# MLG: Datos binarios

Equivalencia entre datos binarios y binomiales

Ejemplo: En un lote de $m$ partes, se observa que $Z_i$ son defectuosos. La colección de $n$ lotes son estudiados, y observaciones $Z_1, \ldots, Z_n$ son registrados. Las partes son independientes dentro y entre los lotes. Si $\pi_i = \mathrm{P}$ (una parte en lotes $i$ es defectuoso), luego la función de verosimilitud es

$$\begin{aligned} \mathcal{L}(\boldsymbol{\pi}; \mathbf{z}) &= \prod_{i=1}^{n} \binom{m}{z_i} \pi_i^{z_i} (1-\pi_i)^{m-z_i} \\ &= \prod_{i=1}^{n} \binom{m}{z_i} \cdot \prod_{i=1}^{n} \pi_i^{z_i} (1-\pi_i)^{m-z_i}. \end{aligned}$$

Note que la función de verosimilitud está definida solo proporcional a una constante ( donde "constante" significa que los parámetros son independientes). Por lo tanto, podemos escribir

$$\mathcal{L}(\boldsymbol{\pi}; \mathbf{z}) \propto \prod_{i=1}^{n} \pi_i^{z_i} (1-\pi_i)^{m-z_i}.$$

Ahora considere el proceso de registro alternativo donde tenemos

$$Y_{ij} = \begin{cases} 1, \ j^{th} \text{ parte de } i^{th} \text{ lote es defectuoso} \\ 0, \ \text{caso contrario} \end{cases}$$

La función de verosimilitud asociada con estos datos es luego

$$\begin{aligned} \mathcal{L}(\boldsymbol{\pi}; \mathbf{y}) &= \prod_{i=1}^{n} \prod_{j=1}^{m} \pi_i^{y_{ij}} (1 - \pi_i)^{1 - y_{ij}} \\ &= \prod_{i=1}^{n} \pi_i^{\Sigma_{j=1}^{m} y_{ij}} (1 - \pi_i)^{m - \Sigma_{j=1}^{m} y_{ij}} \\ &\propto \mathcal{L}(\boldsymbol{\pi}; \mathbf{z}). \end{aligned}$$

En otras palabras, tenemos la misma información acerca de $\boldsymbol{\pi}$ sin importar la forma de como se registren los datos. Si tenemos datos (es decir, aqui donde el tamaño del lote es $m$), podemos simplemente pensar en ellos como una colección de $m$ ensayos de Bernoulli. Por lo tanto desarrollaremos la teoria para este tipo de datos asumiendo que son Bernoulli.

Selección de la función de enlace

Escenario: $Y_1, \ldots, Y_n$ son $n$ v.a.s binarias independientes (tomando valores 0 o 1) con $\pi_i = \mathrm{P}(Y_i = 1)$. Modelamos

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

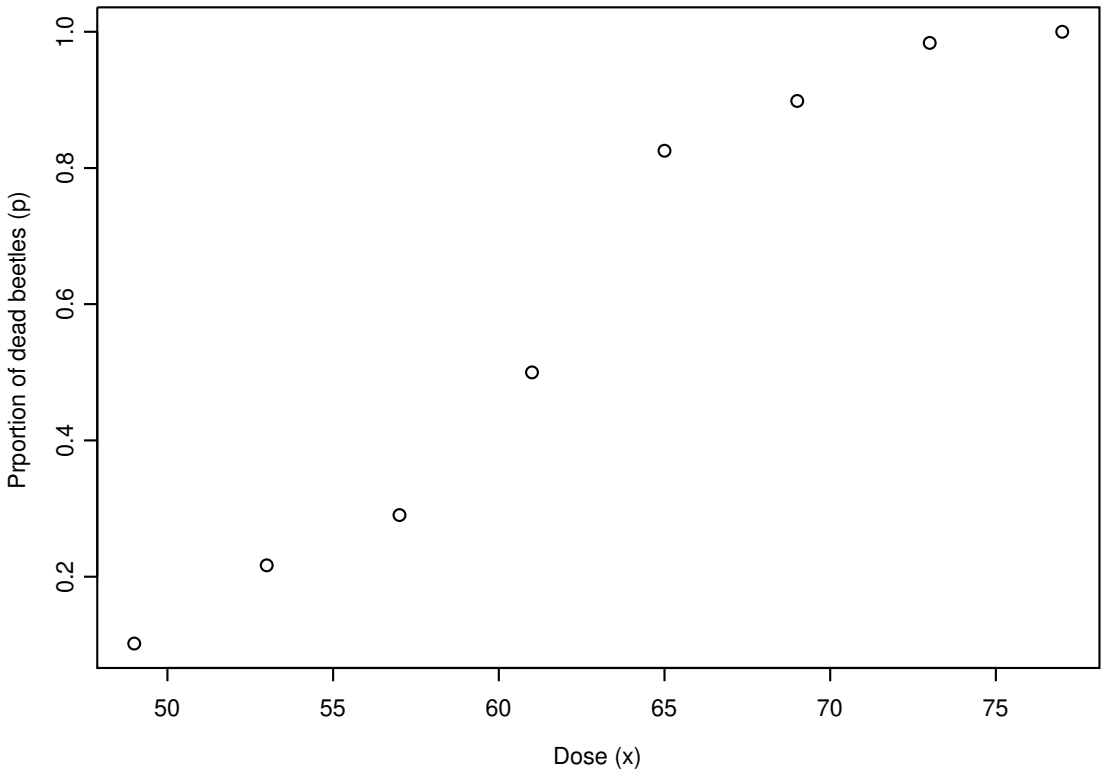
para alguna función de enlace $g$.

Datos de escarabajos

La tabla muestra el número $(y)$ de escarabajos muertos luego de cinco horas de exposición a disulfato de carbono gaseoso en varios niveles de concentración $(x)$ medidos en mg $\mathrm{l}^{-1}$, y el número de escarabajos expuestos $(n)$. Problema: Sugiera una función de enlace adecuada para modelar la relación entre $x$ e $y$.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x$ | 49 | 53 | 57 | 61 | 65 | 69 | 73 | 77 |
| $n$ | 59 | 60 | 62 | 56 | 63 | 59 | 62 | 60 |
| $y$ | 6 | 13 | 18 | 28 | 52 | 53 | 61 | 60 |

Figure 1: Proporción de muertes de escarabajos vs. concentración




La figura 1 nos ayuda a visualizar la relación entre $\pi_i$ y $x_i$.

De la figura 1, una forma razonable de describir la relación entre $\pi_i$ y $x_i$ es

$$\pi_i = F(\eta) = F(\beta_0 + \beta_1 x_i)$$

donde $F$ es la función de distribución (tal que t $0 \leq \pi_i \leq 1$).

Por ejemplo,

$$F(\eta) = \frac{e^{\eta}}{1 + e^{\eta}}$$

es la distribución logistica, generando la función de enlace *logit*, $\eta = \log\left(\frac{\pi}{1-\pi}\right)$. Por otro lado,

$$F(\eta) = \Phi(\eta)$$

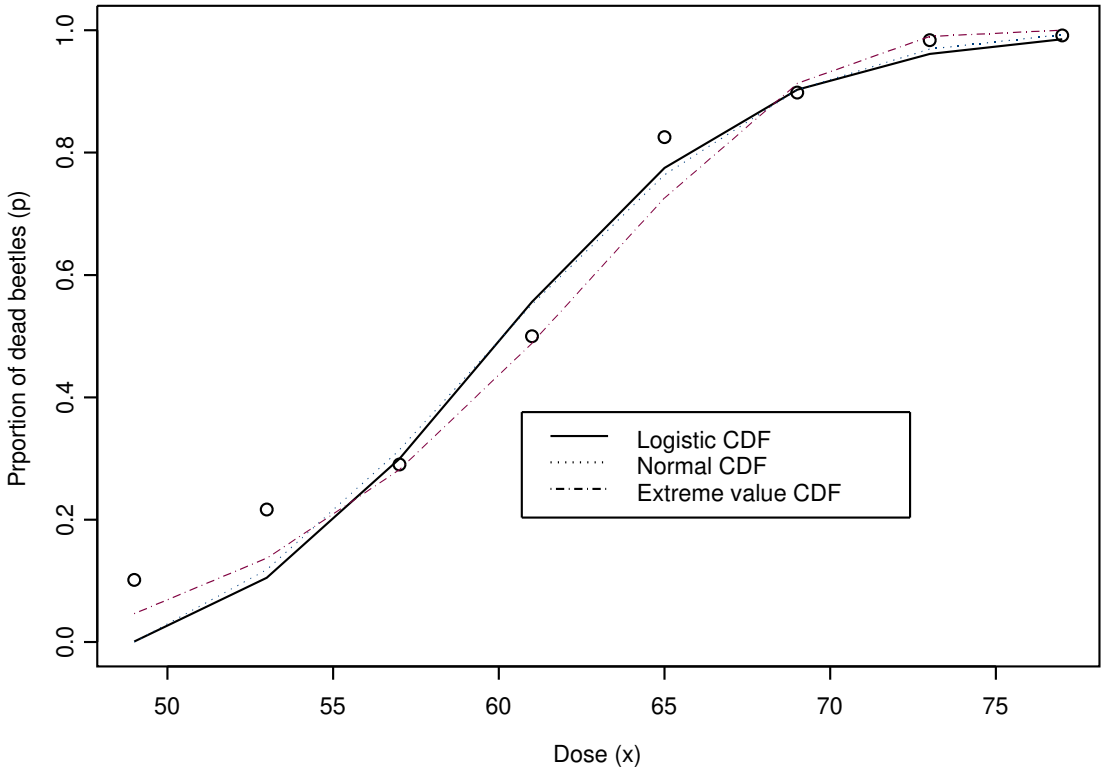
es la distribución $\mathrm{N}(0, 1)$, generando la función de enlace *probit*, $\eta = \Phi^{-1}(\pi)$. Finalmente,

$$F(\eta) = 1 - \exp[-\exp(\eta)]$$

es la distribución de valor extremo, generando la función de enlace *complementary log-log*, $\eta = \log[-\log(1 - \pi)]$.

La figura 2 muestra las formas de estas tres f.d.a,s (para valores seleccionados de valores de los parámetros ), superimpuestos en el diagrama de dispersion de los datos. Es claro que cualquier valor de las f.d.a.s puede proveer un buen ajuste a los datos si encontramos valores adecuados para los parámetros.

Figure 2: Three choices of link functions fit to the beetle data



La sintaxis para ajustas MLG binomiales en R depende de si la respuesta es en la forma binaria o binomial.

Si la variable respuesta, $Y$, es especificada como un vector de 1's y 0's, donde 1=éxito y 0=falla, entonces simplemente se especifica $Y$ como variable respuesta en R con el comando `glm()`.

Si, sin embargo, sus datos son especificados en la forma binomial (como en los datos de escarabajos), necesita proveer una *matriz* como respuesta en el comando `glm()`. En particular, se usa una matriz $n \times 2$ (donde $n$ es el número de observaciones binomiales), con la primera columna conteniendo el número de éxitosy la segunda el número de fallas. Las filas representan distintas combinaciones de variables explicativa(s) (en nuestro ejemplo dosis).

Ajustando estos tres modelos encontramos:

1. El modelo complementary log-log tiene la menor devianza (y por ello es el mejor ajuste en este sentido).
2. El modelo complementary log-log es mejor que los otros en términos de proximidad de los valores ajustados a los datos observados.

**NOTA:** Nosotros no usamos las proporciones de escarabajos muertos como nuestra variable respuesta! Usamos las proporciones solo para nuestro gráfico de diagnóstico. Nuestra respuesta sigue siendo el número de escarabajos en cada grupo.

Si seleccionamos el modelo complementary log-log, el modelo ajustado es entonces

$$\ln\left(-\ln(1-\hat{\pi}_i)\right) = -9.603 + 0.1525x_i,$$