# DOCUMENTACIÓN

# FUNCIÓN DE VALIDACIÓN: CROSS VALIDATION

# Marzo 2021

Autor: Adrián sandoval cordero

Versión 1.0

Inicio PROYECTO: 01-AGOSTO-2020

# Introducción

Durante el año 2020, Pedro de Zuviría (Zurich México *Chief Data Scientist*) solicitó la elaboración de herramientas para mejorar la validación de los modelos del área que serán implementados en el software *Statistical Analysis Software* (SAS); por lo tanto, se decidió crear una macro que implemente una validación mediante la técnica de *Cross Validation (o k-folded validation)* para validar los coeficientes de los modelos.

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 “Cross\_Validation\_v1”).

1. DESCRIPCIÓN DE LA METODOLOGÍA

La metodología de *Cross Validation* sirve para evaluar el desempeño de un modelo y también funciona identificar si el modelo tuvo mucho *overfit* ya que la variabilidad sería alta.

El algoritmo se para evaluarlo es el siguiente:

1. La muestra se divide de forma aleatoria en k-partes iguales de manera aleatoria (en inglés son conocidos como *folds*, por eso la metodología tiene el nombre alternativo de *k-fold validation*)
2. Se toma una Submuestra j (con j= 1, …, k) y se estima el modelo con el resto de las submuestras (es decir con toda la muestra excepto con aquellos datos que correspondan a la Submuestra j).
3. Con el modelo estimado en el paso anterior, se evalúa el modelo en la submuestra excluida (es decir con la Submuestra j).
4. El Paso 2-3 se repite hasta que se hayan excluidas todas las k submuestras (es decir hasta que se haya estimado el modelo y calculado el error excluyendo con cada una de las submuestras)

2. SUPUESTOS Y CODIFICACIÓN EN SAS

**Supuestos**

Para la codificación es SAS se usó la siguiente Métrica de Error:

Para los modelos se cumple lo siguiente:

Esto es válido cuando los valores estimados son de la muestra usada para estimar; sin embargo, como esta métrica de error se calculará con los datos de la muestra excluida.

Así mismo se estimó la desviación estándar para poder calcular la variabilidad entre la estimación de cada submuestra y evaluar si el modelo (o los datos) varían mucho entre cada estimación.

**Codificación en SAS**

La función de SAS se llama de la siguiente forma:

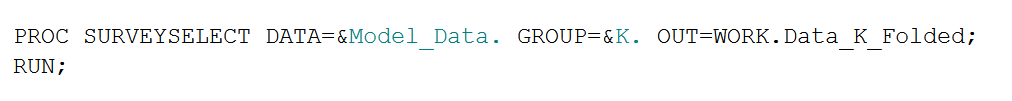
%***Cross\_Validation\_v1***(Model\_Data,Var\_Dep,Var\_Indep,K,Class\_Factor,Offset\_var,Weight,DIST,LINK)

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
* K, número de “*folds*” en la validación
* Class\_Factor, nombre de variables categóricas
* Offset\_var, nombre de variable “*Offset*”
* Weight, nombre de variable de pesos
* DIST, la distribución probabilística del modelo
* LINK, función encaje del modelo

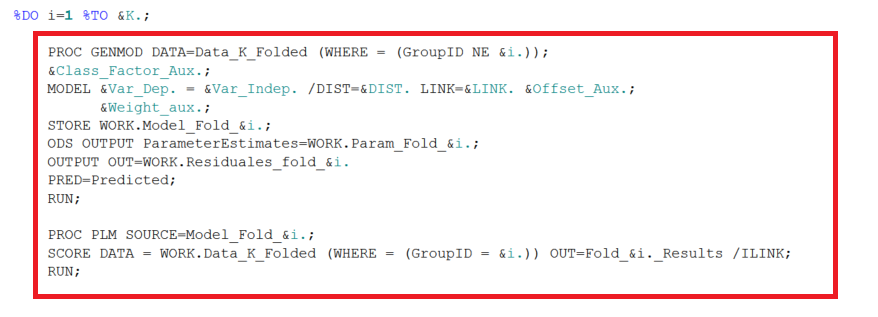
Se adaptó en SAS el algoritmo descrito en el capítulo anterior de la siguiente manera:

1. Para fraccionar la muestra en k-partes iguales, se hizo con el siguiente *PROCEDURE*



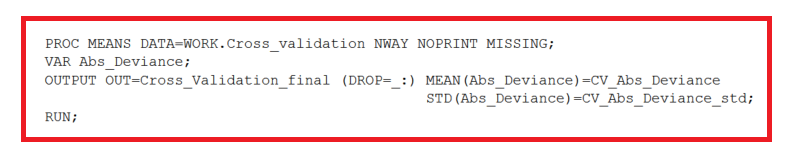
El input *GROUP=&K.* crea una columna con un número donde indica a que Submuestra pertenece la observación

1. Para hacer el Paso 2 del algoritmo, se hace lo siguiente



Se estima el modelo en *PROC GENMOD* y se evalúa en la submuestra excluida con el *PROCEDURE PLM*.

1. Finalmente se calculan las métricas de la siguiente manea



3. INTEREPRETACIÓN DE LOS *OUTPUTS* DE LA MACRO

Ejemplo de uso e interpretación,

* **Modelo Lineal Múltple:**

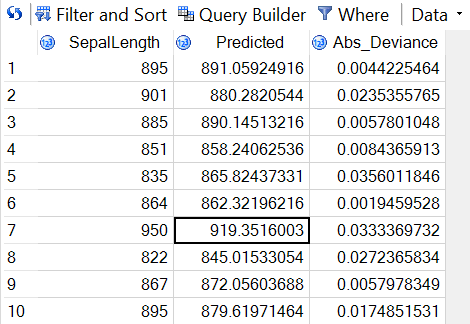
%LET Class\_Factor=Species;

%LET Var\_Indep=SepalWidth PetalLength PetalWidth;

%***Cross\_Validation\_v1***(Model\_Data=SASHELP.IRIS,Var\_Dep=SepalLength,Var\_Indep=&Var\_Indep.,K=**10**,Class\_Factor=&Class\_Factor.,DIST=NORMAL,LINK=IDENTITY)

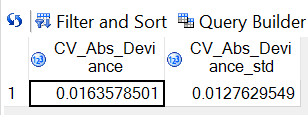
Cuando se calcula la hay 3 outputs:

1. CROSS\_VALIDATION, tiene las métricas de cada una de las 10 submuestras

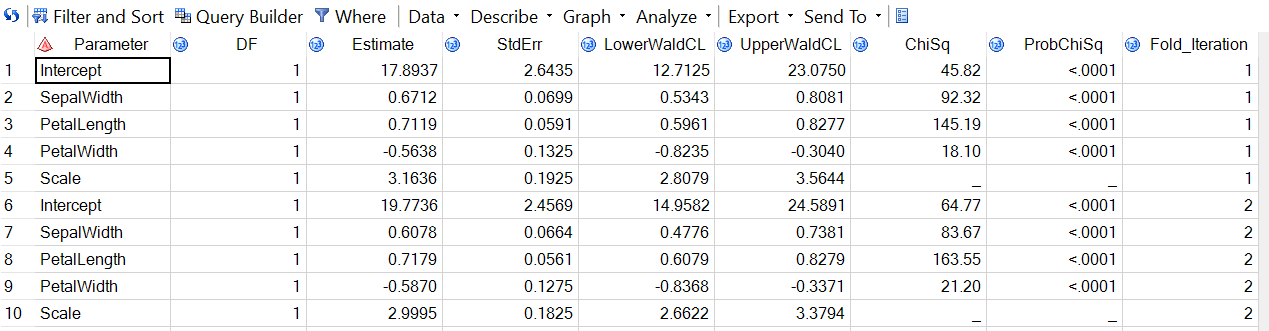


1. CROSS\_VALIDATION\_FINAL, tiene las métricas finales. En este caso el CV\_Abs\_Deviance tiene 1.635% de Error Absoluto y una variabilidad 1.28%, lo cual es significativamente bajo por lo tanto el modelo tiene un ajuste adecuado.

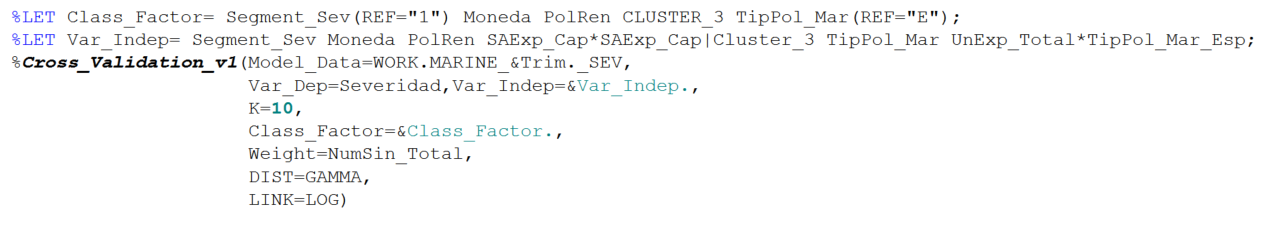
Finalmente, la variable CV\_Abs\_Deviance también funciona identificar si el modelo tuvo mucho *overfit* ya que la variabilidad sería alta.



1. FINAL\_PARAM, es una base con los parámetros hechos por cada iteración



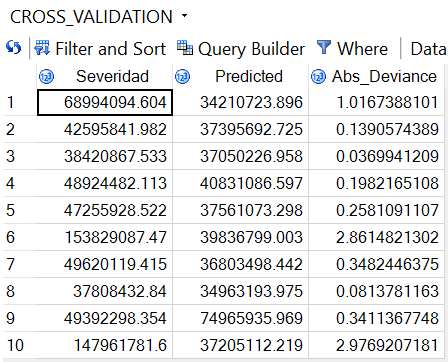
* **Modelo Gamma Múltple:**



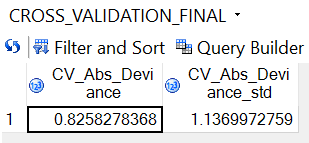
Este modelo tiene variables categóricas (Macro variable Class\_Factor tiene el nombre de esas variable) y así mismo tiene la variable de peso (Macro variable Weight=NumSin\_Total).

Cuando se calcula la hay 3 outputs:

1. CROSS\_VALIDATION, tiene las métricas de cada una de las 10 submuestras



1. CROSS\_VALIDATION\_FINAL, tiene las métricas finales. En este caso el CV\_Abs\_Deviance tiene 82.58% de Error Absoluto y una variabilidad 113.7%, lo cual es alto.



1. FINAL\_PARAM, es una base con los parámetros hechos por cada iteración

