A score-test for testing zero-inflation in Poisson regression models under the identity-link

Jochen Einbeck and Maria Oliveira

in collaboration with
Liz Ainsbury and Kai Rothkamm (PHE)
Manuel Higueras and Pere Puig (UAB)

funded by the National Institute for Health Research
NIHR-RMOFS-2013-03-4

Dortmund, 18th March 2015

Durham University

Background: Radiation biodosimetry

- Radiation accident or incident leading to irradiated blood lymphocytes.
- Need rapid and reliable procedures to determine the radiation dose contracted by individuals.
- Members of the public do not usually wear radiation dosimeters...
- Hence, there is need for techniques which exploit the radiation-induced change in certain biomarkers to estimate the contracted radiation dose.

Background: Radiation biodosimetry

- Radiation accident or incident leading to irradiated blood lymphocytes.
- Need rapid and reliable procedures to determine the radiation dose contracted by individuals.
- Members of the public do not usually wear radiation dosimeters...
- Hence, there is need for techniques which exploit the radiation-induced change in certain biomarkers to estimate the contracted radiation dose.
- Most common: Cytogenetic biomarkers (counts of dicentric chromosome aberrations, micronuclei)

Example

- Frequency of dicentrics after whole body *in vitro* exposure to Co-60 gamma rays (Low LET; sparsely ionising radiation)

		y_{ij}					
x_i	n_i	0	1	2	3	4	5
0.00	2592	2591	1	0	0	0	0
0.25	2193	2185	8	0	0	0	0
0.75	2595	2550	44	1	0	0	0
1.00	2287	2231	54	2	0	0	0
1.50	1811	1712	96	3	0	0	0
2.50	1327	1196	123	7	1	0	0
3.00	1438	1070	320	41	6	1	0
4.50	1396	895	360	110	25	5	1

- x_i: dose (in Gy) used to irradiate blood sample i, $i = 1, \dots 8$.
- y_{ij}: counts of dicentric aberrations in j-th cell of blood sample i, $j = 1, \dots n_i$.

Poisson model

- Count data; that is Poisson model would be first choice:

$$y_{ij} \sim Po(\lambda_i).$$

- Model for mean function

$$g(\lambda_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2.$$

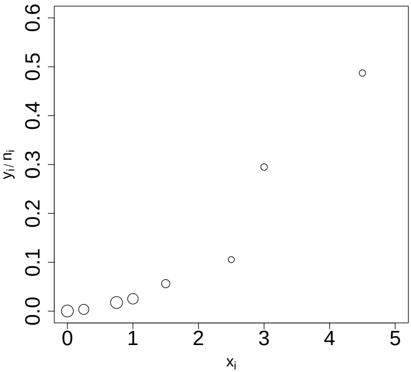
- The straightforward choice for $g(\cdot)$ would be the natural link

$$g(\cdot) = \log(\cdot)$$

- Why consider the identity–link instead?

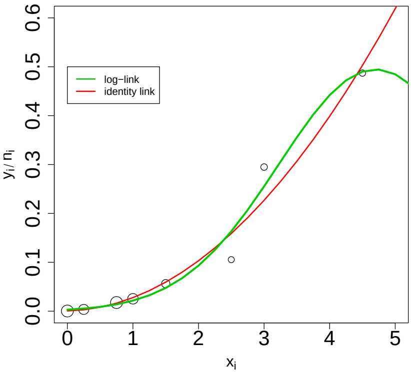
Dose–response curves

- Plot y_i/n_i versus x_i, with $y_i = \sum_j y_{ij}$, and circle sizes $\propto n_i$ [note that y_i and n_i form the sufficient statistics for the Poisson mean λ_i].



Dose–response curves

- Plot y_i/n_i versus x_i, with $y_i = \sum_j y_{ij}$, and circle sizes $\propto n_i$ [note that y_i and n_i form the sufficient statistics for the Poisson mean λ_i].
- Fitted model using log-link and identity-link:



- The log–link model behaves implausibly for higher doses, and is hence not acceptable by practitioners.

Zero–inflation

- Additional problem: There is strong overdispersion.
- The residual deviance of the Poisson (identity–link) model is 56.22 at $8 - 3 = 5$ degrees of freedom, and so the estimated dispersion is $\hat{\phi} = \frac{56.22}{5} = 11.24 \gg 1$.
- Here a plausible source of overdispersion is zero–inflation: Either a cell did not get irradiated (then 0 dicentrics), or it did (then Poisson dicentrics).

Zero–inflation

- Additional problem: There is strong overdispersion.
- The residual deviance of the Poisson (identity–link) model is 56.22 at $8-3=5$ degrees of freedom, and so the estimated dispersion is $\hat{\phi} = \frac{56.22}{5} = 11.24 \gg 1$.
- Here a plausible source of overdispersion is zero–inflation: Either a cell did not get irradiated (then 0 dicentrics), or it did (then Poisson dicentrics).
- Zero–inflated regression model

$$\mathrm{P}(Y_{ij} = y_{ij}) = \begin{cases} p_i + (1-p_i)\exp(-\lambda_i), & y_{ij} = 0, \\ (1-p_i)\exp(-\lambda_i)\lambda_i^{y_i}/y_i!, & y_{ij} > 0, \end{cases}$$

  where $0 \le p_i \le 1$ and $\lambda_i > 0$.
- We use $p_i \equiv p$ and $\lambda_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$.

Score–test

- Zero–inflation is difficult to detect reliably from the data itself.
- A reliable test is required in practice.
- van den Broek (1995, Biometrics **51**) developed a score (Rao) test for testing $H_0 = Po(\lambda_i)$, $H_1 = ZIP(p, \lambda_i)$,
- in other words, $H_0 : p = 0$.
- Score tests are attractive in this context as they do not require an estimation under the alternative!
- However, van den Broek's test is based on a model using the log–link. This would give incorrect results when using the identity link.

Score–test under the identity link

- Let $\lambda_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\theta = p/(1-p)$. Then $H_0 : \theta = 0$.
- Likelihood:

$$L(\theta, \boldsymbol{\beta}) = \frac{1}{(1+\theta)^n} \prod_{i=1}^{n} \left(1_{y_i=0}(\theta + e^{-\lambda_i}) + 1_{y_i \neq 0} e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!} \right).$$

- Note a difficulty: The fitted Poisson mean $\hat{\lambda}_i$ has to be positive so restricted optimization techniques have to be used.
- Score–test statistic

$$T = S(0, \hat{\boldsymbol{\beta}})^T J(0, \hat{\boldsymbol{\beta}})^{-1} S(0, \hat{\boldsymbol{\beta}}).$$

with the Score function S and Fisher information J, evaluated under $H_0 : \theta = 0$.

Score–test under the identity link

- Score–function. Let $\ell = \log L$. Then

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ I_{(y_i=0)} \left(\frac{-\exp(-\lambda_i)}{\theta + \exp(-\lambda_i)} \right) \boldsymbol{x}_i + I_{(y_i>0)} \left(\frac{y_i}{\lambda_i} - 1 \right) \boldsymbol{x}_i \right\}$$

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \left\{ \frac{-1}{1+\theta} + I_{(y_i>0)} \left(\frac{1}{\theta + \exp(-\lambda_i)} \right) \right\}$$

- That is, under $H_0 : \theta = 0$,

$$S(0, \boldsymbol{\beta}) = \left(\sum_{i} \left(\frac{I_{(y_i=0)}}{\exp(-\lambda_i)} - 1 \right), \sum_{i=1}^{n} \boldsymbol{x}_i \left(\frac{y_i}{\lambda_i} - 1 \right) \right)$$

Score–test under the identity link

- Score–function. Let $\ell = \log L$. Then

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ I_{(y_i=0)} \left(\frac{-\exp(-\lambda_i)}{\theta + \exp(-\lambda_i)} \right) \boldsymbol{x}_i + I_{(y_i>0)} \left(\frac{y_i}{\lambda_i} - 1 \right) \boldsymbol{x}_i \right\}$$

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \left\{ \frac{-1}{1+\theta} + I_{(y_i>0)} \left(\frac{1}{\theta + \exp(-\lambda_i)} \right) \right\}$$

- That is, under $H_0 : \theta = 0$,

$$S(0, \boldsymbol{\beta}) = \left(\sum_{i} \left(\frac{I_{(y_i=0)}}{\exp(-\lambda_i)} - 1 \right), \sum_{i=1}^{n} \boldsymbol{x}_i \left(\frac{y_i}{\lambda_i} - 1 \right) \right)$$

- The right hand part is the score vector for a Poisson GLM under identity link. So, under H_0, this term would be zero.

Score–test under the identity link

- Score–function. Let $\ell = \log L$. Then

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ I_{(y_i=0)} \left(\frac{-\exp(-\lambda_i)}{\theta + \exp(-\lambda_i)} \right) \boldsymbol{x}_i + I_{(y_i>0)} \left(\frac{y_i}{\lambda_i} - 1 \right) \boldsymbol{x}_i \right\}$$

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \left\{ \frac{-1}{1+\theta} + I_{(y_i>0)} \left(\frac{1}{\theta + \exp(-\lambda_i)} \right) \right\}$$

- That is, under $H_0 : \theta = 0$,

$$S(0, \boldsymbol{\beta}) = \left(\sum_{i} \left(\frac{I_{(y_i=0)}}{\exp(-\lambda_i)} - 1 \right), \sum_{i=1}^{n} \boldsymbol{x}_i \left(\frac{y_i}{\lambda_i} - 1 \right) \right)$$

- The right hand part is the score vector for a Poisson GLM under identity link. So, under H_0, this term would be zero.
- But as we apply constraints, this term does *not* vanish, and needs to be carried along!

Score–test under the identity link

- After some algebra one finds the components of the Fisher matrix $J(0, \boldsymbol{\beta})$ under $H_0 : p = \theta = 0$

$$
\begin{aligned}
J_{\theta\theta} &= \mathrm{E}\left(- \left.\frac{\partial^2 l_{ZIP}}{\partial \theta^2}\right|_{\theta=0} \right) = \sum_{i=1}^{n} \left(\exp(\lambda_i) - 1 \right), \\
J_{\theta\boldsymbol{\beta}} &= \mathrm{E}\left(- \left.\frac{\partial^2 l_{ZIP}}{\partial \theta \partial \beta_j}\right|_{\theta=0} \right) = - \sum_{i=1}^{n} \boldsymbol{x}_i, \\
J_{\boldsymbol{\beta}\boldsymbol{\beta}^T} &= \mathrm{E}\left(- \left.\frac{\partial^2 l_{ZIP}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right|_{\theta=0} \right) = \sum_{i=1}^{n} \frac{1}{\lambda_i} \boldsymbol{x}_i \boldsymbol{x}_i^T .
\end{aligned}
$$

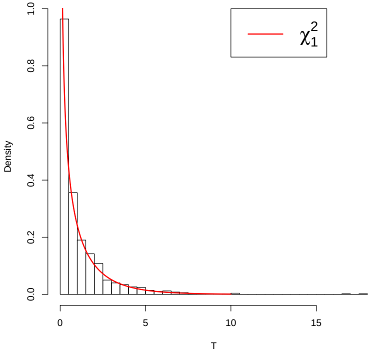
  which completes the test statistic

$$
T = S(0, \hat{\boldsymbol{\beta}})^T J(0, \hat{\boldsymbol{\beta}})^{-1} S(0, \hat{\boldsymbol{\beta}}) \qquad ,
$$

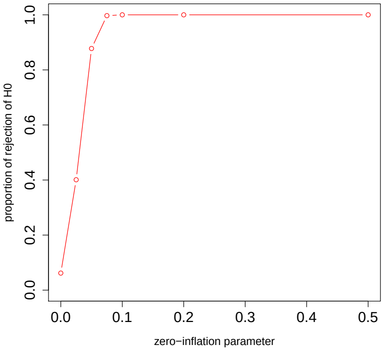
- ... and where $\hat{\lambda}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ is estimated under the Poisson model.

Properties

- Distribution of T under H_0 (values of T for 1000 data sets generated from a true Poisson model):

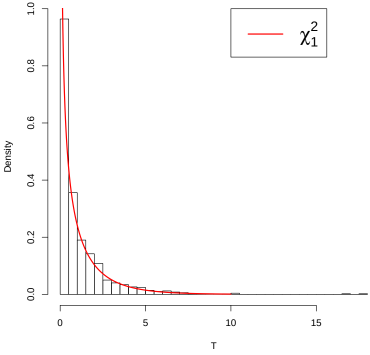


- Power (proportions of rejection of H_0, each for 1000 data sets generated under true p):

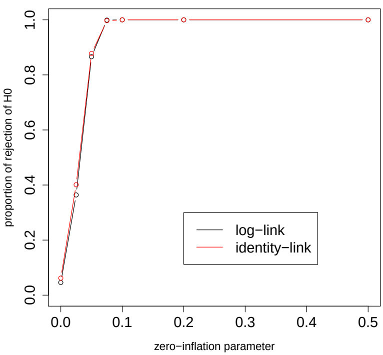


Properties

- Distribution of T under H_0 (values of T for 1000 data sets generated from a true Poisson model):



- Power (proportions of rejection of H_0, each for 1000 data sets generated under true p):



Results

- Test applied on 8 data sets (of the type shown initially).
- Critical value for $\alpha = 0.05$ is $\chi^2_{1,0.95} = 3.84$.

	Whole body exposure				Partial body exposure			
LET	low		high		low		high	
id	18.17	0.92	87.72	61.32	2007.39	1418.28	416.20	387.91
log	*16.89*	*1.00*	*87.16*	*47.20*	*1996.30*	*1417.96*	*421.48*	*398.38*

- All data sets except a single one are zero–inflated!!
- Results for the two link functions are quite similar.

Conclusion

- Driven by practical needs, we developed a score–test for zero–inflation of Poisson models under the identity link.
- Zero–inflated models work generally well for cytogenetic biomarkers.
- The Poisson identity link may be nicer to communicate to the practician. For the Statistician, it is rather troublesome...

Conclusion

- Driven by practical needs, we developed a score–test for zero–inflation of Poisson models under the identity link.
- Zero–inflated models work generally well for cytogenetic biomarkers.
- The Poisson identity link may be nicer to communicate to the practician. For the Statistician, it is rather troublesome...
- References:

    - van den Broek, J. (1995). A score–test for zero–inflation in a Poisson distribution. *Biometrics* **51**, 738–743.
    - Oliveira, M. et al. (2015): Zero–inflated regression models for radiation–induced chromosome aberration data: A comparative study. *Under revision.*