

# Reduciendo el error Monte Carlo en la estimación bayesiana de los riesgos relativos usando modelos de regresión binomial

Salmerón D. y Cano J.A.

Servicio de Epidemiología, Consejería de Sanidad de la Región de Murcia, IMIB-Arrixaca, CIBER Epidemiología y Salud Pública, y Universidad de Murcia

UNIVERSIDAD DE  
MURCIA



ciberesp  
Red de Centros de Investigación en  
Epidemiología y Salud Pública



- Riesgo relativo
  - regresión logística
  - regresión log-binomial
  
- Estimación del modelo de regresión log-binomial
  - Problemas en estimación frecuentista
  - Problemas en estimación bayesiana
    - Funciones en R

# Riesgo relativo y regresión logística

## Riesgo relativo y regresión logística

El modelo de regresión logística se utiliza con mucha frecuencia en estudios Epidemiológicos para analizar una respuesta  $Y \in \{0, 1\}$

$$P(Y = 1 \mid x, \beta) = \frac{\exp(x\beta)}{1 + \exp(x\beta)}$$

$$\text{logit } P(Y = 1 \mid x, \beta) = x\beta$$

$x = (x_1, \dots, x_k)$  es el vector de variables independientes,  $x_k = 1$ ,

$\beta = (\beta_1, \dots, \beta_k)^T$  es el vector de parámetros desconocidos,

$x\beta = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$ .

## Riesgo relativo y regresión logística

$Y = 1$  indica presencia de una enfermedad.

Dos tratamientos que deseamos comparar.

El vector  $x$  incluye el tipo de tratamiento ( $T = 0$  o  $T = 1$ ) y otras variables de ajuste  $Z = (Z_1, \dots, Z_{k-2})^T$ .

El modelo de regresión logística establece que

$$P(Y = 1 | T, Z, \beta) = \frac{\exp(a + bT + cZ)}{1 + \exp(a + bT + cZ)}$$

$$\beta = (a, b, c), \quad c = (c_1, \dots, c_{k-2})$$

## Riesgo relativo y regresión logística

Al comparar los tratamientos, el riesgo relativo de desarrollar la enfermedad

$$\frac{P(Y = 1 \mid T = 1, Z, \beta)}{P(Y = 1 \mid T = 0, Z, \beta)} = \frac{1 + \exp(a + cZ)}{1 + \exp(a + b + cZ)} e^b$$

toma un valor diferente para cada posible configuración de las variables incluidas en  $Z$ .

Dependiendo de la dimensión de  $Z$ , no resulta viable presentar estimaciones de riesgo relativo usando este modelo pues el número de configuraciones puede ser muy alto.

# Riesgo relativo y Modelo log-binomial

# Modelo log-binomial

Cambiamos logit, por la función log

$$\log P(Y = 1 \mid x, \beta) = x\beta$$

En R usaríamos

```
glm(y ~ x, family = binomial(link = "log"))
```



## Modelo log-binomial

Si comparamos los tratamientos usando el modelo log-binomial:

$$\log P(Y = 1 \mid T, Z, \beta) = a + bT + cZ,$$

el riesgo relativo de desarrollar la enfermedad es

$$\frac{P(Y = 1 \mid T = 1, Z, \beta)}{P(Y = 1 \mid T = 0, Z, \beta)} = e^b,$$

independientemente del valor  $Z$ .

## Modelo log-binomial

$$\text{Riesgo Relativo} = \begin{cases} \frac{1+\exp(a+cZ)}{1+\exp(a+b+cZ)} e^b & \text{con regresión logística} \\ e^b & \text{con regresión log-binomial} \end{cases}$$

Si queremos estimar Riesgos Relativos, es preferible usar el modelo de regresión log-binomial, ver McNutt *et al.* (2003), Deddens *et al.* (2003), Greenland (2004), Spiegelman y Hertzmark (2005), Petersen y Deddens (2006), y Deddens y Petersen (2008).

# Estimación del Modelo log-binomial

# Estimación del Modelo log-binomial

Los datos  $y_1, \dots, y_n$  siguen una distribución de Bernoulli

$$y_i \sim \text{Ber}(p_i), \quad \log p_i = x_i \beta, \quad i \in \{1, \dots, n\},$$

y el objetivo es hacer inferencia sobre  $\beta$ .

## Estimación del Modelo log-binomial

El siguiente ejemplo está tomado de Chu y Cole (2010)

```
> y=c(0, 0, 0, 0, 1, 0, 1, 1, 1, 1)
> x=1:10
> summary(glm(y~x,family=binomial(link="log")))
```

```
Error: no valid set of coefficients has
been found: please supply starting values
```

```
>
>
```

## Estimación del Modelo log-binomial

```
> summary(glm(y~x,family=binomial(link="log"),
start=c(-2,0)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-2.09513	0.97057	-2.159	0.0309	*
x	0.20951	0.09706	2.159	0.0309	*

---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

Number of Fisher Scoring iterations: 25

Hubo 29 avisos (use warnings() para verlos)

>

## Estimación del Modelo log-binomial

```
> warnings()
```

Warning messages:

```
1: step size truncated due to divergence
```

```
2: step size truncated: out of bounds
```

```
.
```

```
.
```

```
.
```

```
27: glm.fit: algorithm did not converge
```

```
28: glm.fit: algorithm stopped at boundary value
```

```
29: glm.fit: fitted probabilities numerically
```

```
0 or 1 occurred
```

**¿Donde está el problema?**

# Estimación del Modelo log-binomial

El espacio paramétrico está restringido

$$\begin{aligned} \exp(x_i\beta) &= P(y_i = 1 \mid x_i, \beta), \quad i = 1, \dots, n \\ &\Downarrow \\ x_i\beta &< 0, \quad i = 1, \dots, n, \end{aligned}$$

lo que dificulta el proceso de maximizar la función de verosimilitud para obtener una estimación de  $\beta$ .



## Estimación del Modelo log-binomial

- $\nabla L(\hat{\beta}) \neq 0$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.09513	0.97057	-2.159	0.0309 *
x	0.20951	0.09706	2.159	0.0309 *

- ¿ Podemos fiarnos de las estimaciones en la columna Estimate?
- ¿ Podemos fiarnos de Std. Error, z value y Pr(> |z|)?
- ¿ Vamos a poner en un paper que nuestras estimaciones tienen warnings?

# Estimación bayesiana del Modelo log-binomial

## Estimación bayesiana

- Chu y Cole (2010) han estudiado el problema de ajustar el modelo log-binomial desde el punto de vista bayesiano usando WinBUGS.

$$f(\mathbf{y} \mid \mathbf{X}, \beta) = \prod_{i=1}^n (\exp(x_i\beta)^{y_i} (1 - \exp(x_i\beta))^{1-y_i}) I(x_i\beta < 0)$$

- Concluyen que el método bayesiano produce estimaciones similares al estimador de máxima verosimilitud, y un menor error cuadrático medio.
- Sin embargo, la convergencia de WinBUGS puede ser muy lenta. Proponemos un método MCMC programado en R.

## Ejemplo de Chu y Cole (2010)

```
> y=c(0, 0, 0, 0, 1, 0, 1, 1, 1, 1)
> x=1:10
```

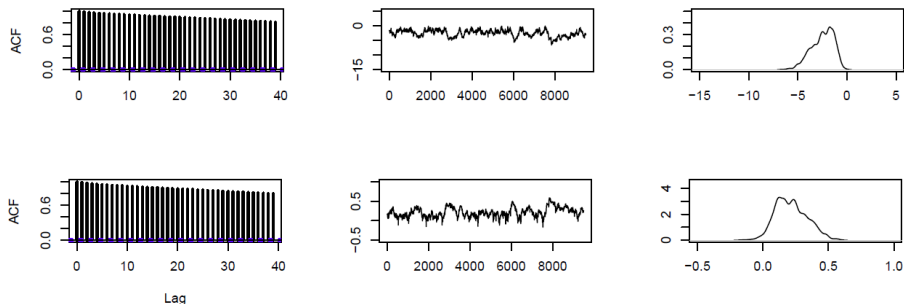


Figure: Resultado obtenido con WinBUGS usando la propuesta de Chu y Cole (2010), con  $\pi(\beta) = 1$ .

Pobre convergencia del algoritmo de Gibbs implementado en WinBUGS, debido en parte a:

- alta correlación a posteriori ( $\approx -0.97$ ) de las componentes de  $\beta$
- las restricciones  $x_i\beta < 0$  no son tenidas en cuenta de manera eficiente

Esto implica

- Gran error Monte Carlo
- Gran número de simulaciones para obtener buenas aproximaciones de la distribución a posteriori

## ¿ Qué hemos hecho?

- 1 Usar una reparametrización para reducir la correlación a posteriori
- 2 Usar un algoritmo de Gibbs que tenga en cuenta las restricciones  $x_i\beta < 0$  de manera eficiente
- 3 El método lo hemos implementado en R, ya que necesitamos
  - ajustar glm de Poisson,
  - operaciones con matrices, y descomposiciones de matrices,
  - simular variables y
  - el paquete coda para el chequeo de la convergencia de cadenas de Markov

# Reparametrización

# Reparametrización

$$\beta \longleftrightarrow \theta$$

$$\beta = L^T \theta, \quad \theta = (\theta_1, \dots, \theta_k)^T$$

Objetivo: Encontrar  $L$  no singular tal que las componentes de  $\theta$  sean aproximadamente independientes bajo la distribución a posteriori.



# Reparametrización

$$\beta = L^T \theta \Rightarrow \text{Cov}(\beta \mid \mathbf{y}) = L^T \text{Cov}(\theta \mid \mathbf{y}) L$$

- La factorización de Choleski de  $\text{Cov}(\beta \mid \mathbf{y})$  proporciona una candidata para  $L$ , pero necesitamos un buen método para aproximar  $\text{Cov}(\beta \mid \mathbf{y})$ .
- Si  $\pi(\beta) = 1$ , entonces  $\text{Cov}(\beta \mid \mathbf{y})$  puede ser aproximada por la matriz de covarianzas del estimador de máxima verosimilitud.

# Reparametrización

Pero calcular dicho estimador, y por tanto, dicha matriz, suele ser complicado

```
> warnings()
```

```
Warning messages:
```

```
1: step size truncated due to divergence
```

```
2: step size truncated: out of bounds
```

```
.
```

```
.
```

```
.
```

```
27: glm.fit: algorithm did not converge
```

```
28: glm.fit: algorithm stopped at boundary value
```

```
29: glm.fit: fitted probabilities numerically
```

```
0 or 1 occurred
```

# Reparametrización

- Poisson: Zou (2004) y Spiegelman y Hertzmark (2005) han sugerido usar un modelo de Poisson sin restricciones para aproximar la inferencia del modelo log-binomial.
- Nosotros proponemos usar dicho enfoque para aproximar  $Cov(\beta | \mathbf{y})$ .
- Si  $S \approx Cov(\beta | \mathbf{y})$ , entonces tomamos  $L$  tal que  $L^T L = S$  (Choleski) y por tanto

$$L^T L = S \approx Cov(\beta | \mathbf{y}) = L^T Cov(\theta | \mathbf{y}) L$$

$$Cov(\theta | \mathbf{y}) \approx \mathbf{I}$$

# Gibbs y las restricciones

## Gibbs y las restricciones

- Gibbs-WinBUGS actualiza cada componente de  $\beta$  simulando un valor que es aceptado con cierta probabilidad en un paso Metropolis-Hastings.
- Este proceso no es eficiente pues las restricciones  $x_i\beta < 0$  solo se consideran en el cálculo de la probabilidad de aceptar la simulación, y no antes.
- Sería mejor si se simulara el valor para que automáticamente todas las restricciones resultaran satisfechas

## Gibbs y las restricciones

- La distribución que el algoritmo de Gibbs tiene simular en cada paso:

$$\theta_j \mid \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_k$$

tiene soporte en un intervalo  $\Theta_j$  determinado por los datos, y por  $\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_k$ .

- Además, ya que  $\text{Cov}(\theta \mid \mathbf{y}) \approx \mathbf{I}$ , dicha distribución es aproximadamente la  $N(\hat{\theta}_j, 1)$ , restringida a  $\Theta_j$ , donde  $\hat{\theta}_j$  se obtiene de ajustar el modelo de Poisson.

## Gibbs y las restricciones

- Metropolis-Hastings: simulando una distribución de Cauchy restringida a  $\Theta_j$ .
- Hemos implementado nuestra propuesta en R y la hemos comparado con WinBUGS mediante algunos ejemplos

## Ejemplo. Greenland (2004).

Relación entre la supervivencia y el nivel de receptores, en una cohorte de 192 mujeres diagnosticadas con cáncer de mama. Las covariables son el nivel de receptores y el estadio.



## Ejemplo. Greenland (2004).

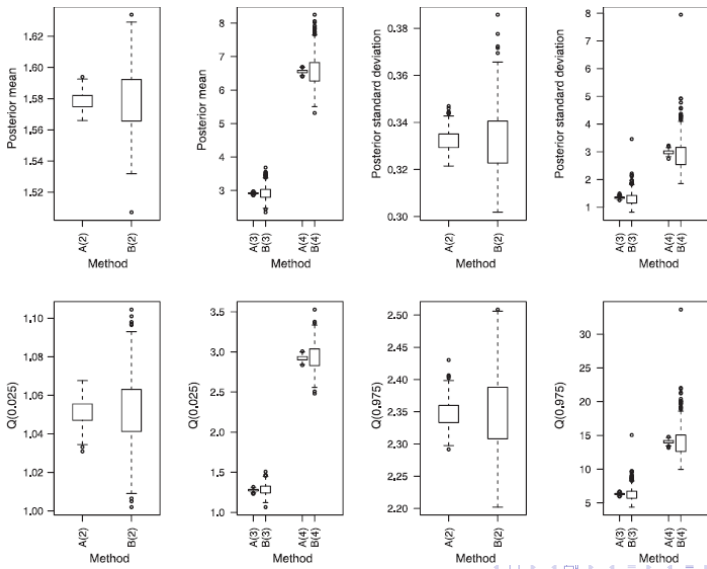
**Table III.** Effective sample size for the breast cancer mortality example.

	WinBUGS Mean (SD)	Proposed method Mean (SD)
$\exp(\beta_1)$	71.9 (13.3)	5192.1 (218.7)
$\exp(\beta_2)$	324.8 (30.6)	4047.4 (209.6)
$\exp(\beta_3)$	60.2 (17.8)	5166.7 (275.3)
$\exp(\beta_4)$	59.2 (16.5)	4486.8 (259.6)

Means and standard deviation (SD) of the 500 effective sample sizes obtained with WinBUGS and the proposed method, based on Markov chains of length 9500.

Figure: .

# Ejemplo. Greenland (2004).



## Ejemplo. Bajo peso al nacer.

Un estudio dirigido a identificar factores de riesgo asociados con el bajo peso al nacer.

189 mujeres embarazadas

Covariables:

irritabilidad del útero

fumar

raza

partos anteriores

edad

## Ejemplo. Bajo peso al nacer.

**Table V.** Effective sample size for the low birth weight example.

	WinBUGS Mean (SD)	Proposed method Mean (SD)
$\exp(\beta_1)$	83.7 (12.8)	4416.3 (237.7)
$\exp(\beta_2)$	432.6 (43.8)	4325.7 (241.6)
$\exp(\beta_3)$	240.1 (39.2)	3842.3 (194.3)
$\exp(\beta_4)$	309.8 (65.5)	4148.8 (215.4)
$\exp(\beta_5)$	245.2 (43.0)	4093.9 (231.0)
$\exp(\beta_6)$	264.0 (68.4)	4928.2 (283.3)
$\exp(\beta_7)$	191.1 (48.9)	4817.9 (286.0)
$\exp(\beta_8)$	431.2 (105.1)	5621.1 (315.3)
$\exp(\beta_9)$	874.0 (227.8)	5443.4 (275.0)
$\exp(\beta_{10})$	315.5 (32.3)	2438.6 (150.1)

Mean and standard deviation (SD) of the 500 effective sample sizes obtained with WinBUGS and the proposed method, based on Markov chains of length 9500.

## Conclusiones

- Si queremos presentar riesgos relativos, es preferible usar modelos de regresión log-binomial.
- Las restricciones dificultan el ajuste del modelo.
- El método bayesiano puede resolver este problema con WinBUGS, y produce estimaciones con un menor error cuadrático medio.

## Conclusiones

- Sin embargo, WinBUGS puede producir un error Monte Carlo grande
- Nosotros proponemos un método MCMC para reducir el error Monte Carlo
- Se basa en una reparametrización y en un paso Metropolis-Hastings con una distribución de Cauchy truncada.

**Gracias por vuestra atención**



McNutt LA, Wu C, Xue X, Haffner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology* 2003; **157**:940–943. DOI: 10.1093/aje/kwg074








Deddens JA, Petersen MR, Lei X. Estimation of prevalence ratios when PROC GENMOD does not converge. *Paper 270-28. Proceedings of the 28th Annual SAS Users Group International Conference, Seattle, Washington, March 30-April 2, 2003*. Downloaded from <http://www2.sas.com/proceedings/sugi28/270-28.pdf> on the 3rd March 2015.



Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology* 2004; **160**:301–305. DOI: 10.1093/aje/kwh221



-  Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology* 2005; **162**:199–200. DOI: 10.1093/aje/kwi188
-  Petersen MR, Deddens JA. RE: "Easy SAS calculations for risk or prevalence ratios and differences". *American Journal of Epidemiology* 2006; **163**:1158–1159. DOI: 10.1093/aje/kwj162
-  Deddens JA, Petersen MR. Approaches for estimating prevalence ratios. *Occupational and Environmental Medicine* 2008; **65**:501–506. DOI: 10.1136/oem.2007.034777
-  Zou GY. A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology* 2004; **159**:702–706.
-  Savu A, Liu Q, Yasui Y. Estimation of relative risk and prevalence ratio. *Statistics in Medicine* 2010; **29**:2269–2281. DOI: 10.1002/sim.3989

-  Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. *Epidemiology* 2010; **21**:855-862. DOI: 10.1097/EDE.0b013e3181f2012b
-  Spiegelhalter DJ, Thomas A, Best NG. WinBUGS User Manual, Version 1.4. Cambridge, United Kingdom: Medical Research Council Biostatistics Unit; 2003.
-  Hamra G, MacLehose R, Richardson D. Markov chain Monte Carlo: an introduction for epidemiologists. *International Journal of Epidemiology* 2013; **42**:627–634. DOI: 10.1093/ije/dyt043
-  R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
-  Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1993; **55**:3-23.

-  Minicozzi P, Caldarella A, Giacomini A, Ponz de Leon M, Cesaraccio R, Falcini F, Fusco M, Iachetta F, Pellegrini C, Tumino R, Capocaccia R, Sant M. Looking at differences in stage and treatment of colorectal cancers across Italy: a EURO CARE-5 high resolution study. *Tumori* 2012; **98**: 671-677. DOI: 10.1700/1217.13488
-  Allemani C, Rachet B, Weir HK, Richardson LC, Lepage C, Faivre J, Gatta G, Capocaccia R, Sant M, Baili P, Lombardo C, Aareleid T, Ardanaz E, Bielska-Lasota M, Bolick S, Cress R, Elferink M, Fulton JP, Galceran J, Gzdz S, Hakulinen T, Primic-Zakelj M, Rachtan J, Diba CS, Sanchez MJ, Schymura MJ, Shen T, Tagliabue G, Tumino R, Vercelli M, Wolf HJ, Wu XC, Coleman MP. Colorectal cancer survival in the USA and Europe: a CONCORD high-resolution study. *BMJ Open* 2013; **3**:e003055. DOI: 10.1136/bmjopen-2013-003055
-  Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* 2006; **6**:7-11.



Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. John Wiley and Sons: New York, 2000.