El modelo de regresión log-binomial: una alternativa al modelo de regresión logística en estudios de cohortes y transversales.

Trabajo Final de Grado

Laura Julià Melis

Director: Klaus Langohr

Julio 2019

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

1. Introducción

Motivación

- El modelo de regresión logística es el más utilizado en estudios epidemiológicos.
- El enlace canónico para el caso binomial es el enlace "logit" o "log-odds" y utilitza el *odds ratio* (OR) como medida de asociación.

Inconvenientes

- OR sobreestima el riesgo cuando la variable de interés es frecuente.
- Su interpretación a menudo es difícil o poco intuitiva.

Objetivo

- Presentar el modelo log-binomial y diversas técnicas para solucionar los problemas de convergencia.

Riesgo relativo

$$RR = \frac{P(D|E)}{P(D|\bar{E})}.$$

Odds ratio

$$OR = \frac{odds(D|E)}{odds(D|\bar{E})} = \frac{P(D|E)/(1-P(D|E))}{P(D|\bar{E})/(1-P(D|\bar{E}))}$$

Relación RR-OR $\longrightarrow$ $$\boxed{OR = \frac{P(D|E)/(1-P(D|E))}{P(D|\bar{E})/(1-P(D|\bar{E}))} = RR \cdot \frac{1-P(D|\bar{E})}{1-P(D|E)}}$$

Sea Y la variable respuesta de interés:

$$Y = \begin{cases} 1 & \text{Presencia de enfermedad (D)} \\ 0 & \text{Ausencia de enfermedad } (\bar{D}) \end{cases}$$

La probabilidad de éxito condiconada al valor de las predictoras es:

$$\pi_{\boldsymbol{X}} = P(Y = 1|\boldsymbol{X}) \quad \text{con} \quad Y \sim B(1, \pi) \quad \text{sujeto a} \quad \pi_{\boldsymbol{X}} \in [0, 1].$$

Pero $\eta = \alpha + \boldsymbol{\beta'X}$ tiene rango $\mathbb{R} \rightarrow$ función de enlace "logit".

Expresión del modelo

$$\text{logit}(\pi_{\boldsymbol{X}}) = \log\left(\frac{\pi_{\boldsymbol{X}}}{1-\pi_{\boldsymbol{X}}}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

O equivalentemente,

$$\pi_{\boldsymbol{X}} = g^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)} = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}{1+\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}$$

- Estimación de los parámetros: criterio de **máxima verosimilitud**.

Función de máxima verosimilitud

$$L(\alpha, \boldsymbol{\beta}|Y, \boldsymbol{X}) = \prod_{i=1}^{n} P(Y = y_i|\boldsymbol{x_i}) f(\boldsymbol{x_i}) \propto \prod_{i=1}^{n} P(Y = y_i|\boldsymbol{x_i})$$

$$= \prod_{i=1}^{n} P(Y = 1|\boldsymbol{x_i})^{\delta_i} P(Y = 0|\boldsymbol{x_i})^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \frac{\exp(\alpha + \boldsymbol{\beta' x_i})^{\delta_i}}{1 + \exp(\alpha + \boldsymbol{\beta' x_i})},$$

- Interpretación a partir del OR, siendo X_i una variable dicotómica:

$$OR_{X_i} = \frac{odds(Y = 1|X_1, ..., X_i = 1, ..., X_k)}{odds(Y = 1|X_1, ..., X_i = 0, ..., X_k)} = \exp(\beta_i)$$

2. Métodos estadísticos (IV)

Regresión log-binomial (i)

La función de enlace que utiliza es el **logaritmo**.

Siendo:

- $\boldsymbol{X} = (X_1, ..., X_k)'$ el conjunto de variables explicativas,
- α el término independiente,
- $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$ los parámetros del modelo y
- $\pi_{\boldsymbol{x}} = P(Y = 1|\boldsymbol{X})$ la probabilidad de éxito

Se define el modelo log-binomial como:

$$\eta = g(\pi_{\boldsymbol{x}}) = \log(\pi_{\boldsymbol{x}}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad (1)$$

Probabilidad de respuesta positiva modelada a partir de (1):

$$\pi_{\boldsymbol{x}} = g^{-1}(\eta) = \exp(\eta) = \exp(\alpha + \boldsymbol{\beta'X}) = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

2. Métodos estadísticos (V)

Regresión log-binomial (ii)

- **Interpretación de parámetros**
  - En estudios de cohortes prospectivos: riesgo relativo (RR).
  - En estudios transversales: razón de prevalencias (PR).

RR/PR asociado a $X_i = 1$ (X_i dicotómica) y ajustado para el resto de covariables

$$RR_{X_i}(\text{o } PR_{X_i}) = \frac{P(Y=1|X_1,...,X_i=1,...,X_k)}{P(Y=1|X_1,...,X_i=0,...,X_k)}$$

$$= \frac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_iX_i+\cdots+\beta_kX_k)}{\exp(\beta_0+\beta_1X_1+\cdots+\beta_kX_k)} = \exp(\beta_i)$$

- **Posibles problemas de estimación**
  - Rango de valores diferentes $\rightarrow \pi_{\boldsymbol{x}} \in [0,1]$ y $\exp(\boldsymbol{\beta'X}) > 0$.
  - Problemas de convergencia al maximizar la función de verosimilitud.
  - Imposibilidad de obtener la estimación de los parámetros del modelo.

1. **Hipótesis:**

$$H_0 : \text{el modelo se ajusta bien a los datos.}$$
$$H_1 : \text{el modelo NO se ajusta bien a los datos.}$$

2. **Estadístico:**

Test de Hosmer-Lemeshow

$$\chi^2_{HL} = \sum_{k=1}^{g} \frac{(O_k - E_k)^2}{E_k} \sim \chi^2_{g-2}$$

Test basado en el estadístico de la devianza

$$D = 2\sum_{j=1}^{J} \left\{ y_j \log\left(\frac{y_j}{\hat{y}_j}\right) + (n_j - y_j) \log\left(\frac{n_j - y_j}{n_j - \hat{y}_j}\right) \right\} \sim \chi^2_{n-p}$$

3. **P-valor y conclusión:**

Si $P(\chi^2_{g-2} > \chi^2_{HL})$ o $P(\chi^2_{n-p} > D)$ inferiores a $\alpha \rightarrow$ rechazar H_0.

2. Métodos estadísticos (VII)

Software

1. Función *"glm"*:

```
> glm(formula = response~terms, family = binomial (link ="log" / "logit"), data, start = NULL, ... )
```

2. Función *"logbin"* (Donoghoe, 2018):

```
> logbin(formula, data, start = NULL, method = c("cem","em","glm","glm2","ab"), warn = TRUE, ...)
```

3. Función *"COPY"* (Deddens, 2002):

```
>  copy <- function(data, Y, vars,n, W) {
+  if (!is.numeric(data[, Y])) {
      data[, Y] <- as.numeric(data[, Y])-1
+  }
+  data$W <- (n-1)/n
+  data.copy <- data
+  data.copy[, Y] <- 1-data.copy[, Y]
+  data.copy$W <- 1/n
+  data.all <- merge(data, data.copy, all = T)
+  formul <- paste(Y, paste(vars, collapse = " + "), sep = "~")
+  mod.mat <- model.matrix(as.formula(formul), data)
+  glm.copy <- glm(as.formula(formul), family = binomial(log), data.all, weights = W, control =
              list(maxit = 100), start = c(-4, rep(0, ncol(mod.mat)-1))
+  return(glm.copy)
+  }
```

3. Estudio sobre la depresión postparto

Metodología

"Is Neuroticism a Risk Factor for Postpartum Depression?"(Martín-Santos y col., 2012)

- 1804 mujeres libres de depresión.
- Evaluaciones mediante cuestionarios.
- Tres variables respuesta diferentes.
- Ajuste de tres modelos de regresión logística siguiendo el método de Hosmer y Lemeshow (2000).
- Interpretación de parámetros en términos del OR ajustado.
- Bondad de ajuste utilizando el test propuesto por Le Cessie y van Houwelingen (1991), implementado en la función *"residuals.lrm"*.

4. Aplicación de los modelos de regresión (I)

Descripción de la base de datos

- 1804 filas y 30 columnas.
- Categorización de variables contínuas.
- Variables explicativas:
  - `epqnT`: Puntuación de neuroticismo.
  - `epds0`: Puntuación basal del cuestionario EPDS.
  - `duke`: Puntuación en el cuestionario sobre el apoyo social.
  - `antpers`: Historial clínico psiquiátrico personal.
- Resumen de las variables explicativas numéricas.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Missings
epqnT	32.0	37.0	43.0	43.6	47.0	73.0	7
epds0	0.0	3.0	6.0	6.2	9.0	27.0	0
duke	5.0	49.0	54.0	52.2	58.0	62	18

Síntomas de depresión a las 8 semanas postparto (EPDS>9)

Variable	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor	adj. OR(IC 95 %)
Constante	-4.305	0.731	-5.888	<0.0001	
Neuroticismo	0.047	0.011	4.119	<0.0001	1.05 (1.02,1.07)
Puntuación EPDS	0.177	0.022	7.903	<0.0001	1.19 (1.14,1.25)
Apoyo social	-0.020	0.009	-2.291	0.022	0.98 (0.96,1)
Historia psiquiátrica personal	0.668	0.190	3.515	0.0004	1.95 (1.34,2.83)

Expresión matemática del modelo

$$\text{logit}(\hat{\pi}_{\boldsymbol{x}}) = \log\left(\frac{\hat{\pi}_{\boldsymbol{x}}}{1-\hat{\pi}_{\boldsymbol{x}}}\right) = -4{,}305 + 0{,}047X_1 + 0{,}177X_2 - 0{,}02X_3 + 0{,}668X_4,$$

$$\hat{\pi}_{\boldsymbol{x}} = \frac{\exp(-4{,}305 + 0{,}047X_1 + 0{,}177X_2 - 0{,}02X_3 + 0{,}668X_4)}{1 + \exp(-4{,}305 + 0{,}047X_1 + 0{,}177X_2 - 0{,}02X_3 + 0{,}668X_4)}$$

4. Aplicación de los modelos de regresión (III)

Modelo de regresión log-binomial (i)

Función "glm":

Variable	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor	$\widehat{RR}$
Constante	-4.835	0.334	-14.459	< 0.0001	
Neuroticismo	0.042	0.006	6.753	< 0.0001	1.043
Puntuación EPDS	0.068	0.007	10.157	< 0.0001	1.070
Apoyo social	0.007	0.001	8.139	< 0.0001	1.007
Historia psiquiátrica personal	0.219	0.138	1.581	0.114	1.245

Expresión matemática del modelo

$$\log(\hat{\pi}_{\boldsymbol{x}}) = -4{,}835 + 0{,}042X_1 + 0{,}068X_2 + 0{,}007X_3 + 0{,}219X_4$$

O alternativamente, en función de $\hat{\pi}_{\boldsymbol{x}}$:

$$\hat{\pi}_{\boldsymbol{x}} = \exp(-4{,}835 + 0{,}042X_1 + 0{,}068X_2 + 0{,}007X_3 + 0{,}219X_4)$$

4. Aplicación de los modelos de regresión (IV)

Modelo de regresión log-binomial (ii)

Función "logbin":

```
Warning message:
nplbin: fitted probabilities numerically 1 occurred
```

Variable	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor	$\widehat{RR}$
Constante	-3.905	NA	NA	NA	
Neuroticismo	3.62e-02	NA	NA	NA	1.037
Puntuación EPDS	3.72e-02	NA	NA	NA	1.038
Apoyo social	7.77e-16	NA	NA	NA	1.000
Historia psiquiátrica personal	2.55e-01	NA	NA	NA	1.291

- Coste computacional muy alto.
- Varias pruebas fijando como tolerancia interior (`bound.tol`) diferentes valores entre 0 y 1 pero no obtención de todos los valores del modelo.

4. Aplicación de los modelos de regresión (V)

Modelo de regresión log-binomial (iii)

Función "COPY":

- Modelo con $n = 10000$ copias.

Variable	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor	$\widehat{RR}$
Constante	-4.734	0.299	-15.815	< 0.0001	
Neuroticismo	0.043	0.005	7.847	< 0.0001	1.044
Puntuación EPDS	0.064	0.006	10.597	< 0.0001	1.066
Apoyo social	0.007	0.001	9.891	< 0.0001	1.007
Historia psiquiátrica personal	0.155	0.1267108	1.226	0.22	1.168

Expresión matemática del modelo

$$\hat{\pi}_{\boldsymbol{x}} = \exp(-4{,}734 + 0{,}043X_1 + 0{,}064X_2 + 0{,}007X_3 + 0{,}155X_4)$$

4. Aplicación de los modelos de regresión (VI)

Pruebas de bondad de ajuste

Prueba de *Hosmer-Lemeshow* (Hosmer & Lemeshow, 2000), con la función `hoslem.test` del paquete *ResourceSelection*.

Modelo	Función de R	Estadístico χ^2	p-valor
Síntomas depresivos a las 8 semanas	logística	10.092	0.259
	log-binomial "glm"	35.526	2.14e-05
	log-binomial "logbin"	55.778	3.12e-09
Síntomas depresivos a las 32 semanas	logística	8.367	0.398
	log-binomial "glm"	8.796	0.360
	log-binomial "logbin"	8.796	0.360
Diagnóstico de depresión mayor	logística	11.442	0.178
	log-binomial "glm"	12.238	0.141
	log-binomial "logbin"	12.487	0.131

5. Discusión y conclusiones (I)

Comparación de resultados

1. Modelo 1:
   - $\widehat{OR}$ y $\widehat{RR}$ llevan a diferentes conclusiones.
   - Cambios en los signos y la significación de parámetros.
   - El modelo de regresión log-binomial no se ajusta bien a los datos.

2. Modelos 2 y 3:
   - $\widehat{OR}$ y $\widehat{RR}$ apuntan en la misma dirección.
   - Parámetros estimados significativos en ambos modelos y signos iguales.
   - Test de *Hosmer-Lemeshow*: ajustes de ambas regresiones son buenos.

Conclusiones

- La bondad del ajuste es de vital importancia para decidir el tipo de regresión a utilizar.
- Se prefiere el modelo log-binomial por la interpretación del riesgo relativo, pero únicamente cuando el ajuste es bueno.

5. Discusión y conclusiones (II)

Consideraciones metodológicas

1. Función *"COPY"*:
   - Resultados con 1000 y 10000 copias muy similares.
   - Coste computacional en ambos casos prácticamente igual.
   - Preferible el ajuste con más copias aunque la ganancia sea mínima.

2. Limitaciones del test de *Hosmer-Lemeshow*:
   - El valor de χ^2_{HL} depende de los puntos de corte que definen los grupos.
   - Poca potencia para detectar falta de ajuste.
   - Alternativa: Test propuesto por Le Cessie y van Houwelingen (1991), función *"`residuals.lrm`"* de R.

Referencias

Deddens, J. A. (2002). Estimation of prevalence ratios when PROC GENMOD does not converge.

Donoghoe, M. W. (2018). *logbin: Relative Risk Regression Using the Log-Binomial Model*. R package version 2.0.4. Recuperado desde https://CRAN.R-project.org/package=logbin

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression* (Second). John Wiley & Sons.

Le Cessie, S. & van Houwelingen, J. C. (1991). A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods. *Biometrics*, *47*(4), 1267-1282. Recuperado desde http://www.jstor.org/stable/2532385

Martín-Santos, R., Gelabert, E., Subirá, S., Gutierrez-Zotes, A., Langohr, K., Jover, M., ... Sanjuan, J. (2012). Research Letter: Is neuroticism a risk factor for postpartum depression? *Psychological medicine*, *42*, 1559-65. doi:10.1017/S0033291712000712

MUCHAS GRACIAS POR SU ATENCIÓN