Modelo Lineal Generalizado

Práctica 4

Ejercicio 1. Consideremos el conjunto de datos dados en el archivo **birthwt**. Los datos corresponden a un estudio de factores de riesgo bajo peso de los recién nacidos. Los datos fueron recogidos en el Baystate Medical Center, Springfield, Massachusetts, en 1986.
El siguiente cuadro describe las variables consideradas.

Variable	**Nombre**
Código de identificación	ID
Bajo peso al nacer (0= peso al nacer ≥ 2500gr) (1= peso al nacer < 2500gr)	LOW
Edad de la madre en años	AGE
Peso en libras en la última menstruación	LWT
Raza(1=Blanca, 2=Negra, 3=Otros)	RACE
Fuma durante el embarazo (1= Si, 0=No)	SMOKE
Historia de trabajo prematuro (premature labor)(0=ninguna, 1=uno, 2=dos, etc.)	PTL
Historia de hipertensión (1=si, 0=no)	HT
Presencia de irritación uterina (1=si, 0=no)	UI
Número de visitas al médico durante el primer trimestre (0=ninguna, 1=uno, 2=dos, etc.)	FTV
Peso al nacer en gramos	BWT

a) Defina las variables factoriales y defina los contrastes para las mismas (en el caso de RACE tome la categoría 1 como referencia).

b) Realice un ajuste univariado para cada una de las variables independientes (excepto a BWT) tomando a LOW como variable de repuesta.

c) Observe cuáles de las variables consideradas son significativas, calcule un estimador del odds ratio e intervalos de confianza de nivel asintótico 0.95 para los mismos.

d) Observe los valores de los estadísticos de bondad de ajuste y su significación. ¿Cuáles son sus conclusiones?

e) Realice un ajuste considerando en su modelo las variables AGE, LWT, RACE, SMOKE, PTL, HT, UI y una constante. Analice la significación de cada una de las variables? ¿Es AGE significativa? Evalúe la significación del modelo.

f) Repita e) sin la variable AGE. ¿Cuáles son sus conclusiones?

g) Se sabe que la variable AGE es biológicamente importante. Es posible que el logit de la variable AGE sea una función cuadrática o con forma de "U". Realice un análisis de cuartiles de la variables AGE para evaluar esta posibilidad. De acuerdo con este análisis le parece razonable suponer que hay una tendencia lineal o cuadrática significativa para esta variable?

h) Repita g) para la variable LWT. ¿le parece que hay evidencias de linealidad o se sostiene un modelo dicotómico para esta variable?

i) A partir de LWT cree una variable dicotómica LWD que tome valores 1 en el primer cuartil y 0 en otro caso. Calcule el odds ratio para esta variable y un intervalo de confianza de nivel 0.95 para la misma. Interprete el resultado.

j) Dado que la variable PTD tiene muy pocos casos con valores superiores a 1, consideraremos una nueva variable dicotómica PTD definida como 0 si PTL=0 y 1 en caso contrario. Calcule el odds ratio para esta variable y un intervalo de confianza de nivel 0.95 para la misma. Interprete el resultado.

k) Realice un nuevo ajuste incluyendo las variables: AGE, LWD, RACE, SMOKE, PTD, HT, UI y constante. ¿Puede compararlo con el ajuste realizado en e)¿. ¿Son los modelos anidados? Evalúe el ajuste del modelo y la significación de cada variable.

Ejercicio 2. Consideremos el conjunto de datos dados en el archivo **vaso**. Los datos fueron obtenidos en un estudio controlado para medir el efecto de la velocidad (RATE) y el volumen (VOLUME) del aire inspirado en la vasoconstricción transitoria de la piel de los dedos. La naturaleza del proceso de medición hizo que sólo se registrara la ocurrencia o no de la vasoconstricción (Y).

a) Realice un ajuste a los datos usando las dos variables independientes RATE y VOLUME. Compute los residuos y los valores predichos.

b) Realice un gráfico de valores predichos vs. residuos ¿Qué observa? Identifique las observaciones que tienen algún comportamiento extraño.

c) Compute los leverage y realice un gráfico de nro. de observación vs. leverage y de leverage vs. residuos. ¿Qué observa? Identifique las observaciones que tienen algún comportamiento extraño.

d) Realice un nuevo ajuste sin las dos observaciones con residuos más grandes. ¿Influyen en la determinación de los estimadores?

e) Realice un nuevo ajuste sin la observación con leverage más grande. ¿Influye en la determinación de los estimadores?

Ejercicio 3. En la siguiente tabla se consignan los datos de admisión en 6 departamentos de U. C. de Berkeley en el otoño de 1973:

Departamento	Hombres Rechazados	Hombres Aceptados	Mujeres rechazadas	Mujeres aceptadas
A	313	512	19	89
B	207	353	8	17
C	205	120	391	202
D	278	139	244	131
E	138	53	299	94
F	351	22	317	24

Schaffer(2000)

En el archivo berkeley.txt se encuentran los datos. Definamos como **éxito** que el individuo es **aceptado** en Berkeley y definamos las variables SEXO y DPTO.
Notemos que el máximo modelo que incluye intercept, efectos principales de SEXO, DPTO y sus interacciones tiene 12 parámetros y por lo tanto es saturado. Ajustaremos algunos modelos buscando uno que ajuste satisfactoriamente.

a) Ajuste un modelo que contenga sólo la intercept (Null model o modelo nulo). Observe cuánto valen los valores predichos y con quien coinciden. Evalúe el ajuste del modelo a través de la deviance.

b) Ahora agregue la variable SEXO a su modelo (Identifique a Mujeres con 0 y Hombres con1). ¿Son los coeficientes estimados significativos? Evalúe el ajuste global del modelo a través de la deviance y del estadístico de Pearson. Este último puede computarlo mediante la suma de cuadrados de los residuos de Pearson que da como resultado la función S-plus **residuals.glm** (residuals.glm(salida,type= aquí se puede elegir "deviance"o "pearson" entre otros).
Compare este modelo con el ajustado en a) usando el test basado en cociente de verosimilitud. ¿A qué conclusión llega?
¿Cómo se interpreta la pendiente en este caso?

c) Considere el modelo usando además de la intercept la variable DPTO como predictor.
Para este caso use la codificación que resulta habitual en ANOVA en que la suma de los efectos es 0 y que corresponde a las variables dummies:

	A	**B**	**C**	**D**	**E**
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1
F	-1	-1	-1	-1	-1

Identifique los efectos significativos. Evalúe la significación conjunta de los efectos comparando con el modelo nulo y el ajuste global del modelo a través de la deviance y del estadístico de Pearson. ¿Parece mejor que el ajuste anterior? ¿El modelo ajusta?

d) Ahora combine los dos modelos anteriores, es decir ajuste un modelo con intercept y las variables SEXO y DPTO. ¿Es SEXO significativo? ¿El modelo ajusta globalmente?
¿En este modelo se supone que la relación entre las variables SEXO y DPTO cambia entre departamentos o no?

Al calcular el test de bondad de ajuste para este modelo se compara la significación conjunta de las 5 interacciones entre SEXO y DPTO, ¿qué resultado da esto? ¿Le parece que la relación entre SEXO y DPTO es constante?

e) Para el modelo planteado en el ítem anterior examine los residuos de Pearson. ¿Qué observa? ¿Le parece adecuado el ajuste del modelo?

f) Ajuste ahora el modelo saturado, pero bajo otro esquema de codificación más fácil de interpretar. Considere el modelo sin intercept y con una variable dummy para cada departamento y agregue sus interacciones con SEXO. De esta manera la matriz de diseño en las 12 observaciones será

dA	dB	dC	dD	dE	dF	sex.A	sex.B	sex.D	sex.E	sex.F
1	0	0	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	1	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	1	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0
0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	1
0	0	0	0	0	1	0	0	0	0	0	0

En esta codificación ,¿cuál es la interpretación del coeficiente de la interacción entre SEXO con DPTO =A, es decir sex.A?

Observe cuá l es la única interacción significativa. ¿Cómo interpretaría esto?

g) Ajuste un nuevo modelo sin intercept, con una variable dummy para cada departamento y la única interacción significativa que encontró en el ítem anterior. Evalúe el ajuste del modelo y estudie los residuos del mismo. ¿Cómo se interpreta el modelo ajustado?