Tarea 1: Fundamentos y Preparación de los Datos

Dr. Marco Antonio Aquino López
Maestría en Probabilidad y Estadística,
CIMAT

Agosto–Diciembre 2025

Parte Teórica

1. **Hat Matrix y propiedades algebraicas.**
   Demuestre que la matriz
   $$H = X(X^\top X)^{-1}X^\top$$
   es idempotente y simétrica. Explique por qué estas propiedades son fundamentales para la interpretación de los *leverages*.

2. **Suma de leverages.**
   Muestre que para un modelo lineal con n observaciones y p parámetros se cumple
   $$\sum_{i=1}^{n} h_{ii} = p.$$
   Interprete este resultado en términos del "número efectivo de parámetros" y discuta su relación con el sobreajuste.

3. **Distribución de los residuos estandarizados.**
   Bajo el modelo lineal clásico con errores normales, demuestre que los residuos estandarizados
   $$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$
   tienen, aproximadamente, distribución t de Student con $n-p-1$ grados de libertad. Explique cómo esta propiedad justifica su uso en la detección de outliers.

4. **Factorización bajo MCAR.**
   Partiendo de la definición de MCAR, pruebe formalmente que
   $$p(Y, R \mid \theta, \psi) = p(Y \mid \theta)\, p(R \mid \psi).$$
   Concluya por qué en este caso el mecanismo de faltantes es ignorable para la inferencia sobre θ.

5. **Insesgadez bajo eliminación de casos (MCAR).**
   Sea $\bar{Y}_{\text{obs}}$ la media muestral basada solo en los casos observados. Demuestre que
   $$\mathbb{E}[\bar{Y}_{\text{obs}}] = \mu$$
   bajo MCAR. Discuta por qué, a pesar de ser insesgado, este estimador pierde eficiencia.

6. **Factorización bajo MAR.**
A partir de la definición de MAR, muestre que
$$L(\theta; Y_{\text{obs}}, R) \propto p(Y_{\text{obs}} \mid \theta).$$
¿Qué suposición adicional en el prior es necesaria en el enfoque bayesiano para concluir ignorabilidad?

7. **Distancia de Cook como medida global de influencia.**
Partiendo de la definición
$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p\hat{\sigma}^2},$$
muestre que se puede reescribir en función de los residuos estandarizados y el leverage como
$$D_i = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1 - h_{ii}}.$$
Discuta la interpretación de esta forma alternativa.

8. **Invarianza afín en Min–Max**
Sea $x_1, \ldots, x_n$ un conjunto de datos y defina la transformación
$$x_i^* = \frac{x_i - \min(x)}{\max(x) - \min(x)}.$$
Pruebe que si $y_i = ax_i + b$ con $a > 0$, entonces $y_i^* = x_i^*$.

9. **Transformación logarítmica y reducción de colas** Considere $X \sim \text{Pareto}(\alpha, x_m)$ con densidad
$$f(x) = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}, \quad x \geq x_m > 0, \quad \alpha > 0.$$
Defina la transformación $Y = \log(X)$.
   a) Encuentre la distribución de Y y su función de densidad.
   b) Discuta cómo cambia el comportamiento de la cola al pasar de X a Y.
   c) Explique por qué la transformación logarítmica "acorta" colas largas y produce distribuciones más cercanas a la simetría.

10. **Robustez de la mediana vs. la media**
Considere $x = \{1, 2, 3, 4, M\}$ con $M \to \infty$.
   a) Calcule la media $\bar{x}$ y la desviación estándar s como función de M.
   b) Calcule la mediana m y el rango intercuartílico RIQ.
   c) Analice: ¿qué medidas permanecen estables y cuáles se distorsionan al crecer M?

11. **Propiedades de la transformación Box–Cox**
Sea $y(\lambda)$ la transformación de Box–Cox definida como:
$$y(\lambda) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log(x), & \lambda = 0, \end{cases} \qquad x > 0.$$

a) Demuestre que $\lim_{\lambda \to 0} y(\lambda) = \log(x)$.

b) Proponga un ejemplo numérico donde x toma valores muy dispersos y compare el efecto de $\lambda = 1$ (sin transformación) frente a $\lambda = 0$ (logaritmo).

12. **Propiedades del histograma**

Sea $x_1, \ldots, x_n$ una muestra i.i.d. de una variable aleatoria continua con densidad $f(x)$. Considere el histograma con k intervalos de ancho h y estimador:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}\{x_i \in I_j\}, \quad x \in I_j.$$

a) Pruebe que $\hat{f}_h(x) \geq 0$ para todo x.

b) Demuestre que $\int_{-\infty}^{\infty} \hat{f}_h(x)\, dx = 1$.

c) Discuta cómo afecta al histograma elegir h muy grande o muy pequeño en términos de sesgo y varianza.

13. **Ejercicio: Estimación de densidad kernel (KDE)** Sea

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

con kernel K integrable, $\int K(u)\, du = 1$, $\int uK(u)\, du = 0$, y segundo momento finito $\mu_2(K) = \int u^2 K(u)\, du$.

- **Normalización**: Demuestre que $\int_{-\infty}^{\infty} \hat{f}_h(x)\, dx = 1$.
- **No negatividad**: Muestre que $\hat{f}_h(x) \geq 0$ si $K(u) \geq 0$ para todo u.
- **Sesgo puntual**: Usando expansión de Taylor de f alrededor de x, derive que

$$\mathbb{E}\{\hat{f}_h(x)\} - f(x) \;=\; \frac{h^2}{2}\, \mu_2(K)\, f''(x) + o(h^2).$$

Proyecto en Equipo

Objetivo:

Aplicar los conceptos de limpieza, imputación, codificación, escalamiento y visualización de datos en un caso real, utilizando la base disponible en https://doi.org/10.5880/GFZ.4.3.2023.002.

El trabajo debe mostrar no solo la implementación técnica, sino también la justificación teórica de cada decisión. Además, se espera que los alumnos investiguen y documenten:

- El origen de los datos: ¿qué institución o proyecto los generó?
- El propósito de la recolección: ¿qué fenómenos científicos buscan representar?
- El significado de las variables medidas: ¿qué información aportan y cómo se relacionan con el contexto dentro del contexto en que fueron capturados?
- Posibles usos e interpretaciones de la base de datos: ¿qué tipo de conclusiones o hipótesis podrían extraerse a partir de su análisis?

Instrucciones Generales:

- El proyecto se realizará en equipos de 3 estudiantes.
- El código deberá entregarse en Python, con comentarios claros.
- Se debe incluir un **informe escrito** (máximo 10 páginas) con explicaciones formales, gráficos y conclusiones.
- Toda afirmación debe estar sustentada ya sea con resultados numéricos, propiedades estadísticas o literatura de referencia.

Lineamientos del proyecto:

1. **Exploración inicial de los datos.** Describan la base de datos: número de variables, número de observaciones, tipos de datos, origen. Clasifiquen cada variable según su escala de medición. Justifiquen cada clasificación.
2. **Detección de problemas en los datos.** Identifiquen:
   - Valores faltantes: cuantifiquen el porcentaje por variable.
   - Posibles outliers: usen al menos dos métodos, de la interpretación adecuada de estos resultados.
   - Inconsistencias o codificación ambigua (ejemplo: valores categóricos mal escritos).
3. **Manejo de datos faltantes.** Discuta que clase de datos faltantes se tienen.

   Seleccionen al menos *dos estrategias* distintas de imputación (si es que esto fuera posible).

   a) Demuestren formalmente por qué bajo MCAR la eliminación de casos completos es insesgada, pero menos eficiente.

   b) Justifiquen cuál estrategia es más apropiada para esta base en términos prácticos y teóricos.

4. **Codificación y escalamiento.**

   a) Si alguna variable categórica requiere transformación, apliquen codificación (one-hot o label encoding). Justifiquen su elección.

   b) Escalen al menos dos variables numéricas usando tanto normalización min-max como estandarización z-score. Comparen los resultados y discutan en qué contextos conviene cada enfoque.

5. **Visualización exploratoria.** Generen al menos tres visualizaciones que permitan diagnosticar:

   - La distribución de una variable continua.
   - La relación entre dos variables (gráfico de dispersión, incluyendo datos imputados).
   - La presencia de outliers o valores extremos.

   Acompañen cada visualización con una interpretación clara.

6. **Reflexión crítica.** Respondan: ¿Cómo influyen las decisiones de limpieza, imputación, codificación y escalamiento en la etapa posterior de modelado estadístico? Den ejemplos concretos del dataset trabajado.

Entrega

- **Código reproducible**: en un archivo `.py`.
- **Informe en PDF**: máximo 10 páginas, incluyendo gráficas y referencias.
- **Un repositorio GitHub**: con un archivo *README.md* y muestra del uso de la herramienta por todo los miembros del equipo.
- **Fecha límite**: (Viernes, 12 de Septiembre 2025 - antes del medio día).