

Taller Práctico Regresión Logística (2) *

Estadística II *Universidad Nacional de Colombia, Sede Medellín*

Este documento corresponde al séptimo taller práctico del curso de **Estadística II** para la *Universidad Nacional de Colombia*, Sede Medellín, en el periodo 2025 - 1. Se brinda una introducción al análisis de regresión. El enfoque de este taller está en la evaluación del rendimiento de la regresión logística. **Monitor:** *Santiago Carmona Hincapié*.

Keywords: regresión múltiple, secuencialidad, interacción

Información general

Con el propósito de profundizar en los conceptos del modelo de regresión logística vistos en clase, se propone afrontar este taller en dos partes, una de teoría básica y otra práctica.

La solución para cada uno de los problemas se efectúa a partir del software estadístico R.

Parte teórica

De respuesta a las preguntas formuladas a continuación en base a la teoría tratada en clase.

Provea una interpretación de ser necesario.

1. Determine el valor de verdad de las siguientes afirmaciones.

- (a) El modelo de regresión logística $\text{logit}(\theta_i) = \psi(\mathbf{x}_i)$ clasifica directamente como éxito o fracaso cada observación provista.
- (b) La tasa de falsos positivos se calcula como $[n_{01}/(n_{00} + n_{01})]$, e indica las observaciones que fueron clasificadas como positivas cuando eran realmente verdaderas.
- (c) La curva ROC muestra cómo cambian la sensibilidad y las especificidad a medida que varía el punto de corte θ^* . De aquí se deriva una métrica importante, que es el AUC.
- (d) La selección del punto de corte θ^* impacta directamente el desempeño del modelo logístico como clasificador, -incluso la matriz de confusión se ve afectada-.
- (e) Un modelo con un AUC igual a 0.5 indica que clasifica las observaciones de manera aleatoria.

*El material asociado a este taller puede encontrarse en el repositorio del curso, (<https://github.com/Itssach/Estadistica-II>)

Ejercicio con datos reales

Considere el siguiente conjunto de datos que agrupa una serie de características enfocadas en *clasificar* la personalidad de múltiples individuos. **Se incluyen variables cuantitativas y cualitativas.** La información puede profundizarse en: <https://www.kaggle.com/datasets/rakeshkapilavai/extrovert-vs-introvert-behavior-data>

Table 1: Información en análisis

Personality	Time_spent_Alone	Stage_fear	Social_event_attendance	Going_outside
Extrovert	4	No	4	6
Introvert	9	Yes	0	0
Introvert	9	Yes	1	2
Extrovert	0	No	6	7
Extrovert	3	No	9	4

Considere a ‘*Personality*’ como la variable respuesta. *Algunas de las covariables en análisis se especifican en la tabla mostrada con anterioridad.* **De respuesta a los siguientes planteamientos:**

1. Analice el rendimiento del modelo a través de la curva ROC y AUC. Interprete.
2. Halle un valor de corte θ^* óptimo que permita mantener una **tasa de clasificación acertada alta**, minimizando el número de falsos negativos.
3. Reporte la matriz de confusión y la tasa de clasificación correcta. ¿Cómo cambian estas matrices con diferentes valores de θ^* ?