

# Ejercicios: Modelos de Regresión Generalizada

## Modelos Estadísticos de Predicción

AUTHOR

Víctor Aceña Gil - Isaac Martín de Diego

PUBLISHED

September 3, 2025

### Ejercicio 1: Conceptual (Fundamentos de GLM)

Explica los tres componentes clave que definen a cualquier Modelo Lineal Generalizado (GLM) y describe brevemente la función de cada uno.

---

### Ejercicio 2: Conceptual (Función de Enlace)

¿Cuál es el propósito fundamental de la función de enlace en un GLM? ¿Por qué la regresión lineal clásica es considerada un caso particular de un GLM? (Pista: piensa en su función de enlace).

---

### Ejercicio 3: Práctico (Ajuste de un Modelo Logístico)

Usa el conjunto de datos `mtcars` de R. La variable `am` indica si la transmisión de un coche es automática (0) o manual (1).

- a) Ajusta un modelo de regresión logística para predecir la probabilidad de que una transmisión sea manual (`am`) en función del peso del coche (`wt`) y los caballos de fuerza (`hp`).
  - b) Utiliza la función `summary()` para examinar el modelo. ¿Qué variables parecen ser significativas?
  - c) Obtén los coeficientes del modelo. ¿Cómo interpretarías el signo del coeficiente para la variable `wt`?
- 

### Ejercicio 4: Interpretación (Odds Ratios)

Basado en el modelo del ejercicio anterior:

- a) Calcula el Odds Ratio (OR) para el coeficiente de la variable `hp`.
  - b) Interpreta este Odds Ratio en el contexto del problema. Específicamente, ¿cómo cambian las "odds" (la razón de probabilidad) de tener una transmisión manual por cada caballo de fuerza adicional, manteniendo el peso constante?
-

## Ejercicio 5: Práctico (Validación del Modelo Logístico)

Continuando con el modelo logístico de `mtcars`:

- a) Genera las predicciones de probabilidad del modelo para los datos.
  - b) Convierte estas probabilidades en clases ("0" o "1") usando un umbral de decisión de 0.5.
  - c) Crea la matriz de confusión comparando las predicciones con los valores reales.
  - d) Calcula la precisión (accuracy) global del modelo.
  - e) (Bonus) Utiliza el paquete `pROC` para calcular y visualizar la curva ROC y obtener el valor del AUC. ¿Qué tan buena es la capacidad discriminativa del modelo?
- 

## Ejercicio 6: Conceptual (Regresión de Poisson)

- a) ¿Qué tipo de variable respuesta está diseñada para modelar la regresión de Poisson?
  - b) ¿Cuál es el supuesto fundamental de la distribución de Poisson respecto a la relación entre la media y la varianza?
  - c) ¿Cómo se llama el problema que surge cuando este supuesto se viola y la varianza es mayor que la media?
- 

## Ejercicio 7: Práctico (Ajuste de un Modelo de Poisson)

El dataset `discoveries` de R es una serie temporal que cuenta el número de "grandes inventos" por año.

- a) Crea un gráfico de la serie temporal. ¿Parece la media del conteo constante a lo largo del tiempo?
  - b) Ajusta un modelo de regresión de Poisson simple donde `discoveries` es la respuesta y el tiempo (`time(discoveries)`) es el predictor.
  - c) Interpreta el coeficiente del tiempo. (Pista: recuerda exponenciarlo para obtener el Incidence Rate Ratio - IRR).
- 

## Ejercicio 8: Diagnóstico (Sobredispersión)

- a) Para el modelo de Poisson del ejercicio anterior, calcula el estadístico de dispersión ( $\hat{\phi}$ ). (Pista:  $\hat{\phi} = \frac{\sum r_i^2}{n-p}$ , donde los  $r_i$  son los residuos Pearson).
  - b) Basándote en el valor de  $\hat{\phi}$ , ¿hay evidencia de sobredispersión?
  - c) Si encuentras sobredispersión, ¿cuál es el modelo alternativo que proponen los apuntes? ¿Qué ventaja teórica ofrece este modelo alternativo?
-

### Ejercicio 9: Conceptual (Deviance)

La deviance es la medida principal de bondad de ajuste en los GLM. Explica conceptualmente qué mide. ¿Cómo se utiliza la diferencia en deviance entre dos modelos anidados para decidir cuál es mejor?

---

### Ejercicio 10: Elección del Modelo Adecuado

Para cada uno de los siguientes escenarios, indica qué tipo de GLM (Logístico, Poisson, Binomial Negativo, Gamma...) sería el más apropiado y por qué.

- a) Quieres modelar el tiempo (en minutos) que tarda un cliente en resolver una consulta en un centro de atención telefónica. El tiempo es siempre positivo y muchos valores se agrupan en tiempos cortos, con una cola larga de tiempos muy largos.
- b) Quieres predecir la presencia o ausencia de una especie de planta en diferentes parcelas de un bosque.
- c) Quieres modelar el número de visitas que cada usuario hace a una página web en un mes. Observas que la varianza del número de visitas es mucho mayor que la media.