Pontificia Universidad Católica de Chile
Facultad de Matemáticas - Departamento de Estadística
Profesora: Ana María Araneda
Ayudantes: Eduardo Vásquez y Vanesa Reinoso
Correos: evasquezt@uc.cl y vcreinoso@mat.uc.cl

EPG3306 - Métodos Estadísticos I
Ayudantía 10
18 de junio del 2022

Contenidos:

- Regresión logística
- Selección de modelos

---

1. La base de datos `breast-cancer-wisconsin.data`, publicada por el doctor William Wolberg, del hospital de la Universidad de Wisconsin, contiene información de diferentes pacientes del hospital que se han hecho exámenes de cáncer de mamas, a través de los años. En estos exámenes se toman diferentes medidas, con lo cual se desea predecir si el cancer es benigno o maligno. Las variables corresponden a:

   - **ct**: espesor de la masa (clump thickness)
   - **ucsize**: uniformidad del tamaño de las células (uniformity of cell size)
   - **ucshape**: uniformidad de la forma de las células (uniformity of cell shape)
   - **ma**: adhesión marginal (marginal adhesion)
   - **sepcs**: tamaño de célula epitelial única (single epithelial cell size)
   - **bn**: bare nuclei
   - **bc**: cromatina blanda (bland chromatin)
   - **nc**: nucleolos normales (normal nucleoli)
   - **mitoses**: mitosis
   - **class**: indica si es benigno o maligno (2 y 4, respectivamente)

   (a) Cargue la base de datos en R y realice cualquier preprocesamiento necesario. Realice gráficos de cada variable con la variable `class`. ¿Se observa que alguna variable **por sí sola** sea un buen predictor?

   (b) Ajuste un modelo de regresión logística con la función de enlace logit, con todos las variables y evalúe la hipótesis H_0: el modelo de regresión propuesto es correcto versus H_1: el modelo saturado es correcto. Compare también con el modelo nulo.

(c) **(Extra)** Ajuste un modelo de regresión logística con la función de enlace probit, con todos las variables y evalúe la hipótesis H_0: el modelo de regresión propuesto es correcto versus H_1: el modelo saturado es correcto. Compare también con el modelo nulo.

(d) Realice una selección de modelos forward utilizando regresión logística. ¿Qué variables quedan?, comente con respecto a lo visto en (a). Considere $\alpha = 0.05$.

(e) **(Extra)** Realice una selección de modelos forward utilizando regresión probit. ¿Qué variables quedan?, comente con respecto a lo visto en (a) y compare con lo obtenido en (d). Considere $\alpha = 0.05$.

(f) **(Extra)** Entre el modelo de regresión logística y el probit, ¿con cuál se quedaría?

(g) Con el modelo elegido anteriormente, y con diferentes puntos de corte, obtenga tanto la sensibilidad como la especificidad del modelo. ¿Qué valores les parecen mejores con respecto al contexto del problema?

(h) Gráfique la curva ROC y obtenga el área bajo la curva.