

# Regresión lineal simple con Bootstrap

Paula Saavedra Nieves

Silvia Suárez Crespo

10 de marzo de 2010

Los modelos de regresión son construidos para representar la dependencia de una variable respuesta,  $Y$ , respecto a otra variable explicativa,  $X$ . En términos generales, la regresión se suele formalizar como la media condicionada de la variable respuesta en función del valor que tome la variable explicativa. Se trataría, pues, de la función siguiente:

$$m(x) = \mathbb{E}(Y|X = x) \text{ para cada posible valor } x \text{ de } X.$$

En consecuencia, podemos descomponer la variable respuesta en función del resultado de  $X$ , más un error de media cero:

$$Y = m(X) + \varepsilon,$$

donde  $\varepsilon$  se conoce como error, verificando  $\mathbb{E}(\varepsilon|X = x) = 0$  para todo  $x$ .

Un caso particular de modelo de regresión, lo constituye el modelo lineal simple que supone variable respuesta y explicativa univariantes y las siguientes hipótesis básicas:

1. **Linealidad.** La función de regresión es una línea recta. Por tanto,

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

donde  $\beta_0$  y  $\beta_1$  son parámetros, en principio desconocidos estimables en base a una muestra, y  $\varepsilon$  es una variable aleatoria no observable, que llamaremos error, y que contiene la variabilidad no achacable a la variable explicativa sino debida a errores de medición u otros factores no controlables.

Esta hipótesis nos sitúa en el contexto paramétrico, dado que supone que la función de regresión es una recta pero deja libertad al valor concreto de la pendiente y la ordenada en el origen. Por supuesto, en la Estadística se han estudiado otros modelos que no requieren suposición paramétrica alguna, conocidos como métodos no paramétricos.

2. **Homocedasticidad.** La varianza del error es la misma cualquiera que sea el valor de la variable explicativa:

$$\text{Var}(\varepsilon|X = x) = \sigma^2 \text{ para todo } x.$$

3. **Normalidad.** El error tiene distribución normal

$$\varepsilon \in N(0, \sigma^2).$$

4. Se distinguen dos tipos de diseño experimental según la naturaleza de la muestra de partida:

- a) **Diseño fijo.** Los valores de la variable explicativa están fijados por el experimentador, de acuerdo a un diseño conveniente de cara a la viabilidad del experimento o a su eficiencia estadística. En este caso los valores de la variable explicativa no son aleatorios, y sólo es aleatorio el error y, en consecuencia, la variable respuesta. Así, la muestra resultante de un diseño fijo sería del tipo:

$$(x_1, Y_1), \dots, (x_n, Y_n).$$

- b) **Diseño aleatorio.** En este caso tanto la variable explicativa como la variable respuesta son aleatorias. Por tanto, la muestra resultante de un diseño aleatorio sería del tipo:

$$(X_1, Y_1), \dots, (X_n, Y_n).$$

5. **Independencia.** Las variables aleatorias que representan los errores  $\varepsilon_1, \dots, \varepsilon_n$  son mutuamente independientes.

Así, bajo el modelo lineal simple, si suponemos diseño fijo, dado un conjunto de observaciones  $(x_1, Y_1), \dots, (x_n, Y_n)$  los estimadores de mínimos cuadrados de  $\beta_0$  y  $\beta_1$  vendrán dados por las fórmulas

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{SS_x}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x},$$

donde  $SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$ . Recordemos que  $\sigma^2$  se estima insesgadamente mediante

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2, \quad i = 1, \dots, n,$$

donde

$$e_i = Y_i - \hat{\mu}_i, \quad i = 1, \dots, n$$

son los residuos y

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots, n,$$

es la estimación de  $\mu_i = \beta_0 + \beta_1 x_i$ ,  $i = 1, \dots, n$ . Es conocido que

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SS_x}.$$

Además, bajo las hipótesis anteriores, si los errores siguen una distribución normal:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}(\hat{\beta}_1)},$$

sigue una distribución t de Student con  $n - 2$  grados de libertad, siendo

$$\hat{\sigma}^2(\hat{\beta}_1) = \frac{s^2}{SS_x}.$$

Así, el intervalo de confianza para  $\beta_1$  de nivel  $(1 - \alpha)$  vendrá dado por

$$IC = \left[ \hat{\beta}_1 - c_u \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 - c_l \hat{\sigma}_{\hat{\beta}_1} \right],$$

donde

$$\mathbb{P}(c_l = -t_{n-2, \alpha/2} \leq T_{n-2} \leq c_u = t_{n-2, \alpha/2}) = 1 - \alpha.$$

Ahora bien, si la hipótesis de normalidad no es cierta, los valores de  $c_l$  y  $c_u$  pueden ser diferentes de los valores críticos de una t de Student. Por supuesto, por el teorema central del límite, esto no ocurrirá si la muestra es grande; sin embargo, para muestras pequeñas y con errores claramente no normales el intervalo anterior puede no ser adecuado. En esta situación puede ser útil utilizar el bootstrap para aproximar la distribución de T. ¿Cómo podemos generar la muestra bootstrap? Una primera idea, si los errores  $\varepsilon_i$  siguen una distribución normal, sería seleccionar los errores bootstrap  $\varepsilon_1^*, \dots, \varepsilon_n^*$  aleatoriamente según una  $N(0, s^2)$  y generar  $y_i^*$  mediante la fórmula

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i^*, \quad i = 1, \dots, n.$$

Esta idea se puede seguir usando en un contexto no paramétrico. Para ello necesitamos tener una buena aproximación de la distribución de los errores. Los valores de los residuos  $\{e_1, \dots, e_n\}$  nos dan una idea de esa distribución. Sin embargo, su distribución no es del todo fiel a la de los errores originales ya que, por ejemplo, su varianza no es constante. Se tiene que

$$\text{Var}(e_i) = \sigma^2(1 - h_i),$$

donde

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_x}, \quad i = 1, \dots, n.$$

Para corregir este problema se construyen los residuos modificados

$$r_i = \frac{e_i}{(1 - h_i)^{1/2}}, \quad i = 1, \dots, n.$$

Estos residuos ya tienen varianza constante  $\sigma^2$ , como los errores  $\varepsilon_i$ . Sin embargo no tienen media cero. Por ello los errores bootstrap se escogen al azar del conjunto  $\{r_1 - \bar{r}, \dots, r_n - \bar{r}\}$ . El plan de remuestreo bootstrap para construir un intervalo de confianza para  $\beta_1$  será el siguiente:

1. Para  $i = 1, \dots, n$ 
  - a) Poner  $x_i^* = x_i$ .
  - b) Seleccionar al azar  $\varepsilon_i^*$  del conjunto  $\{r_1 - \bar{r}, \dots, r_n - \bar{r}\}$ .
  - c) Hacer  $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i^*$ .
2. Estimar  $\hat{\beta}_0^*, \hat{\beta}_1^*$  y los residuos  $e_1^*, \dots, e_n^*$  a partir de  $(x_1^*, y_1^*), \dots, (x_n^*, y_n^*)$ .
3. Evaluar  $T^*$  en la muestra bootstrap. Para cada muestra bootstrap se obtiene

$$t^* = \frac{\hat{\beta}_1^* - \hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1^*)}, \quad \text{donde } \hat{\sigma}^2(\hat{\beta}_1^*) = \frac{s^{*2}}{SSx}, \quad s^{*2} = \frac{1}{n-2} \sum_{i=1}^n e_i^{*2},$$

4. Repetir los pasos anteriores B veces obteniendo  $t_1^*, \dots, t_B^*$ .
5. Ordenar de menor a mayor los valores calculados de  $T^*$  y tomar el valor que ocupa la posición  $\alpha/2 * B$ ,  $c_l^*$  y el que ocupa la posición  $1 - \alpha/2 * B$ ,  $c_u^*$ . El intervalo bootstrap para  $\beta_1$  de nivel  $(1 - \alpha)$  es

$$IC_{\text{boot}} = \left[ \hat{\beta}_1 - c_u^* \hat{\sigma}(\hat{\beta}_1), \hat{\beta}_1 - c_l^* \hat{\sigma}(\hat{\beta}_1) \right].$$

Comprobaremos el funcionamiento del método anterior estudiando detenidamente un caso concreto:

1. Como valores de  $x$  tomaremos quince puntos equiespaciados en el intervalo  $[0, 1]$ ,  $x = 0, 1/15, 2/15, \dots$ , es decir,

```
0.00000000 0.06666667 0.13333333 0.20000000 0.26666667 0.33333333
0.40000000 0.46666667 0.53333333 0.60000000 0.66666667 0.73333333
0.80000000 0.86666667 0.93333333
```

2. Los parámetros  $\beta_0$  y  $\beta_1$  serán ambos iguales a 1.
3. Por último, supondremos que los errores de medida,  $\varepsilon_i$ , siguen una distribución t de Student con 3 grados de libertad:

```

0.36427863 -0.47410508 -2.47999864 -0.13398947  1.69578454  0.01615766
-0.05434797  0.13081385 -0.05699907 -4.66807710  0.97407462  0.84580509
1.18926602  1.41066207  2.06832658

```

Así, una vez fijadas las condiciones anteriores, es posible obtener los valores  $Y_i$ ,  $i = 1, \dots, 15$ , simplemente a través de la generación del modelo,  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ,  $i = 1, \dots, 15$ :

```

1.3642786  0.5925616 -1.3466653  1.0660105  2.9624512  1.3494910
1.3456520  1.5974805  1.4763343 -3.0680771  2.6407413  2.5791384
2.9892660  3.2773287  4.0016599

```

En este punto y teniendo en cuenta la muestra generada, estamos en condiciones de proceder al cálculo del intervalo bootstrap de nivel 0,95 para  $\beta_1$ . A partir del código en R recogido en el Apéndice I, obtenemos los siguientes intervalos que resumimos en el Cuadro 1:

	Ejecución 1	Ejecución 2	Ejecución 3
B=50	[ 0.757,7.098]	[ 0.175,3.837]	[-1.818,5.264]
B=250	[-0.044,6.492]	[-0.735,4.530]	[-7.935,0.438]
B=500	[-0.077,6.549]	[-0.976,6.120]	[-1.056,6.012]
B=1000	[-0.123,6.224]	[ 0.463,3.675]	[-2.512,1.789]
B=10000	[-0.268,5.928]	[-1.309,3.226]	[-2.824,3.462]

Cuadro 1: Tabla intervalos confianza nivel 0.95

Observando los resultados anteriores, es obvio que todos los intervalos obtenidos contienen a  $\beta_1$ , salvo el correspondiente a la tercera ejecución para  $B = 250$ .

Supongamos entonces construido un determinado intervalo de confianza para el parámetro de interés ( $IC$  ó  $IC_{\text{boot}}$ ). A continuación utilizaremos la metodología de Monte - Carlo para la aproximación de la cobertura de dicho intervalo, que se reduce a ser la probabilidad de que nuestro parámetro de interés  $\beta_1$  caiga dentro del mismo.

El proceso de aproximación de Monte - Carlo para este caso particular será:

1. Extraer  $M$  intervalos de confianza para  $\beta_1$  utilizando el método considerado ( $IC^{(m)}$  ó  $IC_{\text{boot}}^{(m)}$ ,  $m = 1, \dots, M$ ).
2. Aproximación por Monte - Carlo:

$$\text{Cobertura} = \frac{1}{M} \sum_{m=1}^M (\beta_1 \in IC^{(m)})$$

ó

$$\text{Cobertura} = \frac{1}{M} \sum_{m=1}^M (\beta_1 \in IC_{\text{boot}}^{(m)}).$$

Consideraremos cuatro casos posibles:

- Caso 1.** Cálculo de la cobertura del  $IC_{\text{boot}}$  para el caso concreto descrito anteriormente, manteniendo que los errores de medida  $\varepsilon_i$  siguen una distribución t de Student con 3 grados de libertad.
- Caso 2.** Cálculo de la cobertura del  $IC$  para el caso anterior.
- Caso 3.** Cálculo de la cobertura del  $IC_{\text{boot}}$  para el caso concreto descrito anteriormente, cambiando la distribución t de Student de los errores  $\varepsilon_i$  por una distribución Normal de media 0 y varianza 3.
- Caso 4.** Cálculo de la cobertura del  $IC$  para el caso anterior.

Los dos primeros casos se encuentran implementados en el Apéndice II, mientras que los dos últimos lo están en el Apéndice III. Además del cálculo de la cobertura, se acompañará en cada caso una gráfica con la representación de los intervalos obtenidos durante el proceso de Monte - Carlo, de forma que visualmente se pueda apreciar el grado de cobertura de los mismos, realizando una única ejecución.

De esta forma, los resultados obtenidos para el **Caso 1** se encuentran recogidos en el Cuadro 2. Les acompañan también las gráficas correspondiente a los intervalos obtenidos variando el índice  $M = 50, 100, 250, 500$  y  $1000$ , fijado un determinado  $B$  ( $B = 50, 250, 500$  ó  $1000$ ). Las Figuras correspondientes a cada valor de  $B$  son, respectivamente, 1, 2, 3 y 4.

	M=50	M=100	M=250	M=500	M=1000
B=50	96.0 %	88.0 %	89.6 %	91.8 %	93.0 %
B=250	98.0 %	92.0 %	94.2 %	95.0 %	94.6 %
B=500	96.0 %	97.0 %	96.0 %	95.2 %	94.8 %
B=1000	94.0 %	94.0 %	92.8 %	94.0 %	94.6 %

Cuadro 2: Cobertura de los intervalos obtenidos en el **Caso 1**.

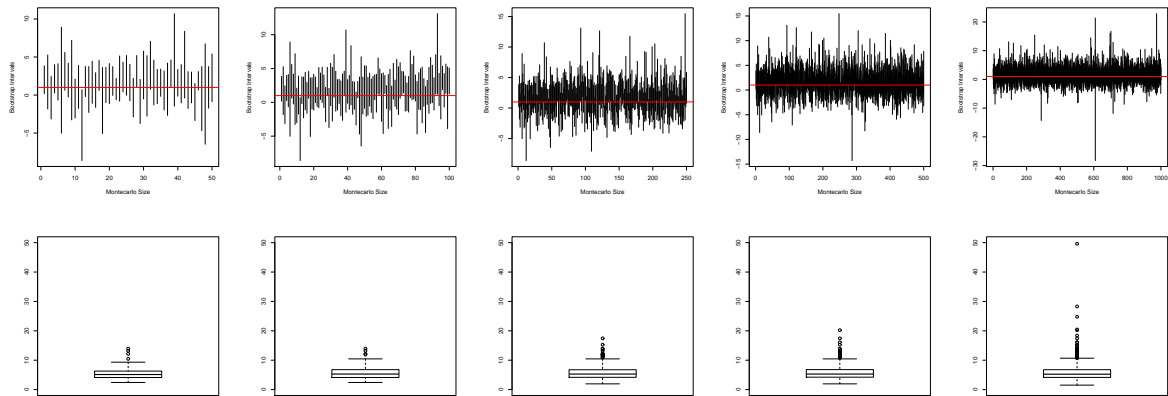


Figura 1: **Caso 1**. En la primera fila, de izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 50$ . En la segunda fila, para los  $M$  anteriores los gráficos Box-Plot de la longitud de los intervalos.

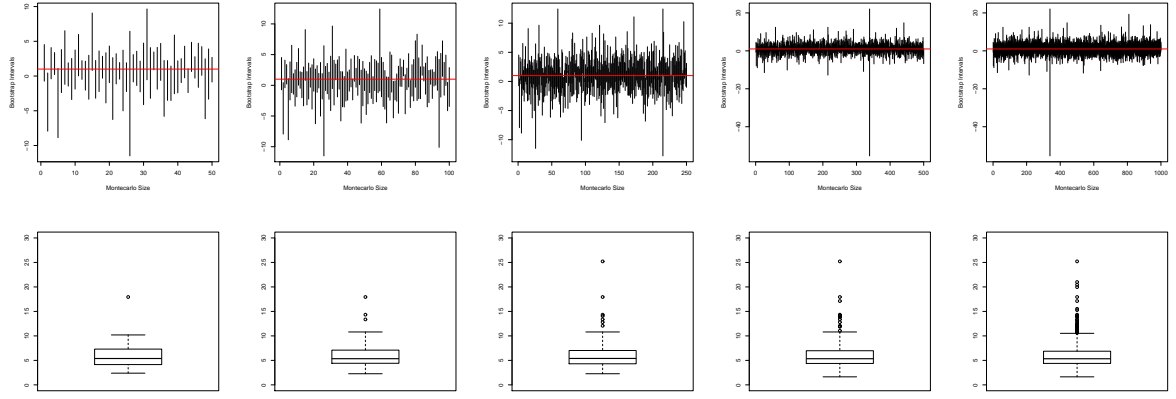


Figura 2: **Caso 1.** En la primera fila, de izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 250$ . En la segunda fila, para los  $M$  anteriores los gráficos Box-Plot de la longitud de los intervalos.

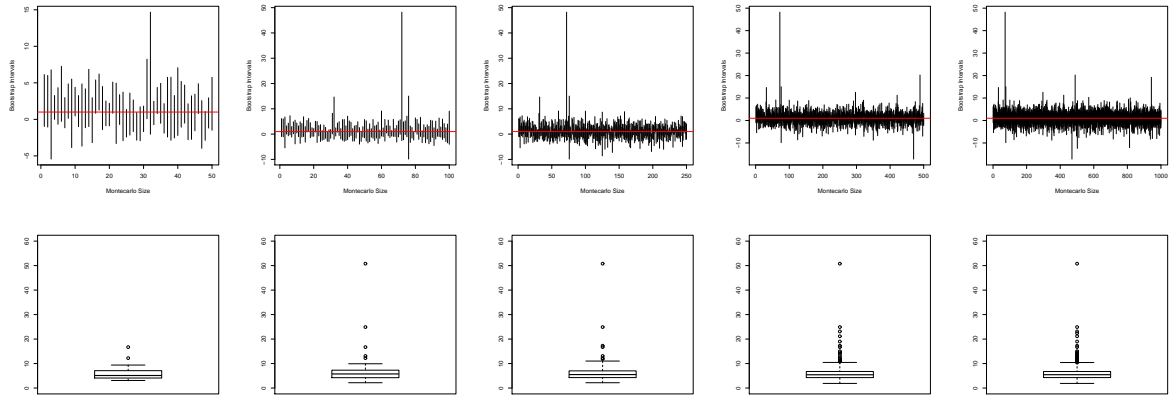


Figura 3: **Caso 1.** En la primera fila, de izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 500$ . En la segunda fila, para los  $M$  anteriores los gráficos Box-Plot de la longitud de los intervalos.



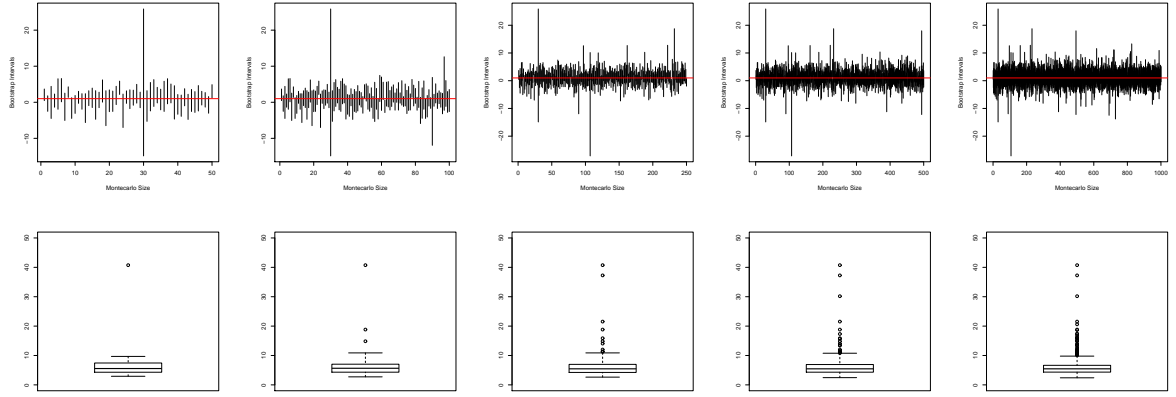


Figura 4: **Caso 1.** En la primera fila, de izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 1000$ . En la segunda fila, para los  $M$  anteriores los gráficos Box-Plot de la longitud de los intervalos.

Bajo los resultados obtenidos en el Cuadro 2 podemos concluir que prácticamente todos las coberturas rondan el 95%, a excepción de algunos casos que resultan de tomar  $B$  y/o  $M$  pequeños. A grandes rasgos se observa que el grado de cobertura aumenta con  $M$ , aunque el esfuerzo computacional que supone pasar de  $B = 500$  y  $M = 500$  a  $B = 1000$  y  $M = 1000$ , respectivamente, no se ve compensado en los resultados. Así mismo, se observa que la longitud de los intervalos en media es muy similar, aumentando puntualmente los datos atípicos al aumentar  $M$ .

En el Cuadro 3 se incluyen los resultados obtenidos para el **Caso 2**. Recordemos que en este caso lo que se realiza es la implementación de los intervalos utilizando la aproximación general de la distribución t de Student. Al igual que para el **Caso 1**, se acompañan las Figuras 5, 6, 7 y 8 correspondientes, respectivamente, a  $B = 50$ ,  $B = 250$ ,  $B = 500$  y  $B = 1000$ .

	M=50	M=100	M=250	M=500	M=1000
B=50	100.0 %	96.0 %	96.0 %	95.8 %	95.4 %
B=250	98.0 %	92.0 %	95.2 %	96.0 %	95.2 %
B=500	96.0 %	97.0 %	96.4 %	95.6 %	95.0 %
B=1000	94.0 %	95.0 %	94.2 %	94.6 %	94.7 %

Cuadro 3: Cobertura de los intervalos obtenidos en el **Caso 2**.

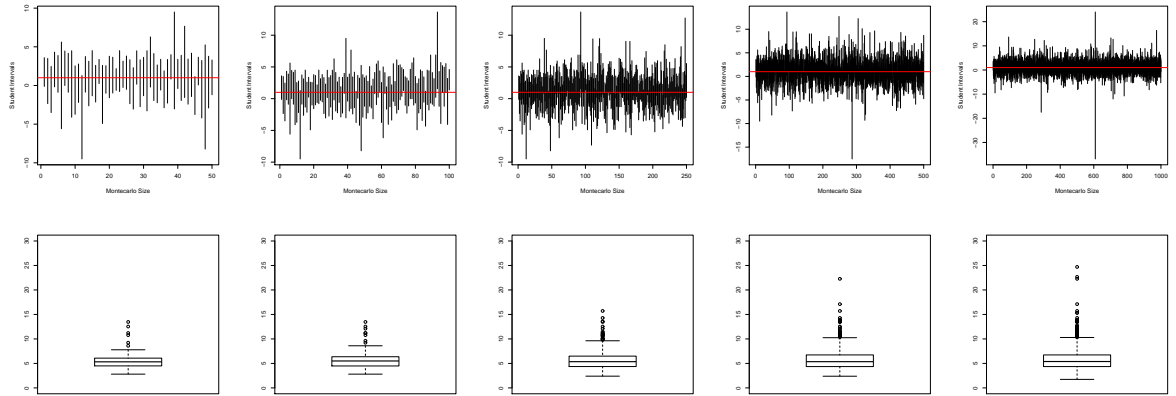


Figura 5: **Caso 2**. En la primera fila, de izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 50$ . En la segunda fila, para los  $M$  anteriores los gráficos Box-Plot de la longitud de los intervalos.

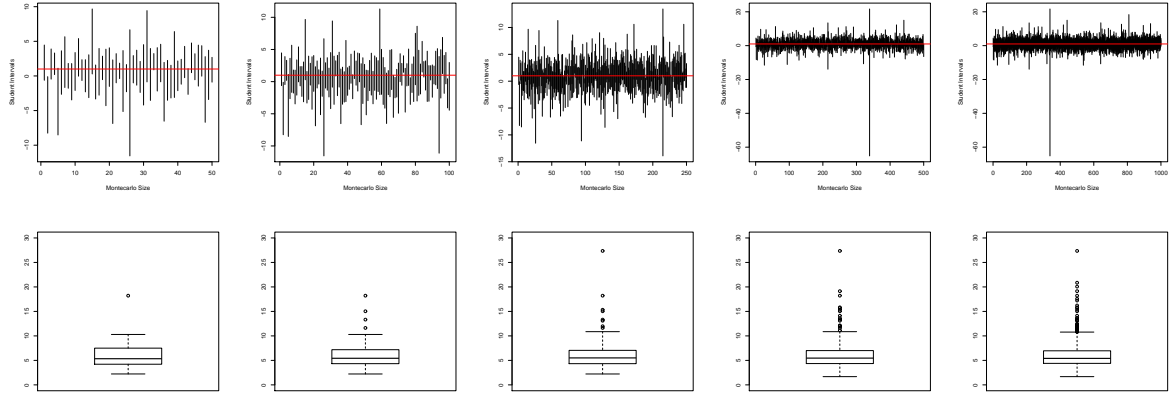


Figura 6: **Caso 2.** En la primera fila, de izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 250$ . En la segunda fila, para los  $M$  anteriores los gráficos Box-Plot de la longitud de los intervalos.

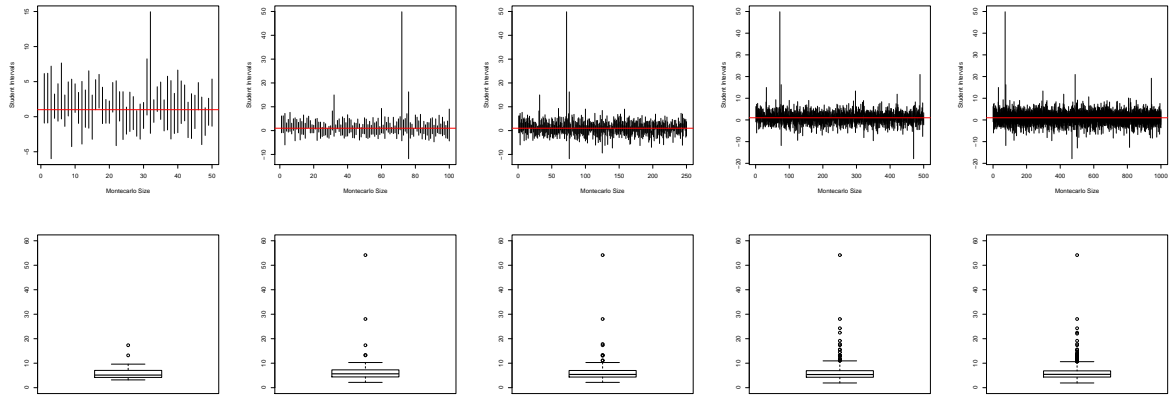


Figura 7: **Caso 2.** En la primera fila, de izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 500$ . En la segunda fila, para los  $M$  anteriores los gráficos Box-Plot de la longitud de los intervalos.

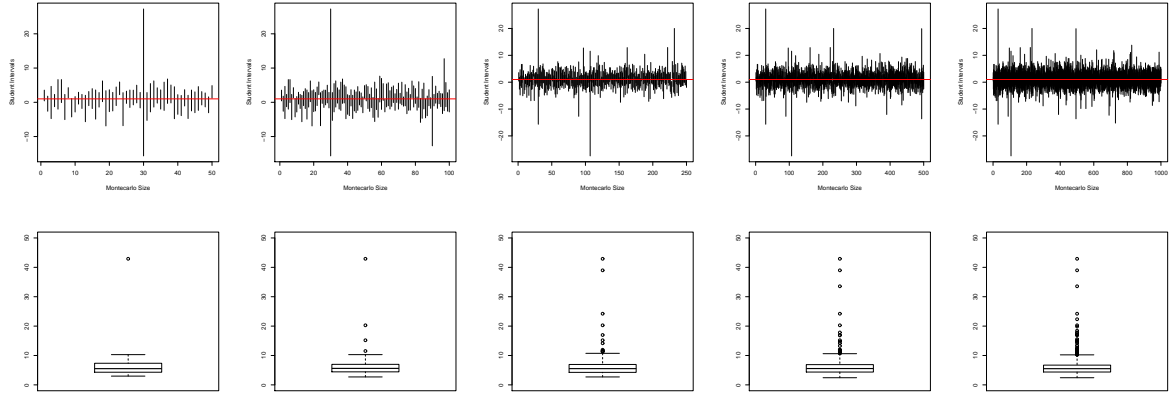


Figura 8: **Caso 2.** En la primera fila, de izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 1000$ . En la segunda fila, para los  $M$  anteriores los gráficos Box-Plot de la longitud de los intervalos.

Para este caso puede verse nuevamente lo comentado para el **Caso 1**: parece existir un balance entre los parámetros  $M$  y  $B$ , de manera que los mejores resultados de cobertura se encuentran en el centro del Cuadro 3. En cuanto a la comparación entre las coberturas obtenidas para el **Caso 1** y para el **Caso 2**, teniendo en cuenta que los errores no son normales sino que provienen de una distribución  $t$  de Student con 3 grados de libertad, se obtiene que, aunque los resultados son relativamente similares, se obtienen mejores coberturas en el **Caso 2**.

En el Cuadro 4 se obtienen las coberturas asociadas a los intervalos construídos para el **Caso 3**, y las Figuras 9, 10, 11 y 12 están asociadas, respectivamente, a  $B = 50$ ,  $B = 250$ ,  $B = 500$  y  $B = 1000$ .

	M=50	M=100	M=250	M=500	M=1000
B=50	84.0 %	86.0 %	88.4 %	92.8 %	91.7 %
B=250	98.0 %	95.0 %	96.4 %	94.4 %	94.7 %
B=500	92.0 %	98.0 %	95.6 %	93.6 %	94.4 %
B=1000	98.0 %	96.0 %	92.8 %	95.0 %	94.7 %

Cuadro 4: Cobertura de los intervalos obtenidos en el **Caso 3**.

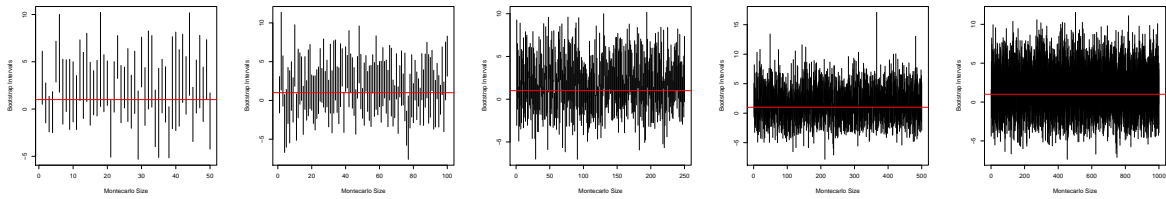


Figura 9: **Caso 3**. De izquierda a derecha, intervalos de confianza bootstrap obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 50$ .

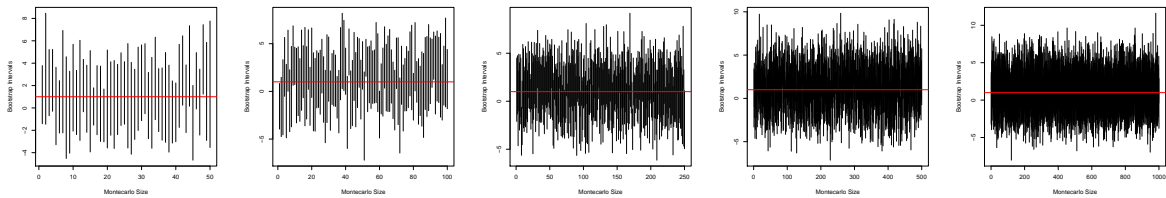


Figura 10: **Caso 3**. De izquierda a derecha, intervalos de confianza bootstrap obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 250$ .

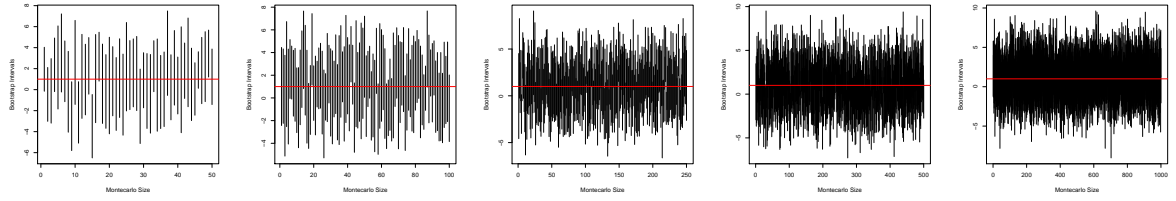


Figura 11: **Caso 3.** De izquierda a derecha, intervalos de confianza bootstrap obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 500$ .

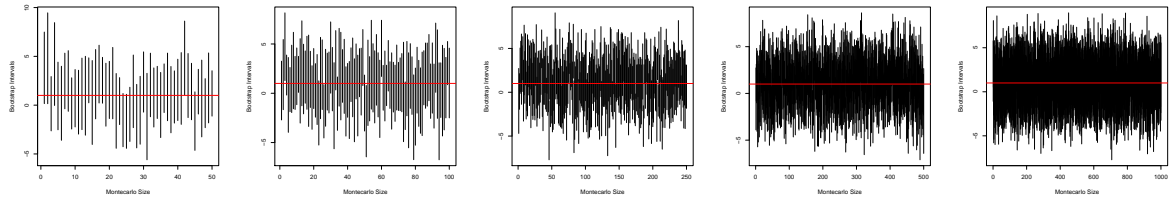


Figura 12: **Caso 3.** De izquierda a derecha, intervalos de confianza bootstrap obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 1000$ .

Al igual que ocurría en los casos anteriores, las coberturas de los intervalos rondan el 95 % excepto casos puntuales (tomando  $B = 50$ ). De la misma forma, se observa que los mejores resultados están situados en el centro del Cuadro 4.

Finalmente, en el Cuadro 5 se obtienen las coberturas asociadas a los intervalos construídos para el **Caso 4**. Al igual que en los casos anteriores, se añaden las gráficas de los intervalos, y las Figuras 9, 10, 11 y 12 son las asociadas, respectivamente, a  $B = 50$ ,  $B = 250$ ,  $B = 500$  y  $B = 1000$ .

	M=50	M=100	M=250	M=500	M=1000
B=50	86.0 %	93.0 %	93.6 %	94.8 %	94.7 %
B=250	100 %	96.0 %	96.4 %	94.8 %	95.6 %
B=500	92.0 %	99.0 %	94.8 %	93.6 %	94.4 %
B=1000	96.0 %	98.0 %	93.2 %	95.2 %	94.6 %

Cuadro 5: Cobertura de los intervalos obtenidos en el **Caso 4**.

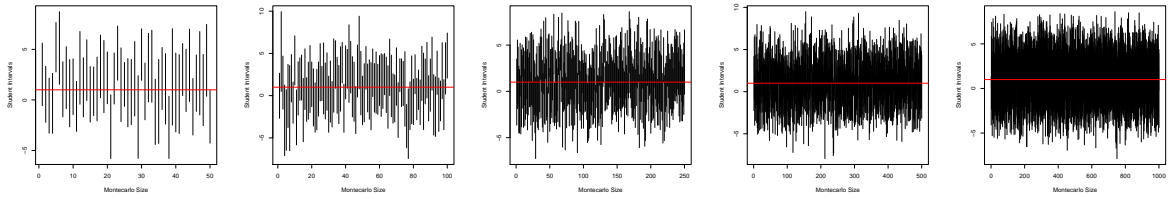


Figura 13: **Caso 4**. De izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 50$ .

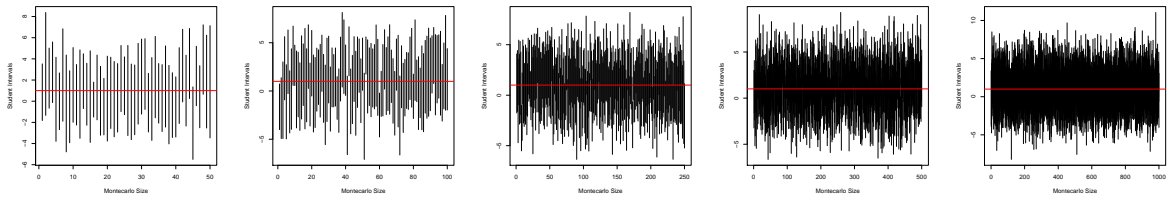


Figura 14: **Caso 4**. De izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 250$ .

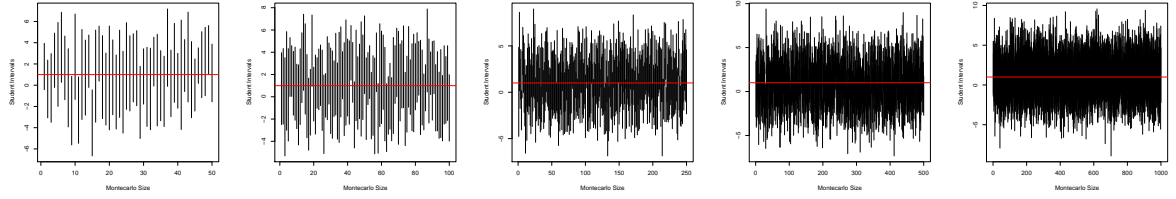


Figura 15: **Caso 4.** De izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 500$ .

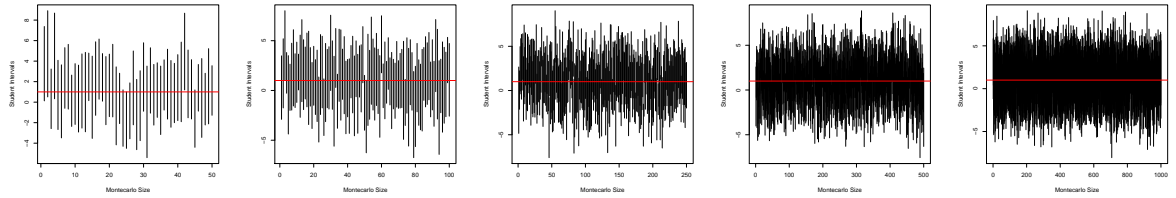


Figura 16: **Caso 4.** De izquierda a derecha, intervalos de confianza obtenidos utilizando  $M = 50$ ,  $M = 100$ ,  $M = 250$ ,  $M = 500$  y  $M = 1000$ , respectivamente, para  $B = 1000$ .

Se obtienen nuevamente coberturas cercanas al 95 %. Comparando las coberturas obtenidas para el **Caso 3** y para el **Caso 4**, obtenemos mejores resultados en el **Caso 2**.



## Apéndice

### Apéndice I. Generación del modelo y determinación intervalo bootstrap

```
#-- Set seed
```

```
set.seed(4)
```

```
#-- Parameters
```

```
Sample<-seq(0,1,by=1/15)[1:15] Sample.size=length(Sample)
```

```
Beta_0<-1 Beta_1<-1
```

```
Bootstrap_resample.size<-500
```

```
Matrix_bootstrap.errors<-matrix(0,nrow=Sample.size,ncol=Bootstrap_resample.size)
```

```
Montecarlo.size<-500
```

```
Alpha<-0.05
```

```
IC.Montecarlo.bootstrap.student.errors<-matrix(0,nrow=Montecarlo.size,ncol=2)
```

```
IC.Montecarlo.student.student.errors<-matrix(0,nrow=Montecarlo.size,ncol=2)
```

```
IC.Montecarlo.bootstrap.normal.errors<-matrix(0,nrow=Montecarlo.size,ncol=2)
```

```
IC.Montecarlo.student.normal.errors<-matrix(0,nrow=Montecarlo.size,ncol=2)
```

```
#-- Head 1), 2) and 3)
```

```
Errors<-rt(Sample_size,3)
```

```
Response_values<-Beta.0+Beta.1*Sample+Errors
```

```
Hat_beta.1<-(sum((Sample-mean(Sample))*Response_values))/((Sample_size-1)*var(Sample))
```

```
Hat_beta.0<-mean(Response_values)-Hat_beta.1*mean(Sample)
```

```
Residuals<-Response_values-Hat_beta.0-Hat_beta.1*Sample
```

```
Residuals_modified<-Residuals/sqrt(1-(1/Sample_size+((Sample-mean(Sample))^2)/((Sample_size-1)*var(Sample))))
```

```
Matrix_bootstrap_errors<-matrix(sample(Residuals_modified-mean(Residuals_modified),size=Sample_size*Bootstrap_resample_size,replace=T),nrow=Sample_size,ncol=Bootstrap_resample_size)
```

```
Response_values_bootstrap<-Hat_beta.0+Hat_beta.1*Sample+Matrix_bootstrap_errors
```

```
Hat_beta.1_bootstrap<-apply((Sample-mean(Sample))*Response_values_bootstrap,2,sum)/((Sample_size-1)*var(Sample))
```

```
Hat_beta.0_bootstrap<-apply(Response_values_bootstrap,2,mean)-Hat_beta.1_bootstrap*mean(Sample)
```

```
Residuals_bootstrap<-t(Response_values_bootstrap)-Hat_beta.0_bootstrap-outer(Hat_beta.1_bootstrap,Sample)
```

```
T_statistic_bootstrap<-(Hat_beta.1_bootstrap-Hat_beta.1)/sqrt((apply(t(Residuals_bootstrap)^2,2,sum)/(Sample_size-2))/((Sample_size-1)*var(Sample)))
```

```
IC_bootstrap<-c(Hat_beta.1-sort(T_statistic_bootstrap)[floor((1-Alpha/2)*Bootstrap_resample_size)]*sqrt((sum(Residuals^2)/(Sample_size-2))/((Sample_size-1)*var(Sample))),
```

```
    Hat_beta.1-sort(T_statistic_bootstrap)[floor(Alpha/2*Bootstrap_resample_size)]*sqrt((sum(Residuals^2)/(Sample_size-2))/((Sample_size-1)*var(Sample))))
```

```
((IC_bootstrap[1]<Beta.1)&(Beta.1<IC_bootstrap[2]))
```

## Apéndice II. Comparación método paramétrico y no paramétrico

```
#-- Head 4) and 5)

for(m in 1:Montecarlo.size){

  Errors<-rt(Sample.size,3)

  Response_values<-Beta.0+Beta.1*Sample+Errors

  Hat_beta_1<-(sum((Sample-mean(Sample))*Response_values))/((Sample.size-1)*var(Sample))
  Hat_beta_0<-mean(Response_values)-Hat_beta_1*mean(Sample)
  Residuals<-Response_values-Hat_beta_0-Hat_beta_1*Sample

  Residuals_modified<-Residuals/sqrt(1-(1/Sample.size+((Sample-mean(Sample))^2)/((Sample.size-1)*var(Sample))))

  Matrix_bootstrap_errors<-matrix(sample(Residuals_modified-mean(Residuals_modified),size=Sample.size*Bootstrap_resample_size,replace=T),nrow=Sample.size,ncol=Bootstrap_resample_size)

  Response_values_bootstrap<-Hat_beta_0+Hat_beta_1*Sample+Matrix_bootstrap_errors
  Hat_beta_1_bootstrap<-apply((Sample-mean(Sample))*Response_values_bootstrap,2,sum)/((Sample.size-1)*var(Sample))
  Hat_beta_0_bootstrap<-apply(Response_values_bootstrap,2,mean)-Hat_beta_1_bootstrap*mean(Sample)
  Residuals_bootstrap<-t(Response_values_bootstrap)-Hat_beta_0_bootstrap-outer(Hat_beta_1_bootstrap,Sample)

  T_statistic_bootstrap<-(Hat_beta_1_bootstrap-Hat_beta_1)/sqrt((apply(t(Residuals_bootstrap)^2,2,sum))/((Sample.size-2) * (Sample.size-1)*var(Sample)))
```

```

IC.Montecarlo.bootstrap.student_errors[m,]<-c(Hat.beta.1-(sort(T.statistic.bootstrap)[floor((1-Alpha/2)*Bootstrap.resample.size)])*sqrt((sum(Residuals^2)/(Sample.size-2))/((Sample.size-1)*var(Sample))),
      Hat.beta.1-sort(T.statistic.bootstrap)[floor(Alpha/2*Bootstrap.resample.size)]*sqrt((sum(Residuals^2)/(Sample.size-2))/((Sample.size-1)*var(Sample))))

IC.Montecarlo.student.student_errors[m,]<-c(Hat.beta.1-qt(1-Alpha/2,df=Sample.size-2)*sqrt((sum(Residuals^2)/(Sample.size-2))/((Sample.size-1)*var(Sample))),
      Hat.beta.1+qt(1-Alpha/2,df=Sample.size-2)*sqrt((sum(Residuals^2)/(Sample.size-2))/((Sample.size-1)*var(Sample))))

}

(Coverage_bootstrap_student_errors<-sum((IC.Montecarlo.bootstrap.student_errors[,1]<=1&(IC.Montecarlo.bootstrap.student_errors[,2]>=1))/Montecarlo.size)

(Coverage_student_student_errors<-sum((IC.Montecarlo.student.student_errors[,1]<=1)&(IC.Montecarlo.student.student_errors[,2]>=1))/Montecarlo.size)

# Intervals plots

ind<-1:Montecarlo.size

windows()

plot(IC.Montecarlo.bootstrap.student_errors[,2]~ind,type='n',ylim=c(min(IC.Montecarlo.bootstrap.student_errors[,1],max(IC.Montecarlo.bootstrap.student_errors[,2])),xlab='Montecarlo
Size',ylab='Bootstrap Intervals')

for(i in 1:Montecarlo.size){
segments(ind[i],IC.Montecarlo.bootstrap.student_errors[i,1],ind[i],IC.Montecarlo.bootstrap.student_errors[i,2])
}

```

```
abline(h=1,col=2,lwd=2)
```

```
windows()
```

```
plot(IC_Montecarlo_student_student_errors[,2]~ind,type='n',ylim=c(min(IC_Montecarlo_student_student_errors[,1],max(IC_Montecarlo_student_student_errors[,2])),xlab='Montecarlo  
Size',ylab='Student Intervals')
```

```
for(i in 1:Montecarlo_size){  
  segments(ind[i],IC_Montecarlo_student_student_errors[i,1],ind[i],IC_Montecarlo_student_student_errors[i,2])  
}
```

```
abline(h=1,col=2,lwd=2)
```

### Apéndice III. Comparación método paramétrico y no paramétrico con errores normales

```
#-- Head 6)

for(m in 1:Montecarlo.size){

  Errors<-rnorm(Sample_size,mean=0,sd=sqrt(3))

  Response_values<-Beta.0+Beta.1*Sample+Errors

  Hat.beta.1<-(sum((Sample-mean(Sample))*Response_values))/((Sample_size-1)*var(Sample))
  Hat.beta.0<-mean(Response_values)-Hat.beta.1*mean(Sample)
  Residuals<-Response_values-Hat.beta.0-Hat.beta.1*Sample

  Residuals_modified<-Residuals/sqrt(1-(1/Sample_size+((Sample-mean(Sample))^2)/((Sample_size-1)*var(Sample))))
  Matrix_bootstrap_errors<-matrix(sample(Residuals_modified-mean(Residuals_modified),size=Sample_size*Bootstrap_resample_size,replace=T),nrow=Sample_size,ncol=Bootstrap_resample_size)

  Response_values_bootstrap<-Hat.beta.0+Hat.beta.1*Sample+Matrix_bootstrap_errors
  Hat.beta.1.bootstrap<-apply((Sample-mean(Sample))*Response_values_bootstrap,2,sum)/((Sample_size-1)*var(Sample))
  Hat.beta.0.bootstrap<-apply(Response_values_bootstrap,2,mean)-Hat.beta.1.bootstrap*mean(Sample)
  Residuals_bootstrap<-t(Response_values_bootstrap)-Hat.beta.0.bootstrap-outer(Hat.beta.1.bootstrap,Sample)

  T_statistic_bootstrap<-(Hat.beta.1.bootstrap-Hat.beta.1)/sqrt(( apply(t(Residuals_bootstrap)^2,2,sum))/((Sample_size-2) * (Sample_size-1)*var(Sample)))

  IC.Montecarlo.bootstrap.normal.errors[m,]<-c(Hat.beta.1-(sort(T_statistic_bootstrap)[floor((1-Alpha/2)*Bootstrap_resample_size)])*sqrt((sum(Residuals^2)/(Sample_size-2))/((Sample_size-1)*var(Sample))),
        Hat.beta.1-sort(T_statistic_bootstrap)[floor(Alpha/2*Bootstrap_resample_size)]*sqrt((sum(Residuals^2)/(Sample_size-2))/((Sample_size-1)*var(Sample))))
```

```

IC.Montecarlo.student_normal_errors[m,]<-c(Hat.beta_1-qt(1-Alpha/2,df=Sample.size-2)*sqrt((sum(Residuals^2)/(Sample.size-2))/((Sample.size-1)*var(Sample))),
      Hat.beta_1+qt(1-Alpha/2,df=Sample.size-2)*sqrt((sum(Residuals^2)/(Sample.size-2))/((Sample.size-1)*var(Sample))))
}

(Coverage_bootstrap_normal_errors<-sum((IC.Montecarlo.bootstrap_normal_errors[,1]<=1)&(IC.Montecarlo.bootstrap_normal_errors[,2]>=1))/Montecarlo.size)
(Coverage_student_normal_errors<-sum((IC.Montecarlo.student_normal_errors[,1]<=1)&(IC.Montecarlo.student_normal_errors[,2]>=1))/Montecarlo.size)

# Intervals plots

windows()

ind<-1:Montecarlo.size

plot(IC.Montecarlo.bootstrap_normal_errors[,2]~ind,type='n',ylim=c(min(IC.Montecarlo.bootstrap_normal_errors[,1]),max(IC.Montecarlo.bootstrap_normal_errors[,2])),xlab='Montecarlo
Size',ylab='Bootstrap Intervals')

for(i in 1:Montecarlo.size){
  segments(ind[i],IC.Montecarlo.bootstrap_normal_errors[i,1],ind[i],IC.Montecarlo.bootstrap_normal_errors[i,2])
}

abline(h=1,col=2,lwd=2)

```

```
windows()
```

```
plot(IC.Montecarlo_student_normal_errors[,2]~ind,type='n',ylim=c(min(IC.Montecarlo_student_normal_errors[,1]),max(IC.Montecarlo_student_normal_errors[,2])),xlab='Montecarlo  
Size',ylab='Student Intervals')
```

```
for(i in 1:Montecarlo.size){  
  segments(ind[i],IC.Montecarlo_student_normal_errors[i,1],ind[i],IC.Montecarlo_student_normal_errors[i,2])  
}
```

```
abline(h=1,col=2,lwd=2)
```