MLG: Datos binarios (cont.)

Funciones de enlace para modelos binomiales

Link	$\eta = g(p_i)$	$p_i = g^{-1}(\eta)$
identity	p_i	η
logarithmic	$\log p_i$	e^{η}
logistic	$\log\left(\frac{p_i}{1-p_i}\right)$	$\frac{e^{\eta}}{1+e^{\eta}}$
probit	$\Phi^{-1}(p_i)$	$\Phi(\eta)$
log-log	$\log(-\log p_i)$	$\exp(-e^{\eta})$
complementary log-log	$\log(-\log(1-p_i))$	$1-\exp(-e^{\eta})$

Comparación de funciones de enlace

- Cuando g es la identidad o la función logarítmica, $\hat{p_i} = g^{-1}(\hat{\eta})$ puede salir del intervalo $[0, 1]$. Por lo tanto, estas funciones de enlace pueden no ser las más adecuadas.
- La logística y probit son de lejos las más usadas.
- El modelo probit requiere integración numérica en el cálculo de los EMVs (porque Φ no

tiene una forma cerrada).

- La logística y probit son simétricas respecto de $p_i = 1/2$,y producen resultados similares a menos que hayan muchas probabilidades muy pequeñas o muy grandes.
- La logística, probit y complementary log-log son similares para pequeños π.
- La complementary log-log es asimétrica.

Pruebas de hipótesis

La expansión de las series de Taylor de una función $f(x)$ alrededor del punto x^* cercano a x es

$$f(x) = f(x^*)+(x-x^*)f'(x^*)+\frac{1}{2}(x-x^*)^2 f''(x^*)+\ldots.$$

Por lo tanto, para una estimación $\hat{\beta}$ que es cercana a β, la aproximación de segundo orden de la log- verosimilitud (en el caso en que β es 1-dimensional) es

$$l(\beta) \approx l(\hat{\beta})+(\beta-\hat{\beta})U(\hat{\beta})+\frac{1}{2}(\beta-\hat{\beta})^2 U'(\hat{\beta}).$$

Podemos ademas aproximar U' por su valor esperado, $-\mathcal{J}(\beta)$, para obtener

$$l(\beta) \approx l(\hat{\beta})+(\beta-\hat{\beta})U(\hat{\beta})-\frac{1}{2}(\beta-\hat{\beta})^2 \mathcal{J}(\beta),$$

donde $\mathcal{J}(\beta)$ es la información evaluada en β.

Cuando $\boldsymbol{\beta}$ es multi-dimensional, tenemos

$$l(\boldsymbol{\beta}) \approx l(\hat{\boldsymbol{\beta}})+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{U}(\hat{\boldsymbol{\beta}})-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathcal{J}(\boldsymbol{\beta})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}), \quad (1)$$

donde $\mathbf{U}$ es el vector de primeras derivadas

de la log-verosimilitud, y $\mathcal{J}$ es la matriz de información.

De forma similar, la primera aproximación de U es

$$\begin{aligned} U(\beta) &\approx U(\hat{\beta}) + (\beta - \hat{\beta})U'(\hat{\beta}) \\ &\approx U(\hat{\beta}) - (\beta - \hat{\beta})\mathcal{J}(\beta) \end{aligned} \qquad (2)$$

en el caso donde β es 1-dimensional, y

$$\mathbf{U}(\boldsymbol{\beta}) \approx \mathbf{U}(\hat{\boldsymbol{\beta}}) - (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathcal{J}(\boldsymbol{\beta})$$

cuando $\boldsymbol{\beta}$ es multi-dimensional.

Distribución aproximada de $\hat{\boldsymbol{\beta}}$

En el caso 1-dimensional, de (2), si $\hat{\beta}$ es el EMV, por definición, tenemos

$$U(\beta) \approx -(\beta - \hat{\beta})'\mathcal{J}(\beta).$$

Por lo tanto,

$$(\hat{\beta} - \beta)\sqrt{\mathcal{J}(\beta)} \approx -\frac{U(\beta)}{\sqrt{\mathcal{J}(\beta)}}.$$

Recuerde que $U(\beta)$ es una suma de variables independientes (dado que las Y_i's son independentes), y que $\mathrm{E}[U(\beta)] = 0$. Ademas,

$\text{Var}[U(\beta)] = \mathcal{J}(\beta)$. Por lo tanto, podemos usar el teorema central del límite para mostrar que, asintóticamente,

$$\frac{U(\beta)}{\sqrt{\mathcal{J}(\beta)}} \sim \text{N}(0,1).$$

Asi, asintóticamente,

$$(\hat{\beta} - \beta)\sqrt{\mathcal{J}(\beta)} \sim \text{N}(0,1).$$

La cantidad $1/\mathcal{J}(\beta)$ es llamada *varianza asintótica* de $\hat{\beta}$. La varianza asintótica es desconocida dado que envuelve parámetros no conocidos β. Sin embargo, para n grande, $\mathcal{J}(\hat{\beta}) \approx \mathcal{J}(\beta)$. Se puede mostrar que, asintóticamente,

$$(\hat{\beta} - \beta)\sqrt{\mathcal{J}(\hat{\beta})} \sim \text{N}(0,1).$$

La estadística

$$(\hat{\beta} - \beta)\sqrt{\mathcal{J}(\hat{\beta})} \tag{3}$$

es llamada *estadística de Wald*.

Cuando $\boldsymbol{\beta}$ es multi-dimensional, $\hat{\boldsymbol{\beta}}$ es asintóticamente normal multivariada con media $\boldsymbol{\beta}$ y matriz de covarianza asintótica $\mathcal{J}^{-1}(\boldsymbol{\beta})$. La versión multidimensional de la estadística de Wald es

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathcal{J}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{4}$$

la cual es asintóticamente distribuida como $\chi^2(p)$.

La estadística de Wald y sus propiedades asintóticas pueden ser usadas para calcular intervalos de confianza aproximados para (o realizar pruebas de hipótesis en base a ellos) $\boldsymbol{\beta}$. Por ejemplo, un intervalo de confianza al 95% para β_j es

$$\hat{\beta}_j \pm z_{0.975}\sqrt{\hat{\text{Var}}[\hat{\beta}_j]},$$

donde $\hat{\text{Var}}[\hat{\beta}_j]$ es el $(j,j)^{th}$ elemento de $\mathcal{J}^{-1}(\hat{\beta})$ y

$$\text{P}(Z \leq z_\alpha) = \alpha,$$

donde $Z \sim$N$(0,1)$.

R da las estimaciones de $\hat{\boldsymbol{\beta}}$ y sus errores estándares en las salidas!

NOTA: Si el modelo es lineal y la variable respuesta está normalmente distribuida, los resultados de (3) y (4) son exactos.

Ejemplo: (cont.)

Recuerde el ejemplo de fraudes bancarios, donde

$$Y_i = \begin{cases} 1, \text{ hay fraude} \\ 0, \text{ caso contrario} \end{cases},$$

y π_i es la probabilidad de que el i^{th} cliente realice un fraude. Asumimos que las Y_i's son v.a.s independientes con distribución Bernoulli. Proponemos el modelo

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

para describir la relación entre la probabilidad de fraude y el balance.

Ahora realizamos una prueba de significancia de los coeficientes de regresión, en particular de la variable explicativa balance.

R estima el coeficiente asociado con el balance es 5.499×10^{-3}, con error estándar 2.204×10^{-4}. Además provee un"z-value" de 24.95, lo cual nos lleva a un p-valor (2-lados) de

$$\text{p-value} = 2\text{P}(Z > 24.95) = 2\text{P}(1-(Z < 24.95)),$$

$$\text{p-value} = 2\text{P}(1-1) = 0,$$

donde Z es una N(0,1). Por lo tanto, hay evidencia de que el balance tiene un efecto en la probabilidad de fraude.

Bondad de ajuste - Devianza

La devianza es importante para MLG ajustados. Se puede usar para probar el ajuste de la función de enlace y de las varibles explicativas globalmente a los datos, o para probar la significancia de una o más variables explicativas en el modelo.

Las siguientes secciones se aplica a MLG (con $\phi = 1$), es decir, Poisson and binomial.

PRUEBA DE BONDAD DE AJUSTE DEL MODELO

Definición de *modelo saturado*:

Sea K el máximo número de parámetros que podemos estimar a partir de los datos, $K \leq n$. En los MLG definimos un modelo saturado como el MLG con la misma distribución que y función de enlace del modelo de interés, pero con $g(\mu_i) = \eta_i^\star$ para $i = 1, \ldots, K$. En otras palabras, el modelo saturado nos permite una respuesta media diferente ($\mu_i = g^{-1}(\eta_i^\star)$), y por lo tanto tiene K parametros para ser

estimados. Vamos a denotar este vector de parámetros por $\boldsymbol{\psi}$.

Podemos pensar en el modelo saturado como el que tiene la estructura de media más general posible para los datos dado que las medias μ_i son *no restringidas*, **podemos asumir que el modelo saturado tiene n parámetros diferentes, uno para cada observación.** El modelo saturado también es llamado *modelo completo* o *modelo maximal*.

Probar la bondad de ajuste de la función de enlace o predictor lineal del MLG a los datos. Una forma de evaluar esto es comparar *el modelo ajustado con el modelo saturado*.

Sean $\mathcal{L}_S(\boldsymbol{\psi}; \mathbf{y})$ y $\mathcal{L}(\boldsymbol{\beta}; \mathbf{y})$ las verosimilitudes correspondientes al modelo saturado y el modelo propuesto, respectivamente.

Sabemos que $\mathcal{L}_S(\boldsymbol{\psi}; \mathbf{y}) \geq \mathcal{L}(\boldsymbol{\beta}; \mathbf{y})$ dado que el modelo de interés es un caso especial del modelo saturado.

Podemos comparar $\mathcal{L}_S(\boldsymbol{\psi}; \mathbf{y})$ y $\mathcal{L}(\boldsymbol{\beta}; \mathbf{y})$ – o equivalentemente $l_S(\boldsymbol{\psi}) \equiv \log \mathcal{L}_S(\boldsymbol{\psi}; \mathbf{y})$ y $l(\boldsymbol{\beta}) \equiv \log \mathcal{L}(\boldsymbol{\beta}; \mathbf{y})$ – para evaluar qué tan bien la función de enlace asumida o el predictor lineal se ajusta a los datos.

En particular, esperamos que $l(\boldsymbol{\beta}) \approx l_S(\boldsymbol{\psi})$ si el modelo es adecuado.

Definimos la *devianza* o *estadística de razón de verosimilitud*, D, como

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

donde $\hat{\boldsymbol{\psi}}$ y $\hat{\boldsymbol{\beta}}$ son los EMVs del modelo saturado y del modelo propuesto, respectivamente.

Bajo ciertas condiciones, asintóticamente

$$D \sim \chi^2_{K-p},$$

donde K y p son el número de parámetros del modelo saturado y propuesto, respectivamente.

Pruebas de hipótesis:
H_0 : Modelo saturado mejor ajuste
H_1 : Modelo propuesto mejor ajuste

$$P(D > \text{valor crítico}) = \alpha,$$

donde

$$D \sim \chi^2_{K-p},$$

Si el modelo propuesto es "pobre", D será mucho mayor que el valor crítico.

Para la distribución Binomial, la devianza es definida por:

$$D = 2 \sum_{i=1}^{n} [y_i \log(y_i/n\hat{i}\mu_i)) + (n_i - y_i) \log(1 - y_i/n_i)/(1 - \hat{\mu}_i)].$$

R calcula la devianza (etiquetada como "residual deviance") en la salida del summary .
En nuestro ejemplo: D = 1596.5

También se puede compara el modelo nulo definido como el MLG propuesto (con la misma función de enlace), pero con $g(\mu_i) = \beta_0$ para algún parámetro β_0. En otras palabras, el modelo nulo asume que todas las observaciones tienen la *misma distribución* con un parámetro común β_0.

Podemos usar la devianza para comparar el modelo propuesto con el modelo nulo. Si $l_N(\hat{\beta}_0)$ es la función de verosimilitud del modelo nulo, entonces asintóticamente:

$$2[l(\hat{\boldsymbol{\beta}}) - l_N(\hat{\beta}_0)] \sim \chi^2_{p-1},$$

donde p y 1 son el número de parámetros en el modelo propuesto y nulo, repectivamente.

R calcula la "null deviance", que es

$$ND = 2[l_S(\hat{\boldsymbol{\Psi}})] - l_N(\hat{\beta}_0)].$$

En nuestro ejemplo: ND = 2920.6

Esta cantidad no es útil en sí, pero

$$ND - D = 2[l(\hat{\boldsymbol{\beta}}) - l_N(\hat{\beta}_0)],$$

es la cantidad que necesitamos para comparar el modelo propuesto al modelo nulo.

De forma similar, podemos usar *el cambio en la devianza*, como el cambio en la suma de los residuos al cuadrado, para probar la significancia de una o más variables explicativas.

En particular, sea el modelo reducido el mismo que el modelo propuesto, pero excluyendo la variable explicativa de interés.

Sea $\boldsymbol{\beta}^*$ el vector de parámetros asociados al modelo reducido, y sea $l_R(\hat{\boldsymbol{\beta}}^*)$ la log-verosimilitud del modelo reducido. Entonces asintóticamente,

$$2[l(\hat{\boldsymbol{\beta}}) - l_R(\hat{\boldsymbol{\beta}}^*)] \sim \chi^2_{p-q},$$

donde p y q son el número de parámetros en el modelo propuesto y reducido, respectivamente.

Prueba de hipótesis:
$H_0 : \beta_j = 0$ (Modelo reducido mejor ajuste)
$H_1 : \beta_j \neq 0$ Modelo propuesto mejor ajuste

El comando `anova` en R provee cambios en la devianza asociados a sacar parámetros secuencialmente del modelo en la columna etiquetada `Deviance`.