UNIVERSIDAD NACIONAL DE COLOMBIA

FACULTAD DE CIENCIAS, SEDE BOGOTÁ
DEPARTAMENTO DE ESTADÍSTICA
ANÁLISIS DE REGRESIÓN

Ejercicio en clase 7. Modelos lineales generalizados: modelo de regresión logística

1. Considere la base de datos **burn1000** incluida en el paquete **aplore3**. Esta base contiene la información de 1000 pacientes tratados por quemaduras en diferentes instituciones. La base también incluye la información de si el paciente falleció o no a causa de las lesiones.

a) Revise cómo está constituida la base de datos usando estadística descriptiva.
b) Construya un primer modelo de regresión logística que busque explicar los decesos (éxito) en función de la edad (age), el género (gender), la raza (race), el porcentaje de área quemada (tbsa), la presencia de heridas por inhalación (inh_inj) y la presencia de flama en el episodio (flame).
c) Revise el valor de devianza. ¿Qué dice este valor sobre el modelo? Calcule el pseudo R^2 ajustado e interprete dicho valor.
d) Calcule los valores predichos para todos los individuos de la muestra. ¿Cómo lucen dichos valores? Compruebe que esos valores son los mismos que se obtienen si se hace el calo manual con las estimaciones de los parámetros. Grafique los valores predichos versus los valores observados y comente.
e) Interprete las estimaciones puntuales y por intervalo del modelo en términos de razones de chances.
f) Estime modelos variando la función de enlace. Use criterios de información para establecer cuál es la “mejor” función de enlace.
g) Considere un modelo con interacciones entre age e inh_inj, y tbsa e inh_inj. Use la selección hacia atrás (backward) y el BIC para seleccionar el “mejor” modelo.
h) Con la especificación del modelo anterior, revise cómo cambia el ajuste cuando se cambia la función de enlace.
i) Con el mejor modelo del punto anterior, revise el cumplimiento de supuestos.

2. Considere el conjunto de datos **Smarket** del paquete **ISLR2**. Este conjunto de datos contiene información salarial de 1250 movimientos de bolsa. Las 9 variables medidas son: el año de la observación (Year), el porcentaje de retorno del día anterior (Lag1), el porcentaje de retorno de dos días atrás (Lag2), el porcentaje de retorno de tres días atrás (Lag3), el porcentaje de retorno de cuatro días atrás (Lag4), el porcentaje de retorno de cinco días atrás (Lag5), el volumen de acciones tranzadas en billones (Volume), el porcentaje de retorno del día en cuestión (Today), la dirección del mercado el día en cuestión (Direction), es decir si el mercado va al alza (“Up”) o a la baja (“Down”).

a) Explore el conjunto de datos. Haga gráficos de dispersión por pares de variables.
b) Construya un modelo de regresión logística que prediga la dirección del mercado. No use la variable (Today), ya que claramente esa variable es la que determina la dirección.
c) Construya la matriz de confusión de ese modelo usando como límite de clasificación 0.5. ¿Diría que es un modelo que captura la dinámica del mercado? Calcule la tasa de error aparente.

d) Particione el conjunto de datos de modo que los datos antes del 2005 se usen para entrenar el modelo y el resto se usen para testearlo. Construya la matriz de confusión sobre la parte de testeo usando 0.5 como límite de clasificación. Calcule la tasa de error aparente.
e) Con el mismo esquema de validación, construya un modelo con interacciones hasta de orden 6. ¿Cómo es su desempeño en términos de la matriz de confusión?
f) Compare el desempeño general de ambos modelos, usando la curva ROC.
g) Para el modelo seleccionando en el item anterior, escoja el valor óptimo de tau, monitoreando varios indicadores: AER, recall, precisión, F1.

3. Retome la base de datos del punto 1.

a) Verifique si la proporción de casos no está balanceada. Haga entonces un muestreo estratificado para realizar una partición de los datos.
b) Entrene nuevamente el modelo descrito en el punto 1-g) y sus variantes, cambiando la función de enlace.
c) Revise el desempeño global de estos modelos sobre la parte de testeo usando la curva ROC.
d) Para el modelo con el mejor desempeño en c), haga una selección óptima del valor de tau.