PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
Facultad de Matemática
Diplomado en Estadística
Profesor: Jonathan Acosta
Fecha Entrega: Miércoles 21 de agosto a las 18:00 horas

Taller Evaluado de Regresión Logística

Objetivo: Ajustar un modelo de regresión logística en **R** e interpretar sus resultados.

Instrucciones: Para este taller pueden trabajar en parejas y se utilizará el conjunto de datos 'prostate' de la librería 'faraway'.

Este conjunto de datos contiene información sobre 97 pacientes con cáncer de próstata. Considere la variable respuesta como 'svi' (seminal vesicle invasion). Esta variable indica si el cáncer se ha extendido a la vesícula seminal (1) o no (0). Las demás variables serán consideradas como variables explicativas. El objetivo es ajustar un modelo de regresión logística para predecir la invasión de la vesícula seminal.

Realizar una descomposición aleatoria de la base de datos con la proporción 90%-10% para train y test, respectivamente.

1. Realice un análisis de inflación de varianzas (VIF) y de ser necesario elimine todas aquellas necesarias de modo de garantizar un vif menor a 8.
2. Utilizando la data 'prostate.train' realice una gráfica apropiada entre la variable respuesta y cada una de las covariables. Según esta perpespectiva gráfica, ¿Existe alguna de ellas que pueda explicar la variable respuesta?
3. Utilizando el criterios de Akaike (AIC), la metodología stepwise (puede ser backward, forward o both, indique explícitamente cuál utilizará) y la función de enlace 'logit' para determinar el modelo de regresión logística que mejor ajusta a la variable respuesta.
4. Analice la significancia del modelo obtenido luego del proceso de selección, y responda si
   (a) ¿Es el modelo obtenido significativo?
   (b) ¿Existe alguna covariable no significativa?
   (c) ¿En caso de existir alguna covariable no significativa, la quitaría del modelo?. Fundamente.
   (d) Utilice los odd-ratios para interprete las variables del modelo final.
5. Repita el proceso de las preguntas 3 y 4 con la función de enlace 'probit'.
6. Utilice algún criterio apropiado para definir cuál de los dos modelos finales (logit o probit) es el mejor.
7. Realice la predicción para los datos de la muestra 'prostate.test' con ambos modelos. Incluya un intervalo de confianza para las predicciones.
8. Utilice el punto de corte 0.5 para realizar la clasificación. Reporte las dos matrices de confusión (una de cada modelo). Utilice la exactitud (accuracy) para indicar cuál modelo es mejor. ¿El modelo mejor es el mismo indicado en la pregunta 6?

Observación: En caso de no tener la librería instalada, siga las siguientes instrucciones:

- install.packages("faraway") para instalar la librería
- library(faraway) para cargar la librería.