(formula=y x); summary(fit)
beta=as.numeric(lm(formula=y x)$coefficients)

B=1000; betabnp=matrix(0,nrow=2,ncol=B)
for(j in 1:B){
ind=sample(1:n,size=n,replace=TRUE)
xb=x[ind]; yb=y[ind]
betabnp[,j]=as.numeric(lm(formula=yb xb)$coefficients)
}

mean(betabnp[1,]); sd(betabnp[1,]); quantile(betabnp[1,],c(0.025,0.975))
mean(betabnp[2,]); sd(betabnp[2,]); quantile(betabnp[2,],c(0.025,0.975))
```

Bootstrap

Bootstrap en modelos de regresión

- Otra posibilidad también muy utilizada para hacer bootstrap en modelos de regresión consiste en hacer bootstrap de los resíduos.

- El procedimiento es el siguiente:

  1. Estimar $\hat{\beta}_0$, $\hat{\beta}_1$, y $\hat{\sigma}^2$ con base en la muestra original y calcular los residuos

$$\hat{\varepsilon}_i = \frac{y_i - \beta_0 - \beta_1 x_i}{\hat{\sigma}}, \quad i = 1, \ldots, n,$$

     con

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}.$$

  2. Obtener una muestra de los resíduos, $(\hat{\varepsilon}_1^{(b)}, \ldots, \hat{\varepsilon}_n^{(b)})$, y a partir de ellos calcular

$$y_i^{(b)} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i^{(b)}, \quad i = 1, \ldots, n.$$

  3. Usar la muestra $\{(x_i, y_i^{(b)})\}$ y estimar $(\beta_0^{(b)}, \beta_1^{(b)})$.

Bootstrap

Bootstrap en modelos de regresión

```
# usamos los mismos dados generados en el bootstrap anterior.
# como primero enfoque, simularemos una muestra de los resíduos
asumiendo para ellos una dist. normal.

se=sqrt(sum((y-beta[1]-beta[2]*x)**2)/(n-2))
betabp=matrix(0,nrow=2,ncol=B)
for(i in 1:B){
resp=rnorm(n,0,se)
ybp=beta[1]+beta[2]*x+resp
betabp[,i]=as.numeric(lm(formula=ybp x)$coefficients)
}

mean(betabp[1,]); sd(betabp[1,]); quantile(betabp[1,],c(0.025,0.975))
mean(betabp[2,]); sd(betabp[2,]); quantile(betabp[2,],c(0.025,0.975))
```

Bootstrap

Bootstrap en modelos de regresión

```
# ahora no asumimos ninguna distribución para los residuos y
muetsreamos desde su funcion de dist. emprica.

res=(y-beta[1]-beta[2]*x)/se
betabsp=matrix(0,nrow=2,ncol=B)
for(i in 1:B){
ressp=sample(res,size=n,replace=TRUE)
ybsp=beta[1]+beta[2]*x+ressp
betabsp[,i]=as.numeric(lm(formula=ybsp x)$coefficients) }

mean(betabsp[1,]); sd(betabsp[1,]); quantile(betabsp[1,],c(0.025,0.975))
mean(betabsp[2,]); sd(betabsp[2,]); quantile(betabsp[2,],c(0.025,0.975))
```