

# TRABAJO FINAL - MODELO GLM

Casos de Estudios Aplicados al Sector Seguros  
y Financieros en Colombia

Por: Sergio Alejandro Burbano

Docentes: Julian Caro

Ximena Quiroga

## CÁLCULO DE LA COLUMNA 'RESP'

- Se crea una nueva columna llamada '**resp**' dividiendo los valores de `stro_Corr_AGUAACAGBC` entre los valores de `exp_corr_ACAGBC`.
- **Variables:**
  - **var\_resp:** `stro_Corr_AGUAACAGBC`
  - **peso:** `exp_corr_ACAGBC`
  - **ano:** `year`

## IDENTIFICACIÓN DE VALORES ÚNICOS EN LA COLUMNA 'CATEGORÍA'

- Se obtiene una lista de valores únicos de la columna 'CATEGORÍA'.

## IDENTIFICACIÓN DE COLUMNAS CATEGÓRICAS

- Se seleccionan todas las columnas con tipo de dato **object** o **category**.

# Manejo de Valores Faltantes y Creación de Variables Dummy

## ► Manejo de valores faltantes

- **Columnas numéricas:** Se reemplazan los valores NaN con la media de la columna.
- **Columnas categóricas:** Se reemplazan los valores NaN con la moda (valor más frecuente) de la columna 'catage\_glob'.

## ► Creación de variables dummy

- Se generan variables dummy para las columnas categóricas usando `pd.get_dummies()`, eliminando la primera categoría para evitar multicolinealidad (`drop_first=True`).

# Ajuste del Modelo GLM y Resultados

## ▶ Preparación de variables para el modelo

- ▶ **X:** Se seleccionan las columnas necesarias para el modelo, convirtiéndolas a tipo **float** y rellenando valores NaN con 0.
- ▶ **y:** Se selecciona la columna 'resp' como variable dependiente.

## ▶ Definición de exposición (offset)


- ▶ Se define el offset como el logaritmo de exp\_corr\_ACAGBC, reemplazando valores infinitos o NaN por 0.


## ▶ Ajuste del modelo GLM

- ▶ Se crea un modelo **GLM** con distribución de Poisson y enlace logarítmico.
- ▶ El modelo se ajusta con el método `fit()`, y se imprime el resultado con `resultado.summary()`.

## ▶ Resultados del modelo

- ▶ **Pseudo R-cuadrado:** 1.000 (posible sobreajuste).
- ▶ Las variables significativas incluyen factores geográficos y socioeconómicos.
- ▶ **Método IRLS:** Estimación de coeficientes.

- 
- modelo de **Regresión Generalizada (GLM)** con distribución **Poisson** y **enlace logarítmico** para predecir la variable "resp", empleando 97 variables explicativas y 371,390 observaciones. El modelo revela que diversas variables, especialmente aquellas relacionadas con características geográficas, demográficas y socioeconómicas (como "PERIFERIA METROPOLITANA DE NIVEL MEDIO" y la categoría de hogares), tienen un impacto significativo en la variable dependiente.

- 
- ▶ El **Pseudo R-cuadrado** es 1.000, lo que indica un ajuste perfecto, aunque esto puede sugerir **sobreajuste** al capturar patrones específicos o ruido. La **deviance** y el **log-likelihood** muestran una buena adecuación del modelo, pero aún pueden indicar que el ajuste puede mejorarse.
  - ▶ El modelo emplea el algoritmo **IRLS (Iterative Reweighted Least Squares)**, adecuado para distribuciones Poisson, comúnmente usadas para contar eventos raros. En resumen, el modelo GLM demuestra una relación fuerte entre las variables y la variable "resp", pero el ajuste perfecto puede requerir un análisis adicional de la estructura de los datos.