Surrogate marker evaluation when data are small, large, or very large

Geert Molenberghs

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

Universiteit Hasselt & KU Leuven, Belgium

geert.molenberghs@uhasselt.be & geert.molenberghs@kuleuven.be

www.ibiostat.be

London, March 30, 2017

Broad Principle: Pseudo-likelihood

- Arnold and Strauss (1991)
- Geys, Molenberghs, and Ryan (1999)
- Molenberghs and Verbeke (2005)
- **Units:** clusters, repeated measures, spatial data, microarrays,...

$$f(y_1, y_2, y_3) \quad \longleftrightarrow \quad f(y_1|y_2, y_3) \cdot f(y_2|y_1, y_3) \cdot f(y_3|y_1, y_2)$$

$$f(y_1, y_2, y_3) \quad \longleftrightarrow \quad f(y_1, y_2) \cdot f(y_1, y_3) \cdot f(y_2, y_3)$$

$$f(y_{i1}, \ldots, y_{in_i})$$

replaced by a product of convenient factors

- The **wrong** likelihood used
- The **right** results obtained:
  - ▷ Consistent, asymptotically normal estimators
  - ▷ Often minor loss of statistical efficiency
  - ▷ Often major gain of computational efficiency

Specific Use 1:
Pseudo-likelihood for HD Multivariate Longitudinal Data

- Fieuws and Verbeke (2006); Fieuws *et al* (2006)
- M sequences of repeated measures
- **Example:** 44 sequences of hearing variables

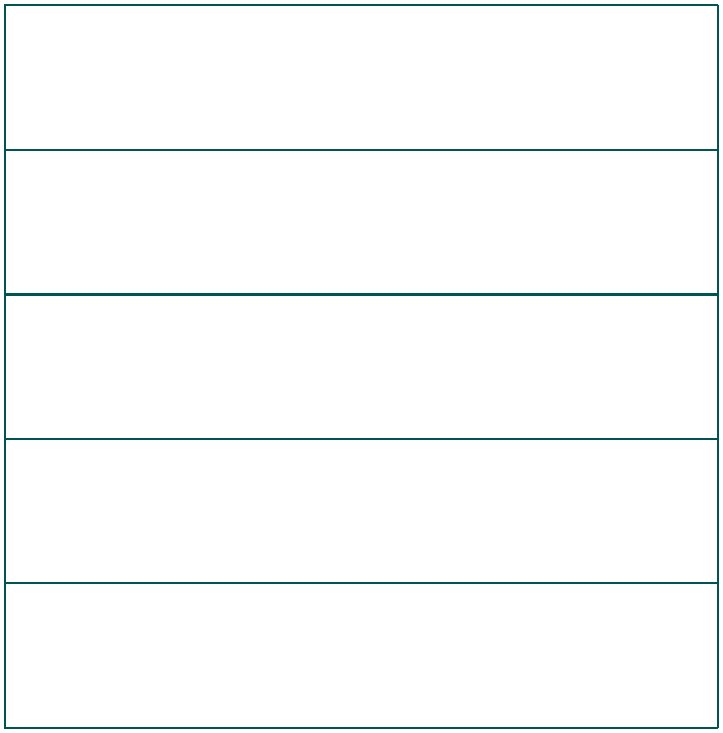
- Data for patient i:

Y_{i11}	Y_{i12}	Y_{i13}	$\cdots$	Y_{i1n_i}
Y_{i21}	Y_{i22}	Y_{i23}	$\cdots$	Y_{i2n_i}
Y_{i31}	Y_{i32}	Y_{i33}	$\cdots$	Y_{i3n_i}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$Y_{i,44,1}$	$Y_{i,44,2}$	$Y_{i,44,3}$	$\cdots$	$Y_{i,44,n_i}$

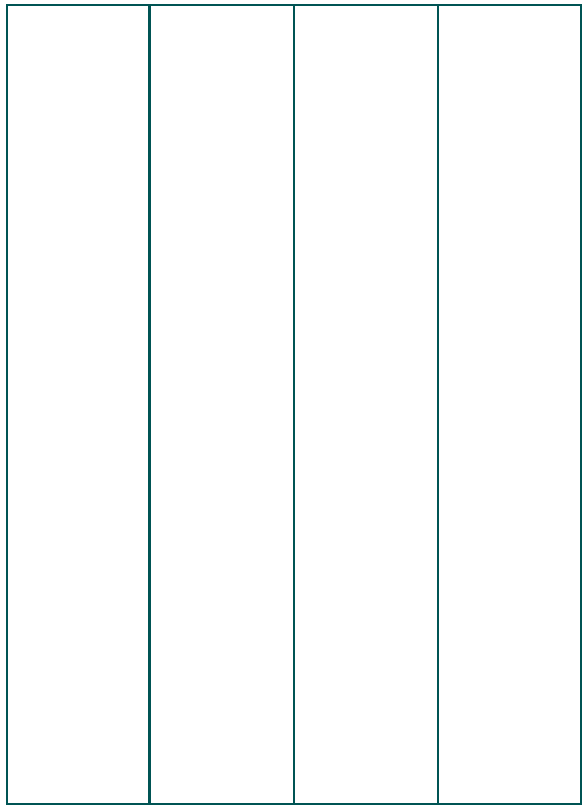
- Fit model to each of the $M(M-1)/2$ pairs

- Use PL to reach valid conclusions

Specific Use 2:
Split Sample Method:
(In)dependent Subsamples



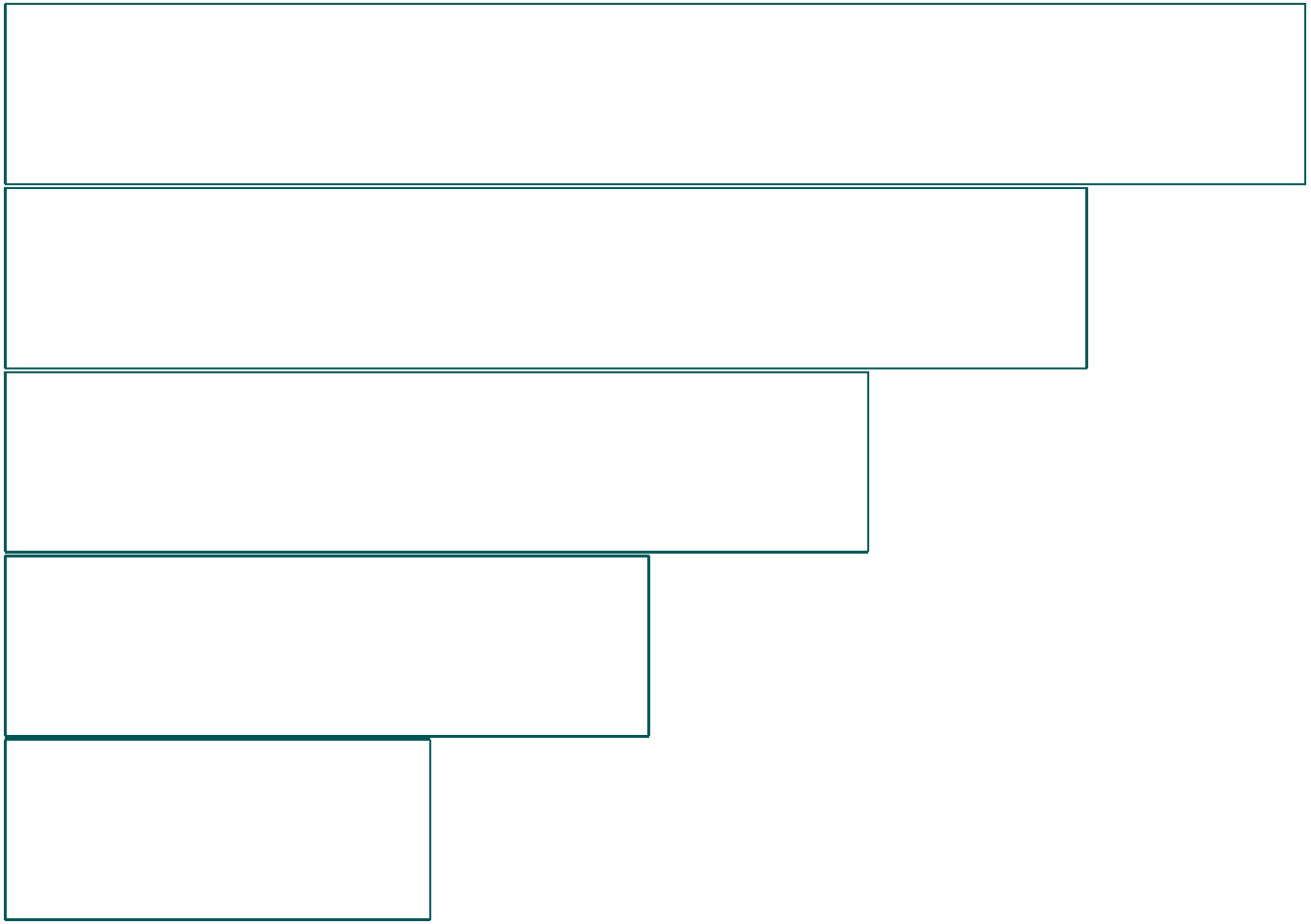
or



Behavior

- **Univariate normal: equivalent**
- **Univariate Bernoulli (probability): equivalent**
- **Univariate Bernoulli (logit): different estimator, same precision**
- **Compound symmetry: different estimator, some precision loss**

Specific Use 3:
Per Cluster Size



Fixed Cluster Size $\longleftrightarrow$ Variable Cluster Size

- **Fixed cluster size:** closed-form maximum likelihood estimator: **easy**

- **Variable cluster size:**

  ▷ Estimate parameters per cluster size

  ▷ Average these to find

  ▷ **But:** Now weighted average needed

Which Weights and Why?

Constant weights
Proportional weights
Optimal weights
Scalar weights
Iterated optimal weights
Approximate optimal weights

Surrogate Markers

- **Model:**

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}$$
$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}$$

- **Error structure:**

  - ▷ **Individual level:**
    - ∗ Deviations ε_{Sij} and ε_{Tij} are correlated
  - ▷ **Trial level:**
    - ∗ Treatment effects α_i and β_i are correlated
    - ∗ (Information from intercepts μ_{Si} and μ_{Ti} can be used as well)

- Estimation can be problematic:
  - ▷ especially in small studies
  - ▷ especialy when studies are of differing sizes
- **Solution 1:** Use multiple imputation to make all studies equally large
- **Solution 2:**
  - ▷ Analyze trial-by-trial: it can be shown that this is valid
  - ▷ Combine results across trials using weighted averages
  - ▷ When some (or all) trials are very large: sub-sampling is allowable
- **Solution 2-advantages:**
  - ▷ **:** Very stable $\longleftarrow$ small trials
  - ▷ **:** Very fast $\longleftarrow$ very large trials

Statistics

Applied Surrogate Endpoint Evaluation Methods with SAS and R provides an overview of contemporary meta-analytic and information-theoretic methodology to evaluate candidate surrogate endpoints from clinical trials and beyond. The book strongly focuses on user-friendly software in both SAS and R for a variety of outcome types.

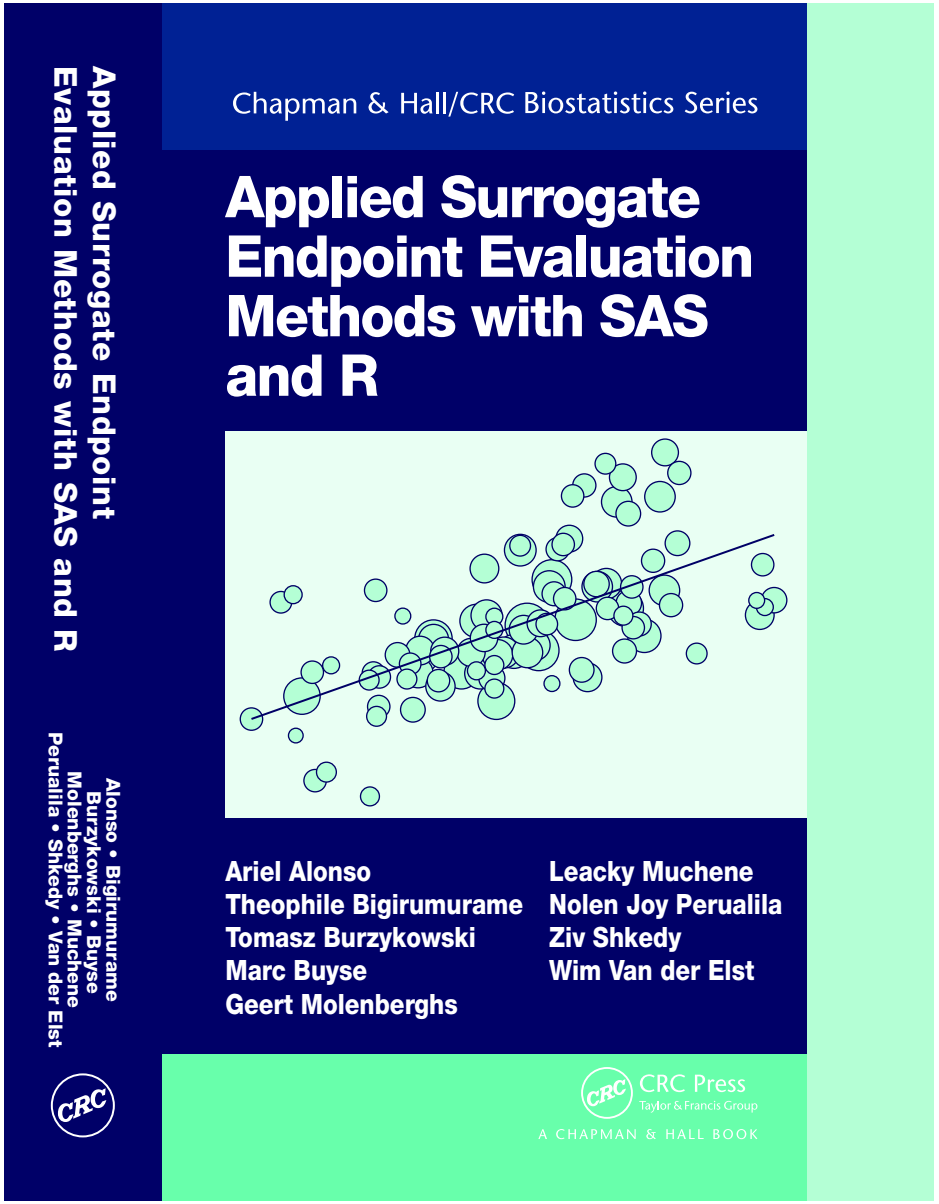
The book is aimed at researchers and practitioners who want to study and apply methodology for surrogate endpoint and biomarker evaluation. Methodology is described while keeping mathematical detail to a minimum. Throughout the book, a suite of generic case studies is used to illustrate the concepts and methodology. A large part of the book is devoted to the description and illustration of SAS macros, R language libraries, and R Shiny Apps. The software tools can be downloaded from the authors' web pages. Methodology, applications, and software encompass continuous, binary, categorical, time-to-event, and longitudinal outcomes.

The University of Hasselt and KU Leuven-based editor team, supplemented by a fine group of chapter authors, has over twenty years of experience in the field of surrogate marker evaluation in clinical and other studies. The book is rooted at the same time in methodological research, regular and short courses taught on the topic, as well as in vast experience with the design and conduct of clinical trials. The team's prolific contributions have led to numerous papers, chapters, and books on this topic. This book was written in a coherent fashion, with common notation, conventions, and case studies throughout all chapters.

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

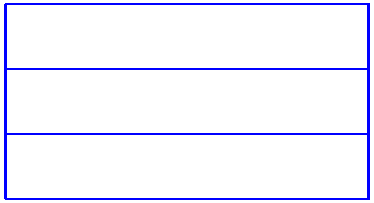
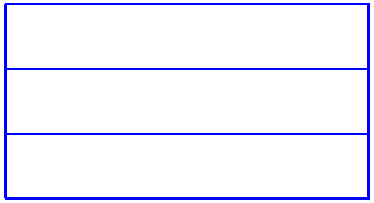
K23717
ISBN: 978-1-4822-4936-1
90000
9 781482 249361
www.crcpress.com



Application: Leuven Diabetes Study

- 120 general practitioners — 2495 patients
- **Outcomes**
  - ▷ **LDL:** low-density lipoprotein cholestrol
  - ▷ **HbA1C:** glycosylated hemoglobin
  - ▷ **SBP:** systolic blood pressure
- **Ordinal targets**
- Multiple outcomes & measured repeatedly & ordinal

$\Longrightarrow$ **joint modeling**

Method	3 sequences	Partitioning	CPU
1 ≡ ML	(123)		7'13"
2 ≡ PLp	(12)(13)(23)		1'23"
3 ≡ PLs	(123)		1'21"
4 ≡ PLps	(12)(13)(23)		0'20"

CPU Gain / Efficiency Loss

- Subsamples can be analyzed in parallel
- Base model above, with numerical integration over $Q = 3$ quadrature points:

  7'13" $\longrightarrow$ 0'20"

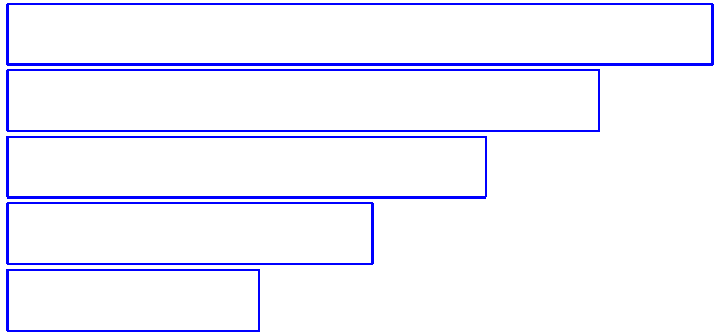
- More demanding integration: $Q = 15$

  10h02'42" $\longrightarrow$ 0h4'17"

- Statistical efficiency: almost always $\geq 95\%$
- For PLps occasionally $85\% - 87\%$

Application: Quantifying Expert Opinion

- Janssen Pharmaceutica
- chemical compound acquisition to diversify library
- 22,015 compounds presented to 147 experts
- **Outcome:** recommended (1) $\longleftrightarrow$ not recommended (0)
- Variable #compounds per expert



- 'Simple' model:

$$\text{logit}\left[P\left(Y_{ij}=1|b_i\right)\right]=\beta_j+b_i$$

  - ▷ b_i: normal random effect of expert i
  - ▷ β_j: potential of compound j
  - ▷ **there are 22,015 β_j's**

Modified Procedure

- Partition β_j's into S mutually exclusive, exhaustive sets
- Fit model to each of the $S = 30$ subsets
- Repeat this $W = 20$ times
- $\simeq$ 96 hours on HPC (Nehalem)
- Can be brought down to 1 hour when parallelized
- Can be optimized further
- Weighted analysis by differing numbers of compounds per expert

Conclusions

- Broad framework based on:
  - ▷ pseudo-likelihood
  - ▷ pairwise modeling
  - ▷ split sample
- Statistically valid procedures: consistent, asymptotically normal
- Can lead to tremendous CPU gain
- Statistical efficiency loss mostly acceptable