# DOCUMENTACIÓN

# BOOTSTRAPPING

# Marzo 2021

Adrián sandoval cordero

Inicio PROYECTO: 01-AGOSTO-2020

Introducción**1**

Obtención de Códigos2

Definición de Coberturas3

**Procesamiento de Base de Datos4**

Obtención de Flatfile por Ramo5

Analisis de Información6

# Introducción

El siguiente documento contienen la descripción técnica sobre la macro así como los supuestos e interpretación.

# Estructura General del Documento

El documento se compone de la siguiente forma:

1. Descripción de la metodología
2. Supuestos y Codificación en SAS
3. Interpretación de los *outputs* de la macro

Al mes de la creación de la primera versión de este documento (Marzo 2021), la macro se encuentra en su tercera versión (adjunta en el anexo bajo el nombre de “Bootstrapping\_v3”).

1. DESCRIPCIÓN DE LA METODOLOGÍA

La metodología de *Bootstrapping* es una técnica de remuestreo usado para inferencia estadística, se basa en la creación de una nueva muestra (de nuestra muestra observada original) para poder hacer una estimación sobre las estadísticas poblacionales (entiéndase estadística como función de algún parámetro poblacional es decir ).

La idea general de *Bootstrapping* se puede resumir de la siguiente forma: La población es a la muestra así como ésta muestra es a la submuestra.

Supongamos que sacamos una muestra de una población y estamos interesados en estimar un parámetro poblacional, digamos (este parámetro puede ser un vector).

Para poder hacer inferencia sobre el parámetro poblacional mencionado arriba, se necesita crear una submuestra con reemplazo basado en y que llamaremos *Bootstrap sample* denotada , algunos métodos alternativos sugieren que ésta nueva muestra sea creada de la misma forma en la que se creó la muestra original; sin embargo, la macro implementada en SAS crea la submuestra usando el método de Muestreo Simple con Reemplazo, que consiste en que cada elemento de la muestra tiene la misma probabilidad de ser seleccionado (la probabilidad exacta es ) y cada elemento se puede elegir más de una vez.

La creación de esta nueva submuestra se replica b-veces, la b-ésima *Bootstrap sample* se denotará.

Se estimará el parámetro poblacional con cada una de los *Bootstrap sample*.

La distribución de para estimar es análoga a la distribución para estimar .

Con el parámetro se puede generar una estimación para la muestra poblacional.

Aplicado en términos de los Modelos Linelaes Generalizados (GLM por sus siglas en inglés *Generalized Linear Models*), le técnica aplicada sigue el siguiente algoritmo.

**Algoritmo de SAS**

1. Tenemos una muestra y mediante la técnica de Muestreo Simple con Reemplazo (MScR) se encuentra el *Bootstrap sample*
2. Se repite el **Paso 1** b-veces
3. Con cada muestra se crea una estimación de los coeficientes del modelo
4. Se calculan las siguientes métricas
5. Con las métricas se calcula los siguientes intervalos de confianza (IC)
6. Metodología 1, hace uso de los promedios calculados en el paso anterior para encontrar el IC para el j-ésimo coeficiente del modelo

Donde es el cuantil de la distribución normal estándar (por ejemplo si entonces )

1. Metodología 2, se hace uso de los cuantiles obtenidos mediante el remuestreo

Donde es el cuartil inferior y el cuartil superior donde los cuartiles.

Como se calculan b submuestras; entonces, existen b-coeficientes () que se ordenan de forma creciente:

Finalmente, se calculan los coeficientes de la siguiente manera

2. SUPUESTOS Y CODIFICACIÓN EN SAS

Para hacer el remuestreo se usó la técnica de: Remuestro Simple con Reemplazo

La función de SAS se llama de la siguiente forma:

%***BOOTSTRAP\_v3***(Model\_Data,Var\_Dep,Var\_Indep,N,Class\_Factor,alpha,Offset\_var,Weight,DIST,LINK)

Donde cada input significa lo siguiente:

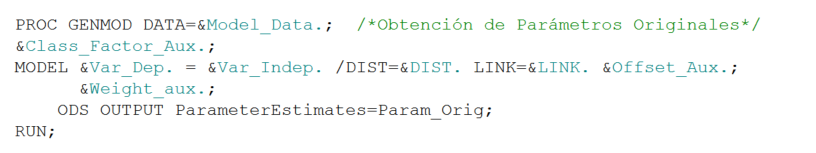
* Model\_Data, la tabla de datos donde se encuentra las variables a usar
* Var\_Dep, el nombre de la variable dependiente
* Var\_Indep, vector con los nombres de variables independientes
* N, número de iteraciones en el remuestreo, se recomienda que éste parámetro sea 100 ya que así se obtienen buenos resultados y el programa no toma mucho tiempo
* Class\_Factor, nombre de variables categóricas
* alpha, nivel de significancia de los intervalos de confianza
* Offset\_var, nombre de variable “*Offset*”
* Weight, nombre de variable de pesos
* DIST, la distribución probabilística del modelo

LINK, función encaje del modelo

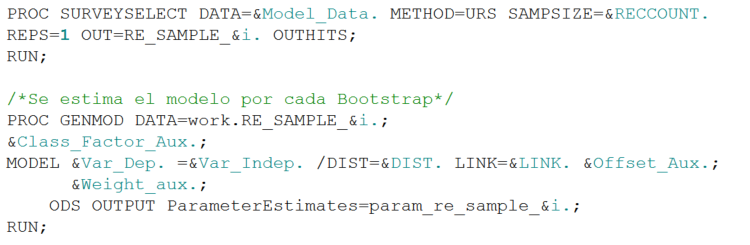
Para la codificación en SAS se aplicaron los siguientes pasos.

1. Se obtienen los percentiles para la metodología b) mencionada anteriormente así como los parámetro del modelo GLM como referencia





1. Se hace el cálculo de los parámetros con los remuestreos

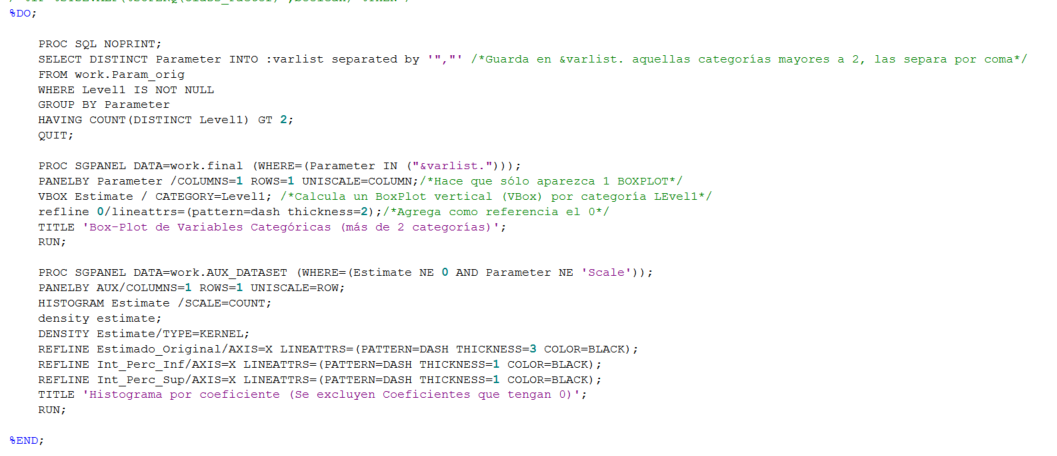


1. Se calculan las medias y desviación estándar de los coeficientes para el metodología a) mencionada anteriormente y se estima los primeros coeficientes



1. Finalmente, se calculan los *Boxplot* para los coeficientes que tienen variables categóricas (aquellas variables categóricas que tengan más de dos clases) así como los histogramas por cada coeficiente para poder hacer más análisis.

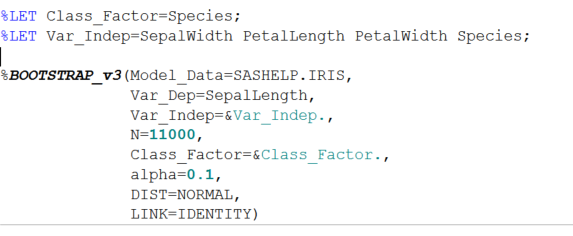
Los *Boxplot* se hacen con más de dos categorías para poder identificar mejor el efecto de la variable, ya que si tomamos una variable con dos categorías observamos un coeficiente siempre tendrá cero y por lo tanto no se podrá estudiar el efecto de ésta variable en el modelo.



3. INTEREPRETACIÓN DE LOS *OUTPUTS* DE LA MACRO

Ejemplo de uso e interpretación

* Modelo Lineal Múltiple



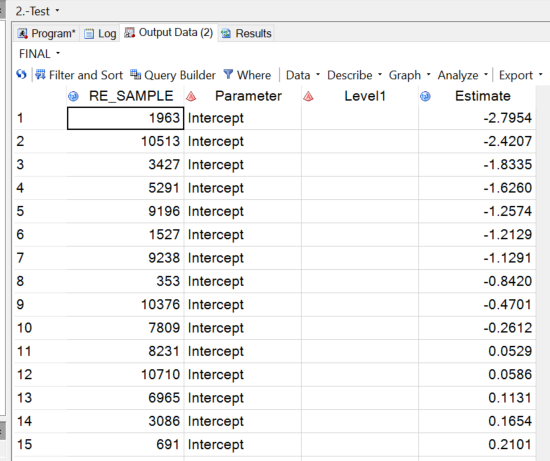
Cuando se calcula la hay 2 outputs:

1. FINAL\_CI, esta base contiene los IC con las 2 metodologías por cada parámetro, así como los parámetros estimados originalmente con las siguientes métricas:
   * Estimado\_Original, parámetro original
   * StdErr, error estándar original originales
   * LowerWaldCL y UpperWaldCL, IC con el método de Wald de la estimación original
   * ChiSq y ProbChiSq, los parámetros que indican significancia de las variables (como convenio deben ser menores a 5% para que la variable sea significativa)
   * Mean\_Re y Std\_Re, tienen la media y desviación estándar de los coeficientes que se estimaron mediante las remuestras
   * Bias, es el sesgo del coeficiente que se mide de la siguiente forma
   * Int\_Norm\_Inf y Int\_Norm\_Sup, tiene los intervalos de confianza con la metodología a) descrita arriba
   * Int\_Perc\_Inf y Int\_Perc\_Sup, tiene los intervalos de confianza con la metodología b) descrita arriba

Se puede interpretar que como ningún IC contiene el 0; por lo tanto, todas las variables son significativas.



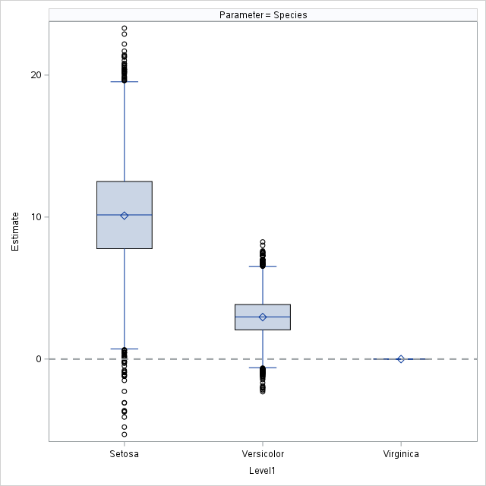
1. FINAL, esta base tiene los coeficientes de todos los remuestreos, además la variable RE\_SAMPLE tiene el número el número de remuestreo al que pertenece el coeficiente



1. Finalmente contiene un BoxPlot e Histograma
   * *Boxplot*

Este se hace para determinar si las categorías en el modelo tienen adecuada separación, en este ejemplo se usó la variable “Species” que tiene 3 categorías

Setosa, Versicolor y Virginica.



De el gráfico de *Boxplot* de arriba se puede observar que aunque existen muchos coeficientes outlier el rango intercuartil permanece diferente y arriba del cero; por lo tanto, las categorías hacen una buena separación

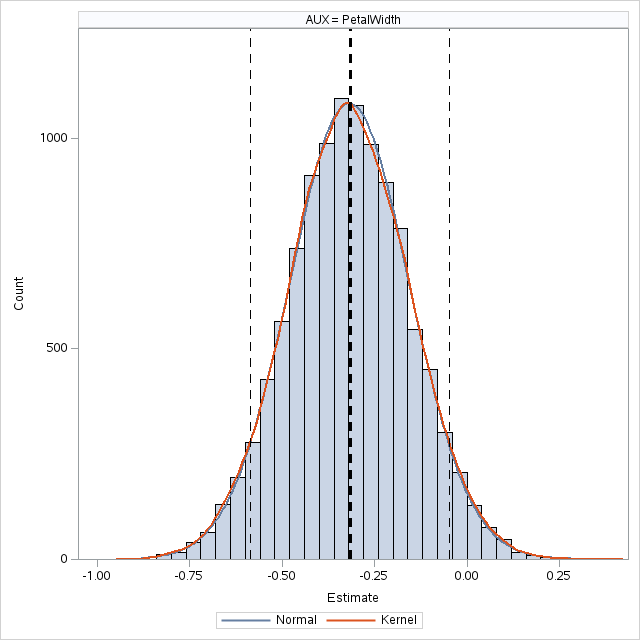
* + Histograma

El histograma se muestra para cada una de las variables, con el fin de identificar algún sesgo en el coeficiente. Se mostrarán algunos ejemplos de histograma.

El histograma de abajo es para la variable “PetalWidth”, la línea roja identificada como “Kernel” es la distribución de la muestra y la línea azul identificada como “Normal”.

Las líneas punteadas de los extremos son los percentiles del IC y la línea de en medio es la mediana.

Se espera que los coeficientes tengan una distribución normal para considerar que el coeficiente no tiene sesgo, esto implica que tanto el “Kernel” tenga un comportamiento similar a “Normal”, la mediana sea igual a la media y esté equidistante a los percentiles.



A continuación, viene el histograma de la variable “PetalLength”, éste muestra un leve sesgo; sin embargo, no es significativo ya que la mediana está cerca de la media y se muestra equidistance.

