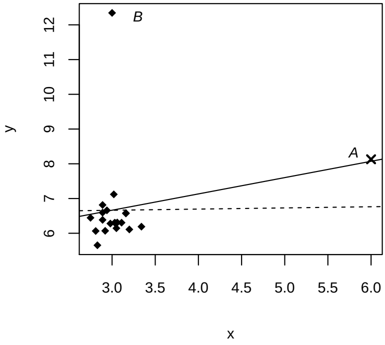
Explicit influence analysis for count data under AB–BA crossover trials

Chengcheng Hao
Department of Statistics
Stockholm University
Sweden

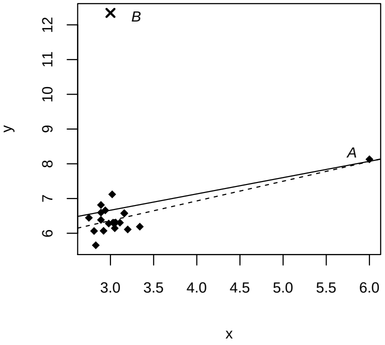
Joint work with Tatjana von Rosen (Stockholm University) and Dietrich von Rosen (Swedish University of Agricultural Sciences)

LinStat2014, LiU, 2014.08.28

Influential observations



(a) The removal of point A



(b) The removal of point B

Possible reasons for **'influential observations'**:

- (i) possible gross **errors** due to data processing or due to measurement errors;
- (ii) responses that due to **chance** occur at the tail of the distribution;
- (iii) inaccuracy of the model in describing **small subpopulations** of the data;
- (vi) inadequacy of the model when modelling **small subpopulations** of the data.

In summary, assessing influence of cases could help researchers

- to **measure** the reliability of the scientific conclusions and
- to **identify** gross errors or important subpopulations.

In this sense, for **every** statistical conclusion, it needs such **explorative data analysis**, which gives analysts an deeper insight into the data.

Aim

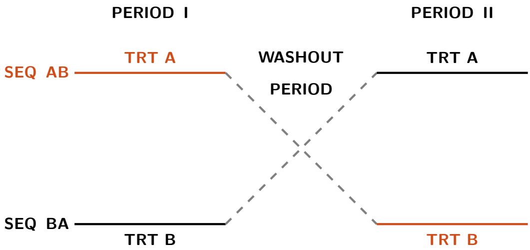
- In this work, we aim to propose tools for conducting **influence analysis** in crossover models with random individual effects.

- **Crossover designs:** designs in which subjects receive different treatments in certain orders.

- The term **crossover model** is a general name for models which are developed to investigate data obtained from studies with crossover designs.

Crossover designs

- Example: AB–BA crossover designs
- Suppose that $2n$ subjects are in an experiment in order to compare the effects of two treatments, **TRT A** and **TRT B**.



- AB–BA crossover study with count data

Layard and Arvesen (1978) discussed a crossover clinical trial to test a standard anti-nausea treatment (drug A) against a proposed treatment (drug B). Twenty subjects were tested, ten for each order of administration of drugs. The response variable is the number of episodes of nausea during the first two hours after cancer chemotherapy, and for a given patient this is approximately Poisson distribute. Wash-out period existed between treatment periods.

Crossover modelling

- Let y_{ijk} represent a count response observed during the k-th period on jth subject within the i-th sequence in a crossover study, with $i = 1, 2;\ j = 1, 2, \ldots, m;\ k = 1, 2$. The Poisson crossover model is set up as

$$\begin{cases} y_{ijk} \,|\, \gamma_{j(i)} \sim Po\,(\lambda_{ijk}),\ \gamma_{j(i)} \sim N(0, \sigma_\gamma^2), & \text{for all } i, j, k, \\ \ln \lambda_{ijk} = \mu_i + \pi_k + \tau_{d(i,k)} + \gamma_{j(i)}, \end{cases} \tag{1}$$

  - μ_i is the **general mean** of the ith **sequence**;
  - ϕ_k is the effect of the kth **period**;
  - The value of $d(i, k)$ is the treatment assigned under sequence i in the kth period; $\tau_{d(i,k)}$ is the **treatment effect** due to treatment $d(i, k)$;
  - $\gamma_{j(i)}$ represents **random individual effect** of the jth subject within sequence i, which is assumed to be $\gamma_{j(i)} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\gamma^2)$;
  - The variances σ_γ^2 is supposed to be unknown.

Generalized mixed linear model (GMLM; Pierce et al. 1975).

Influence analysis in GMLMs

- In the generalized mixed linear models, seldom studies discussed influence analysis in generalized mixed linear models.
  - The likelihood function includes high-dimensional integrals.
  - Ouwen et al. (2001): numerical curvature of the likelihood displacement as an influence measure.
  - Using the Q-function of the EM-algorithm instead of likelihood funtion, Zhu et al. (2003) and Xu et al. (2006) derived an analytic approximated curvature and one-step estimate, respectively.
- Remarks on crossover modelling
  - Most applications **focus on comparisons of the treatments**, while controlling for the nuisance effects.
  - Hao el al. (2014) showed influence analysis of crossover models for continuous data can carry out in two independent fixed-effect models.

- Thus, the overall purpose is to propose a new method to carry out influence analysis in crossover models for count data with several novel features.

- to enable evaluations of influence for different estimates or tests of the interest.
- to underline an important group of mixed models being disregarded by the previous literature on influence analysis.
- to extend our understanding and interpretation of the proposed influence measures.

- **Tools:**
  - Perturbation scheme
  - Objective function of influence

Case-weighted perturbation

Perturbation: the possible deviations of the observed data, which are defined by the **perturbation scheme**.
ω the **perturbation weight**.
K the subset of **indices for perturbed observations**.
$[K]$ the subset of **indices for unperturbed observations**.

A case-weighted perturbation scheme exists with respect to K if

(i) $\hat{\beta}(0)$ is the same as the estimate under the unperturbed model based on observations corresponding to $[K]$;

(ii) there is some null perturbation weight ω_0 such that $\hat{\beta}(\omega_0)$ is the same as the estimate under the unperturbed model based on all observations.

Example: If $\boldsymbol{y}_K$ and $\boldsymbol{y}_{[K]}$ are independent, log-likelihood $l(\omega) = \omega l_K + l_{[K]}$.

Perturbations for count data

- **Scheme I: case-weighted perturbation.**

$$\boldsymbol{y}(\omega) = \left(\omega \boldsymbol{y}_K^{\mathrm{T}}, \boldsymbol{y}_{[K]}^{\mathrm{T}}\right)^{\mathrm{T}}, \boldsymbol{\eta}(\omega) = \left(\boldsymbol{\eta}_K^{\mathrm{T}} + (\ln \omega)\mathbf{1}_p^{\mathrm{T}}, \boldsymbol{\eta}_{[K]}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \omega = 1, 2, \ldots, \tag{2}$$

where p is the size of set K.

For Poisson crossover model and $ijk \in K$,

$$y_{ijk}(\omega) = \omega y_{ijk},$$
$$\eta_{ijk}(\omega) = \ln \omega + \eta_{ijk} = \ln \omega + \mu_i + \boldsymbol{x}_{2,ijk}^{\mathrm{T}} \boldsymbol{\beta} + \gamma_{j(i)},$$

Interpretation: the length of the period for each treatment for the jth subject within sequence i increases by $(\omega - 1)$ times. For $\omega \to 0$, the length of periods reduces to 0 and, therefore, both the expectation and variance of the perturbed responses are 0.

If a single subject is perturbed $K = \{ij1, ij2\}$, the contribution of the perturbed subject to the likelihood is given by

$$\int g(\gamma_{j(i)}) \prod_{ijk \in K} \exp\left[y_{ijk}(\omega)\eta_{ijk}(\omega) - \exp\{\eta_{ijk}(\omega)\} - \ln y_{ijk}(\omega)!\right] d\gamma_{j(i)},$$

where $g(\gamma_{j(i)}) = (2\pi\sigma_\gamma^2)^{-1/2} \exp\{-\gamma_{j(i)}^2/(2\sigma_\gamma^2)\}$ is the density of $\gamma_{j(i)} \sim N(0, \sigma_\gamma^2)$.

Other perturbation schemes

- **Scheme II: perturbation of shifting response.**

$$\boldsymbol{y}(\omega) = \left(\boldsymbol{y}_K^{\mathrm{T}} + \omega \mathbf{1}_p^{\mathrm{T}}, \boldsymbol{y}_{[K]}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \omega = 0, 1, 2, \ldots,$$

- **Scheme III: perturbation of shifting response.**

$$\boldsymbol{y}(\omega) = \left(\omega \boldsymbol{y}_K^{\mathrm{T}}, \boldsymbol{y}_{[K]}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \omega = 0, 1, 2, \ldots,$$

where p is the size of K.

Delta-beta influence

- An appropriate **objective function of influence** is supposed to be proposed according to the inferential interest and the application.
- **Delta-beta influence:**

$$\Delta\widehat{\beta} \quad = \widehat{\beta}(\omega) - \widehat{\beta}(\omega_0).$$

Crossover modelling: AB–BA designs

- Reparametrization of the parameters:

$$\pi_1 = -\pi_2 = \pi/2, \quad \tau_A = -\tau_B = \tau/2.$$

- In the matrix notation, linear predictor can be rewritten as

$$\ln \boldsymbol{\lambda} = \mathbf{X}_1 \boldsymbol{\mu} + \mathbf{X}_2 \boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \tag{3}$$

$\boldsymbol{\mu} = (\mu_1, \mu_2)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\pi, \tau)^{\mathrm{T}}$, $\boldsymbol{\gamma} : 2m \times 1$, $\mathbf{Z} = \mathbf{I}_{2m} \otimes \mathbf{1}_2$ and

$$\mathbf{X}_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & -1/2 \\ 1/2 & 1/2 \\ -1/2 & -1/2 \\ \vdots & \vdots \\ 1/2 & -1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

Crossover modelling: AB–BA designs

Let Y_1 and Y_2 be independent discrete random variables, where

$$Y_1 \sim Po(\lambda_1)$$

and

$$Y_2 \sim Po(\lambda_2).$$

Then their sum $N = Y_1 + Y_2$ is distributed as

$$N \sim Po(\lambda_1 + \lambda_2)$$

and the conditional distribution

$$Y_1 | N = n \sim Bin(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}).$$

Crossover modelling: AB–BA designs

Two separate models:

Let $n_{ij} = y_{ij1} + y_{ij2}$, $i = 1, 2$, $j = 1, \ldots, m$.

$$\begin{cases} n_{ij} \mid \gamma_{ij} \sim Po\left(\delta_{ij}\right), \; \gamma_{ij} \sim N(0, \sigma_\gamma^2), \text{ for all } i, j, \\ \ln \boldsymbol{\delta} = \tilde{\mathbf{X}}_1 \boldsymbol{\mu}^* + \tilde{\mathbf{Z}} \boldsymbol{\gamma}, \end{cases} \tag{3}$$

and

$$\begin{cases} y_{ij1} \sim Bin\left(n_{ij}, p_{ij}\right), \text{ for all } i, j, \\ \mathrm{logit}(\boldsymbol{p}) = \tilde{\mathbf{X}}_2 \boldsymbol{\beta}, \end{cases} \tag{4}$$

where

- $\tilde{\mathbf{X}}_1 = \mathbf{T}_1 \mathbf{X}_1$, $\tilde{\mathbf{X}}_2 = \mathbf{T}_2 \mathbf{X}_2$ and $\tilde{\mathbf{Z}} = \mathbf{T}_1 \mathbf{Z}$,
- $\mathbf{T}_1 = \left(\frac{1}{2} \quad \frac{1}{2}\right) \otimes \mathbf{I}_{2m}$ and $\mathbf{T}_2 = (1 \; -1) \otimes \mathbf{I}_{2m}$,
- The parameters $\boldsymbol{\mu}$, β and σ_γ^2 are the same as the ordinal model.

Proposed methodology

- **Principal ideas:**
  - Split a crossover model with random subject effects into several independent models, where at least the interested parameters are included in fixed effects models.
  - Check whether the perturbation scheme affect division after perturbation.
  - Utilising the explicit updating formula for the fixed effects models.

Results for case-weighted perturbation

Consider the case-weighted perturbation in (2) for the jth subject in sequence i i.e. $K = \{ij1, ij2\}$.

The change in the estimates of period and treatment effect $\hat{\beta}$ due to perturbation is given by

$$\Delta\hat{\beta} = \hat{\beta}(\omega) - \hat{\beta}(1) = \frac{1}{2}\ln\left\{\frac{1 + (\omega - 1)y_{ij1}(m\bar{y}_{i\cdot 1})^{-1}}{1 + (\omega - 1)y_{ij2}(m\bar{y}_{i\cdot 2})^{-1}}\right\}\begin{pmatrix} 1 & (-1)^{i+1} \end{pmatrix}^{\mathrm{T}},$$

for $i = 1, 2$; $j = 1, \ldots, m$, where m is the number of subjects within each sequence and $\bar{y}_{i\cdot k} = \frac{1}{m}\sum_{j=1}^{m} y_{ijk}$.

A useful measure for assessing influence of subjects on the estimates of period and treatment effects is suggested from the above result. We can do the following series expansion when $(\omega - 1)y_{ij1}(m\bar{y}_{i\cdot1})^{-1} < 1$ and $(\omega - 1)y_{ij2}(m\bar{y}_{i\cdot2})^{-1} < 1$,

$$\ln\left\{\frac{1+(\omega-1)y_{ij1}(m\bar{y}_{i\cdot1})^{-1}}{1+(\omega-1)y_{ij2}(m\bar{y}_{i\cdot2})^{-1}}\right\}$$

$$= \sum_{r=1}^{\infty}\frac{(-1)^{r-1}}{r}\left(\frac{y_{ij1}}{m\bar{y}_{i\cdot1}}\right)^r(\omega-1)^r - \sum_{r=1}^{\infty}\frac{(-1)^{r-1}}{r}\left(\frac{y_{ij2}}{m\bar{y}_{i\cdot2}}\right)^r(\omega-1)^r$$

$$= \sum_{r=1}^{\infty}\frac{(-1)^{r-1}}{r}\left\{\left(\frac{y_{ij1}}{m\bar{y}_{i\cdot1}}\right)^r - \left(\frac{y_{ij2}}{m\bar{y}_{i\cdot2}}\right)^r\right\}(\omega-1)^r \tag{5}$$

$$= \left(\frac{y_{ij1}}{\bar{y}_{i\cdot1}} - \frac{y_{ij2}}{\bar{y}_{i\cdot2}}\right)\sum_{r=1}^{\infty}\sum_{s=1}^{r}\frac{(-1)^{r-1}}{m^r r}\left(\frac{y_{ij1}}{\bar{y}_{i\cdot1}}\right)^{r-s}\left(\frac{y_{ij2}}{m\bar{y}_{i\cdot2}}\right)^{s-1}(\omega-1)^r. \tag{6}$$

The influence on the estimate of β is largely dependent to

$$
\begin{aligned}
d_{ij} &= y_{ij1}(\bar{y}_{i\cdot 1})^{-1} - \; y_{ij2}(\bar{y}_{i\cdot 2})^{-1} \\
&= \frac{1}{\bar{n}_{i\cdot}}\left(\frac{y_{ij1}}{\bar{y}_{i\cdot 1}/\bar{n}_{i\cdot}} - \frac{y_{ij2}}{\bar{y}_{i\cdot 2}/\bar{n}_{i\cdot}}\right) = \frac{1}{\bar{n}_{i\cdot}}\left(\frac{y_{ij1}}{\hat{p}_{ij}} - \frac{n_{ij}-y_{ij1}}{1-\hat{p}_{ij}}\right) \\
&= \frac{y_{ij1}-n_{ij}\hat{p}_{ij}}{\bar{n}_{i\cdot}\hat{p}_{ij}(1-\hat{p}_{ij})},
\end{aligned}
$$

where $\bar{n}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} n_{ij}$ and $\hat{p}_{ij} = \mathrm{logit}^{-1}(\tilde{\boldsymbol{x}}_{2,ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \bar{y}_{i\cdot 1}/\bar{n}_{i\cdot}$. Note that d_{ij} is proportional to the Pearson residual of the unperturbed model for the jth subject within sequence i given by

$$
\tilde{\chi}_{ij} = \frac{y_{ij1}-n_{ij}\hat{p}_{ij}}{\sqrt{n_{ij}\hat{p}_{ij}(1-\hat{p}_{ij})}}.
$$

Conclusions

- This work is an example that influence measures are based on residuals even in some generalized mixed linear models. This simplifies the understanding and interpretation of the proposed influence measures.
- A new feature of this method is that we first decompose the mixed model into two independent models, where one is fixed effects model, and then the explicit measures of influence for model parameters are derived. For this reason, the proposed approach is both statistically and computationally effective.
-
- Although not shown here, graphical tools according to the terms in series expansion can be used to explore influential observations in crossover model for count data.

Further research problems

- Mixed linear model with explicit maximum likelihood estimates;
- Generalized linear mixed model with dispersion parameters;
- Influence on the predictions of random effects.

Thanks for your attention!