Facultad de Ingenierías
Inferencia Estadística
Maestría en Inteligencia Artificial y Ciencia de Datos
**Prof**. Cristian E Garcia.

# Desafío 3

**Condiciones:**

- Subir la tarea en formato pdf en la plataforma UAO-Virtual.
- Es necesario incluir el código como anexo. Mostrar los resultados a partir de tablas, gráficos o indicadores que les permita dar respuesta a los planteamientos.
- Se debe realizar un informe estadístico de los resultados, en el informe se debe omitir los apartados de código, si es necesario utilce repositorios o anexos, si su desarrollo es computacional genere una guía técnica de los resultados.
- Deben interpretar los resultados obtenidos en cada situación de acuerdo al contexto.
- Realizar la actividad en grupos máximo de 4 personas.

## Situación 1

Sea

$$Y_1 = \theta + \epsilon_1$$
$$Y_2 = 2\theta - \varphi + \epsilon_2$$
$$Y_3 = \theta + 2\varphi + \epsilon_3$$

donde $E[\epsilon_i] = 0$ para $i = 1,2,3$. Encuentre las estimaciones de los parámetros $\theta$ y $\varphi$ por mínimos cuadrados.

## Situación 2

En un estudio se desean investigar los factores que influencian la elección de alimento primario de los caimanes. El estudio se basa en la información obtenida de 219 caimanes capturados en cuatro lagos de Florida. La variable respuesta de escala nominal es el tipo primario de alimento, en volumen, encontrado al interior del estómago de un caimán. En la tabla a continuación se presenta la clasificación de la elección de alimento primario según lago y tamaño del caimán.

| Lago | Tamaño (m) | Peces | Invertebrados | Reptiles | Aves | Otros |
|---|---|---|---|---|---|---|
| Hancock | ≤ 2.3 | 23 | 4 | 2 | 2 | 8 |
| Hancock | > 2.3 | 7 | 0 | 1 | 3 | 5 |
| Ocklawaha | ≤ 2.3 | 5 | 11 | 1 | 0 | 3 |
| Ocklawaha | > 2.3 | 13 | 8 | 6 | 1 | 0 |
| Trafford | ≤ 2.3 | 5 | 11 | 2 | 1 | 5 |
| Trafford | > 2.3 | 8 | 7 | 6 | 3 | 5 |
| George | ≤ 2.3 | 16 | 19 | 1 | 2 | 3 |
| George | > 2.3 | 17 | 1 | 0 | 1 | 3 |

(a) Fije la categoría de referencia en: Peces. Ajuste el modelo.
(b) Calcule las razones de chances para los lagos y el tamaño.
(c) Presente de manera ordenada dichas razones en una tabla.
(d) ¿Tiene algún efecto el tamaño del caimán sobre la elección del alimento primario? Justifique e interprete los posibles efectos en términos de razones de chances.Interprete el efecto del tamaño del caimán sobre la elección primaria de alimentos invertebrados.
(e) ¿Tiene algún efecto el lago sobre la elección del alimento primario? Justifique e interprete los posibles efectos en términos de razones de chances

# Situación 3

Usted ha sido contratado como científico de datos por la Comisión de Taxis y Limusinas (TLC) de Nueva York. Su tarea es desarrollar modelos para entender los factores que influyen en tres aspectos de los viajes:

1. La probabilidad de recibir una propina.
2. La cantidad de pasajeros por viaje.
3. El monto de la propina recibida.

Se debe utilizar el conjunto de datos de **Enero de 2023** de los Taxis Amarillos. Puede descargarlo directamente desde el sitio de la TLC de NYC.

**Modelo 1: Modelado de la Ocurrencia de Propinas**

- **Objetivo**: Identificar los factores que aumentan la probabilidad de que un cliente deje propina.
- **Modelo Base**: Ajuste un modelo para predecir la variable `recibio_propina` (1 si `tip_amount` > 0, 0 en caso contrario).

- **Alternativa a Investigar**: ¿Qué sucede con los tiempos de cómputo y la estabilidad de los coeficientes al usar un algoritmo más rápido como `speedglm` (R) o al ajustar el modelo sobre submuestras de diferentes tamaños (e.g., 1%, 10%, 50% de los datos)?

**Modelo 2: Modelado del Número de Pasajeros**

- **Objetivo**: Entender cómo varían el número de pasajeros (`passenger_count`) según la hora del día, la distancia del viaje o la zona.
- **Modelo Base**: Ajuste un modelo adecuado. Evalúe el supuesto de equidispersión (media ≈ varianza).
- **Alternativa a Investigar**: Si detecta sobredispersión, ajuste un modelo **Binomial Negativo** y compare los resultados (coeficientes, errores estándar y bondad de ajuste). Justifique por qué este modelo es teóricamente más apropiado.

**Modelo 3: Modelado del Monto de la Propina**

- **Objetivo**: Predecir el monto de la propina (`tip_amount`) condicionado a que esta sea mayor a cero.
- **Modelo Base**: Filtre los datos para `tip_amount` > 0 y ajuste un **GLM con distribución Gamma** y función de enlace logarítmica.
- **Alternativa a Investigar**: Explore el impacto de usar una función de enlace diferente (e.g., `inverse`). Adicionalmente, considere un enfoque de **regularización (LASSO o Ridge)** usando `glmnet` para manejar la posible multicolinealidad y realizar selección de variables. Discuta las ventajas de este enfoque en un contexto de "big data".

# Situación 4

En la sesión 7 del curso se introduce la estimación por máxima verosimilitud (MV) en modelos lineales generalizados (GLM), haciendo énfasis en métodos de optimización numérica como el algoritmo de *Fisher Scoring*. Este tipo de métodos, conocidos como algoritmos **batch**, requieren el uso completo del conjunto de datos en cada iteración para calcular el gradiente y la matriz de información. En contextos con bases de datos de gran tamaño, esta condición se convierte en una limitación computacional cuando los datos exceden la capacidad de almacenamiento en memoria del sistema.

El objetivo es modeloar el riesgo de crédito utilizando un conjunto de datos hipotecarios cuyo tamaño impide su carga completa en memoria. Dado este escenario, se propone la implementación de un estimador de máxima verosimilitud mediante un algoritmo de optimización **online**, es decir, uno que procese los datos de forma secuencial o en mini-lotes (*mini-batches*). El objetivo es Diseñar, implementar y evaluar un estimador de máxima verosimilitud para un modelo de regresión logística que prediga la probabilidad de incumplimiento (default) de un préstamo hipotecario. La estimación debe realizarse a partir de un conjunto de datos masivo, procesado sin cargarlo completamente en memoria principal. Para ello deberá desarrollar las siguientes actividades

**1. Modelo Teórico y Derivación del Gradiente Estocástico**

- Partiendo de la función de log-verosimilitud del modelo logístico, derive la expresión del gradiente correspondiente a una única observación.
- Este gradiente individual será la base para el algoritmo de descenso de gradiente estocástico (SGD), el cual se utiliza en la estimación online.

**2. Implementación del Optimizador Online**

- Implemente una función en el elnguaje de su preferencia que resuelva el problema de optimización utilizando SGD o *mini-batch* SGD.
- La función debe:
  - Leer el archivo de datos en fragmentos (chunks).
  - Actualizar el vector de coeficientes $\beta$ luego de procesar cada fragmento.
  - Liberar el fragmento procesado antes de cargar el siguiente.
- Utilice la siguiente regla de actualización para cada observación $i$ en un mini-lote:

  $\beta^{(t+1)} = \beta^{(t)} - \alpha \nabla \ell_i(\beta^{(t)})$

  donde $\alpha$ representa la tasa de aprendizaje y $\nabla \ell_i$ es el gradiente con respecto a la observación $i$

**3. Análisis de Convergencia y Evaluación del Modelo**

- Diseñe un criterio de convergencia adecuado para este enfoque. Dado que no es viable computar la log-verosimilitud total en cada iteración, proponga métricas alternativas como:
  - La norma del vector de gradiente.
  - La variación del vector de coeficientes.
  - La pérdida media sobre un subconjunto fijo de validación.
- Realice un análisis de sensibilidad con respecto a los hiperparámetros clave:
  - La tasa de aprendizaje $\alpha$
  - El tamaño del mini-lote.
- Evalúe el comportamiento del algoritmo en términos de:
  - Velocidad de convergencia.
  - Estabilidad de los estimadores obtenidos.
  - Robustez frente a diferentes configuraciones.

**4. Validación e Interpretación**

- Una vez finalizado el proceso de optimización, presente el vector de coeficientes estimado $\hat{\beta}$
- Entrene un modelo logístico estándar sobre una muestra aleatoria que sí pueda cargarse completamente en memoria, y compare los coeficientes obtenidos con los generados por el algoritmo online.
- Interprete los coeficientes en términos de su relación con la probabilidad de incumplimiento, indicando el efecto de cada predictor sobre la razón de *odds* (odds ratio).

**Datos Sugeridos**

Se recomienda utilizar los datos públicos de desempeño de préstamos unifamiliares a tasa fija de Fannie Mae: **Single-Family Fixed-Rate Loan Performance Data**.

## Instrucciones de Acceso:

1. **Portal de Datos**:
   Acceder al sitio oficial: https://capitalmarkets.fanniemae.com

2. **Estructura de los Archivos:**
   - **Acquisition Data**: Contiene información al momento de originar el préstamo (e.g., puntaje de crédito, monto del préstamo, etc.).
   - **Performance Data**: Contiene el historial mensual del préstamo (e.g., estado del préstamo, si está en mora, etc.).

3. **Descarga de Archivos:**
   - Seleccionar todos los archivos de adquisición y desempeño correspondientes a un año completo (por ejemplo, 2018).
   - Los archivos están disponibles en formato `.zip` por trimestre (e.g., `2018Q1.zip`).

4. **Consideraciones:**
   - Los archivos pueden ser de gran tamaño (varios GB).
   - Se recomienda verificar el espacio disponible en disco y una conexión de red estable.

## Parte del Desafío: Preparación de los Datos

- Descomprimir los archivos descargados.
- Realizar la unión de los datos de adquisición y rendimiento utilizando el identificador de préstamo (`Loan Identifier`).
- Construir la variable de respuesta binaria. Por ejemplo, un préstamo será considerado en *default* si presenta más de 90 días de morosidad en cualquier periodo de su historial.
-

## El informe deberá incluir:

- La derivación analítica del gradiente estocástico.
- La descripción e implementación del algoritmo.
- El análisis de convergencia y sensibilidad.
- La validación cruzada con un modelo estándar.
- La interpretación de resultados en el contexto de riesgo de crédito.