

Desafío - Clasificación desde la econometría

- Para poder realizar esta actividad debes haber revisado la lectura y videos correspondiente a la unidad.
- Crea una carpeta de trabajo y guarda todos los archivos correspondientes (notebook y csv).
- Una vez terminado el desafío, comprime la carpeta y sube el `.zip`

Descripción

En esta sesión trabajaremos con:

- `sbp` : Presión Sanguínea Sistólica.
- `tobacco` : Promedio tabaco consumido por día.
- `ldl` : Lipoproteína de baja densidad.
- `adiposity` : Adiposidad.
- `famhist` : Antecedentes familiares de enfermedades cardíacas. (Binaria)
- `types` : Personalidad tipo A
- `obesity` : Obesidad.
- `alcohol` : Consumo actual de alcohol.
- `age` : edad.
- `chd` : Enfermedad coronaria. (dummy)

Desafío 1: Preparar el ambiente de trabajo

- Se detallan los pasos a seguir
- *tip*: Los tips o sugerencias preceden de *tip*
- Se generan dos notebooks, uno con las soluciones y otro con los ejercicios.

Desafío 2

A continuación se presenta el siguiente modelo a estimar:

$$\log\left(\frac{\Pr(\text{chd} = 1)}{1 - \Pr(\text{chd} = 1)}\right) = \beta_0 + \beta_1 \cdot \text{famhist}$$

Para ello ejecute los siguientes pasos:

1. Recodifique `famhist` a dummy, asignando 1 a la categoría minoritaria.
2. Utilice `smf.logit` para estimar el modelo.
3. Implemente una función `inverse_logit` que realice el mapeo de log-odds a probabilidad.
4. Con el modelo estimado, responda lo siguiente:
 - ¿Cuál es la probabilidad de un individuo con antecedentes familiares de tener una enfermedad coronaria?
 - ¿Cuál es la probabilidad de un individuo sin antecedentes familiares de tener una enfermedad coronaria?
 - ¿Cuál es la diferencia en la probabilidad entre un individuo con antecedentes y otro sin antecedentes?
 - Replique el modelo con `smf.ols` y comente las similitudes entre los coeficientes estimados.
tip: Utilice $\beta/4$
 - Estime el mismo modelo con LPM

Desafío 3: Estimación completa

Implemente un modelo con la siguiente forma

$$\log\left(\frac{\Pr(\text{chd} = 1)}{1 - \Pr(\text{chd} = 1)}\right) = \beta_0 + \sum_{j=1}^N \beta_j \cdot X_j$$

- Depure el modelo manteniendo las variables con significancia estadística al 95%.
- Compare los estadísticos de bondad de ajuste entre ambos.
- Reporte de forma sucinta el efecto de las variables en el log-odds de tener una enfermedad coronaria.

Desafío 4: Estimación de perfiles

A partir del modelo depurado, genere las estimaciones en log-odds y posteriormente transfórmelas a probabilidades con `inverse_logit`. Los perfiles a estimar son los siguientes:

- La probabilidad de tener una enfermedad coronaria para un individuo con características similares a la muestra.
- La probabilidad de tener una enfermedad coronaria para un individuo con altos niveles de lipoproteína de baja densidad, **manteniendo todas las demás características constantes**.
- La probabilidad de tener una enfermedad coronaria para un individuo con bajos niveles de lipoproteína de baja densidad, **manteniendo todas las demás características constantes**.