GENERALIZED LINEAR AND LOG-LINEAR MODELS

El GLM se define por tres componentes:

1. Variantes aleatorias independientes de respuesta $Y_1, \ldots, Y_n$ se asume que estas variantes siguen una distribución de probabilidad de la familia exponencial, con valor esperado $E[Y_i] = \mu_i$, que, en los modelos logarítmicos lineales, es la media del logaritmo de las frecuencias esperadas en las celdas; la variable de respuesta, Y, es la primera parte del componente aleatorio del GLM; la segunda parte es el residual.

2. Un predictor lineal basado en las variables predictoras $X_{i,1}, \ldots, X_{i,p-1}$ y los correspondientes parámetros, β: este es el componente sistemático de un GLM.

3. Una función de enlace monótona, g, que relaciona el predictor lineal con la respuesta esperada, $x_i'\beta = g(\mu_i)$.

Ejemplos de GLM incluyen el modelo logit binomial para datos binarios, el modelo Poisson, y el modelo lineal general para variables de resultado continuas. La función de enlace natural para la distribución de Poisson es la función logarítmica. Resulta en una componente línea

$$\log E\,[Y_i] = x_i'\beta$$

donde X_i es el vector i-ésimo de la matriz de diseño X.

Los modelos logarítmicos lineales usualmente se expresan por los parámetros involucrados en un modelo. Por ejemplo, un modelo logarítmico lineal para la clasificación cruzada de las variables A y B que considera solo los efectos principales de ambas variables, o en otras palabras, el modelo de independencia de A y B, se expresa como:

$$\log \widehat{m} = \lambda + \lambda_r^A + \lambda_s^B$$

donde $\widehat{\mathrm{m}} = \mu, \lambda$ es la constante del modelo, λ_r^A es el parámetro r-ésimo de los $k_A - 1$ parámetros de la variable A (con k_A indicando el número de categorías de A), y λ_s^B es el parámetro s-ésimo de los $k_B - 1$ parámetros de la variable B (con k_B indicando el número de categorías de B). Un modelo a menudo utilizado como base es el modelo nulo, $\log \widehat{m} = \lambda$. La comparación de modelos puede realizarse utilizando la diferencia entre los coeficientes de razón de verosimilitud G^2 entre el modelo objetivo y el modelo base. El coeficiente es:

$$G^2 = 2\sum_i m_i \log \frac{m_i}{\widehat{m_i}}$$

donde m_i es la frecuencia observada para la celda i, $\widehat{m}_i$ es la frecuencia esperada para esta celda, e i recorre todas las celdas de la clasificación cruzada. Cuando se evalúa un solo modelo, el X^2 de Pearson puede ser la mejor opción, que es:

$$X^2 = \sum_i \frac{(m_i - \widehat{m_i})^2}{\widehat{m_i}}$$

Tanto G^2 como X^2 están distribuidos aproximadamente como χ^2 con:

$$df = \text{número de celdas} - \text{número de parámetros estimados}$$