Herramientas de diagnóstico para datos binarios

No hay medidas "ideales" para evaluar el ajuste de un MLG binomial con

$$g(\pi) = \sum_{j=1}^{p} x_{ij}\beta_j.$$

Sin embargo, hay algunos resultados asintóticos que proveen aproximaciones razonables bajo ciertas circunstancias.

En esta clase, asumimos que $Y_i \sim$Binomial(m_i, π_i), donde Y_i es el número de "éxitos" (de m_i ensayos) para la i-ésima combinación de niveles de variables explicativas, $i = 1, \ldots, n$.

Pruebas de bondad de ajuste formales

Devianza. Podemos calcular el modelo saturado como sigue, definimos n parámetros $\theta_1, \ldots, \theta_n$ donde $\theta_i = \pi_i$. Luego

$$L_S(\boldsymbol{\theta}) = \prod_{i=1}^{n} \binom{m_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{m_i - y_i}$$

$$l_S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \binom{m_i}{y_i} + \sum_{i=1}^{n} y_i \log \theta_i + \sum_{i=1}^{n} (m_i - y_i) \log(1 - \theta_i)$$

$$\begin{aligned} \frac{\partial}{\partial \theta_i} l_S(\boldsymbol{\theta}) &= \frac{y_i}{\theta_i} + \frac{y_i - m_i}{1 - \theta_i} \\ 0 &= \frac{y_i}{\hat{\theta}_i} + \frac{y_i - m_i}{1 - \hat{\theta}_i} \\ \hat{\theta}_i &= \frac{y_i}{m_i} \end{aligned}$$

Sea $p_i = \hat{\theta}_i$, entonces p_i es la proporción de éxitos para la i-ésima combinación de niveles de variables explicativas. Sea $\hat{\pi}_i$ el EMV de π_i bajo el modelo propuesto (es decir, $\hat{\pi}_i = g^{-1}(\Sigma_{j=1}^{p} x_{ij}\hat{\beta}_j)$).Podemos calcular la devianza como:

$$D = 2[l_S(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\beta}})]$$

$$= 2\left[\sum_{i=1}^{n} y_i \log p_i + \sum_{i=1}^{n} (m_i - y_i) \log(1 - p_i) - \sum_{i=1}^{n} y_i \log \hat{\pi}_i - \sum_{i=1}^{n} (m_i - y_i) \log(1 - \hat{\pi}_i)\right]$$

$$= 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{p_i}{\hat{\pi}_i}\right) + (m_i - y_i) \log\left(\frac{1 - p_i}{1 - \hat{\pi}_i}\right)\right]$$

D puede ser evaluada en base a los datos

observados. Bajo la hipótesis nula que nuestro modelo ajustado se ajusta tan bien como el modelo saturado, $D \sim \chi^2_{n-p}$ asintóticamente.

Podemos también comparar el modelo propuesto al modelo nulo. El modelo nulo tiene un parámetro, θ, donde $Y_i \sim$Binomial(m_i, θ). Es fácil calcular el EMV $\hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} \equiv p$. Luego

$$\Delta D = 2[l(\hat{\boldsymbol{\beta}}) - l_N(\hat{\theta})]$$

$$= 2\left[\sum_{i=1}^{n} y_i \log \hat{\pi}_i + \sum_{i=1}^{n}(m_i - y_i)\log(1-\hat{\pi}_i) - \sum_{i=1}^{n} y_i \log p - \sum_{i=1}^{n}(m_i - y_i)\log(1-p)\right]$$

$$= 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{\hat{\pi}_i}{p}\right) + (m_i - y_i)\log\left(\frac{1-\hat{\pi}_i}{1-p}\right)\right]$$

Bajo la hipótesis nula que el modelo nulo se ajusta tan bien como el modelo propuesto, $\Delta D \sim \chi^2_{p-1}$ asintóticamente.

Finalmente, podemos comparar nuestro modelo propuesto a un modelo con un subconjunto de variables explicativas (modelo anidado).

Por ejemplo, si sacamos q variables explicativas del modelo, el cambio en la devianza (bajo la hipótesis nula de que el modelo reducido se ajusta tan bien como el modelo completo) tiene una distribución χ^2_q asintóticamente.

NOTA IMPORTANTE: que esta distribución es solamente asintótica. Por lo tanto, pruebas de hipótesis basadas en esta estadística puede ser pobre si los m_i's son pequeños (es decir hay pocas réplicas en cada grupo).

Datos:

Dose (x)	49	53	57	61	65	69	73	77
# alive	53	47	44	28	11	6	1	0
# dead (y_i)	6	13	18	28	52	53	61	60
Total (m_i)	59	60	62	56	63	59	62	60

Estadística de Pearson chi-cuadrado

Para datos provenientes de una distrbución Binomial(m_i, π_i),la estadística de Pearson chi-cuadrado está normalmente definida por:

$$X^2 \equiv \sum_{i=1}^{n} \frac{(y_i - m_i\pi_i)^2}{m_i\pi_i(1-\pi_i)}.$$

donde Y_i es el número de "éxitos" (en el ejemplo muertes) y π_i es la probabilidad de éxito en la i-ésima combinación de niveles de variables explicativas (en el ejemplo dosis).

Si loa π_i's son desconocidos bajo la hipótesis nula H_0, usamos la estadística

$$X^2 \equiv \sum_{i=1}^{n} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

Nota: X^2 y D tienen la misma distribución asintótica.

Idea de la prueba:

Usamos la expansión de Taylor de $s\log(s/t)$ alrededor de $s = t$:

$$s\log\left(\frac{s}{t}\right) = (s-t) + \frac{1}{2}\frac{(s-t)^2}{t} + \cdots.$$

Luego, de (1), tenemos

$$\begin{aligned} D &= 2\sum_{i=1}^{n}\left\{(y_i - m_i\hat{\pi}_i) + \frac{1}{2}\frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i} + \right. \\ &\quad \left. [(m_i - y_i) - (m_i - m_i\hat{\pi}_i)] + \frac{1}{2}\frac{[(m_i - y_i) - (m_i - m_i\hat{\pi}_i)]^2}{m_i - m_i\hat{\pi}_i} + \cdots\right\} \\ &\approx \sum_{i=1}^{n}\frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} \\ &= X^2 \end{aligned}$$

Por lo tanto, bajo la hipótesis nula el modelo propuesto se ajusta tan bien como el modelo saturado, $X^2 \sim \chi^2_{n-p}$.

Note que las pruebas de Devianza y Pearson son solamente asintóticas. Por lo tanto, las hipótesis basadas en estas dos estadísticas pueden dar resultados distintos para muestras pequeñas. Alguna evidencia sugiere que X^2 puede ser más próxima a una distribución χ^2_{n-p} que D.

Residuales

Los *residuales de Pearson* son definidos como

$$X_k = \frac{y_k - m_k\hat{\pi}_k}{\sqrt{m_k\hat{\pi}_k(1-\hat{\pi}_k)}}, \qquad k = 1, \ldots, n.$$

Los *residuales de devianza* son definidos como

$$d_k = \text{sign}(y_k - m_k\hat{\pi}_k)\left\{2\left[y_k \log\left(\frac{p_k}{\hat{\pi}_k}\right) + (m_k - y_k)\log\left(\frac{1-p_k}{1-\hat{\pi}_k}\right)\right]\right\}^{1/2},$$

$k = 1, \ldots, n.$

Valores grandes de los residuales indican observaciones atípicas (ya que valores grande de la devianza o de f X^2 indican una falta de ajuste en el modelo).

PRECAUCIÓN: Puede ser difícil interpretar los gráficos de estos residuales! *No* espere que ellos luzcan como los que uno espera ver en el contexto de regresión lineal – especialmente si los m_i's son pequeños! En particular, *no* deberiamos esperar a que estén normalmente distribuidos, o que tengan variance común.

Por ejemplo, considere el caso donde $m_i = 1$ (es decir tenemos datos binarios). Pretenda por un momento que conocemos el valor verdadero de los β_j's, entonces conocemos los π_i's exactamente. En este caso, podriamos calcular los residuales de Pearson como

$$X_k = \frac{y_k - \pi_k}{\sqrt{\pi_k(1-\pi_k)}}, \qquad k = 1, \ldots, n.$$

Aquí, Y_k es la única variable aleatoria, y tiene una distribución Bernoulli. Entonces, claramente, X_k no está normalmente distribuida.

En el caso usual donde los β_j's (y aquí los π_k's) son estimados, la distribución de X_k es aún más compleja.

Más grandes m_k, más cercanamente Y_k pueden ser aproximados a una distrinución normal.