When likelihood goes wrong

Nancy Reid
University of Toronto

Nov 20 2024

UCLA Statistics & Data Science

Department of Biostatistics

Outline

1. Examples: a haphazard selection
2. Models and parameters
3. Some approaches to misspecification
4. Conclusion

Examples: a haphazard selection

Climate change

Pfister et al 2024

Clim. Past, 20, 1387–1399, 2024
https://doi.org/10.5194/cp-20-1387-2024

Climate of the Past
Open Access
EGU

600 years of wine must quality and April to August temperatures in western Europe 1420–2019

Christian Pfister[1], Stefan Brönnimann[2], Andres Altwegg[3], Rudolf Brázdil[4], Laurent Litzenburger[5], Daniele Lorusso[6], and Thomas Pliemon[7]

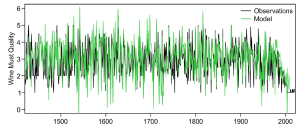


Figure 4. Observed series of wine quality (average; black) from 1420 to 2019 and series obtained with a statistical model calibrated in 1781–1800 (green). The model is explained in Sect. 3.

Scientific question: Can historical records of wine quality be used as temperature proxies?

observational data

Statistical model: "we used a statistical [linear regression] model for wine quality based on local temperature and precipitation"

yes, if used carefully

Income supplements

Vivalt et al 2024

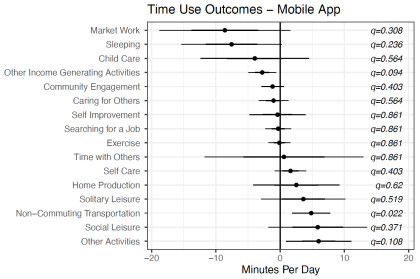
NBER WORKING PAPER SERIES

THE EMPLOYMENT EFFECTS OF A GUARANTEED INCOME:
EXPERIMENTAL EVIDENCE FROM TWO U.S. STATES

Eva Vivalt
Elizabeth Rhodes
Alexander W. Bartik
David E. Broockman
Sarah Miller

Working Paper 32719
http://www.nber.org/papers/w32719

Figure 5: Time Use Results: Mobile App



Scientific question: Does guaranteed income supplement affect labor market measures?

randomized controlled trial

Statistical model: $Y_i = \alpha + \beta Treated_i + \gamma^{\mathrm{T}} X_i + \epsilon_i$

"support for both sides of this debate"

Breast cancer mortality

Giannakeas et al 2024

JAMA Oncology | Original Investigation

Bilateral Mastectomy and Breast Cancer Mortality

Vasily Giannakeas, PhD, MPH; David W. Lim, MDCM, MEd, PhD; Steven A. Narod, MD

IMPORTANCE The benefit of bilateral mastectomy for women with unilateral breast cancer in terms of deaths from breast cancer has not been shown.

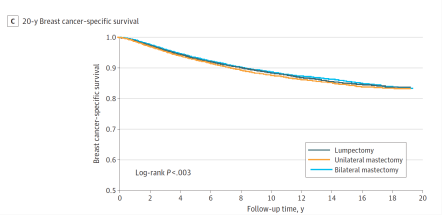
OBJECTIVES To estimate the 20-year cumulative risk of breast cancer mortality among women with stage 0 to stage III unilateral breast cancer according to the type of initial surgery performed.

DESIGN, SETTINGS, AND PARTICIPANTS This cohort study used the Surveillance, Epidemiology, and End Results (SEER) Program registry database to identify women with unilateral breast cancer (invasive and ductal carcinoma in situ) who were diagnosed from 2000 to 2019. Three closely matched cohorts of equal size were generated using 1:1:1 matching according to surgical approach. The cohorts were followed up for 20 years for contralateral breast cancer and for breast cancer mortality. The analysis compared the 20-year cumulative risk of breast cancer mortality for women treated with lumpectomy vs unilateral mastectomy vs bilateral mastectomy. Data were analyzed from October 2023 to February 2024.

EXPOSURES Type of breast surgery performed (lumpectomy, unilateral mastectomy, or bilateral mastectomy).

Editorial

Supplemental content



Scientific question: Does bilateral mastectomy for unilateral breast cancer improve 20-year survival?

matched case-cohort study

Statistical model: “We used the Kaplan-Meier method to estimate survival”

“preemptive surgery did not appear to reduce the risk of death”

Dementia and mortality

Jang et al 2023

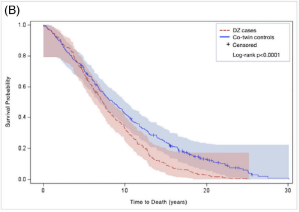
DOI: 10.1002/alz.13553

RESEARCH ARTICLE

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

Dementia and mortality in older adults: A twin study

Jung Yun Jang[1] | Christopher R. Beam[2,3] | Ida K. Karlsson[4] | Nancy L. Pedersen[2,4] | Margaret Gatz[4,5]



Scientific question: Relationship between dementia and mortality

observational study of discordant twin pairs

Statistical model: multi-level Cox regression with random effects

"genetic variance contributes to the association between dementia risk and mortality"

Article

https://doi.org/10.1038/s41550-023-01983-1

Variability of extragalactic X-ray jets on kiloparsec scales

Received: 17 May 2022

Accepted: 27 April 2023

Published online: 29 May 2023

Eileen T. Meyer[1], Aamil Shaik[1], Yanbo Tang[2], Nancy Reid[3], Karthik Reddy[1,4], Peter Breiding[2], Markos Georganopoulos[1], Marco Chiaberge[5,6], Eric Perlman[7], Devon Clautice[7], William Sparks[8,9], Nat DeNigris[1,10] & Max Trevor[1,11]

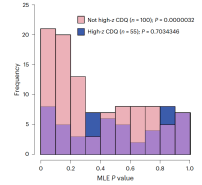


Fig. 3 | Histogram of the single-region *P* values from the directional test, not adjusted for multiple comparison. In pink, the subset of sources that

Yanbo Tang

Scientific question: Are observations of *X*-ray jets consistent with current theory?

observational data

Statistical model: compare background and sources measurements using Poisson:

$$x_i \sim Po(a_i\beta_i), \quad y_i \sim Po\left(b_i(\beta_i + f_i\mu_i)\right) \qquad H : \mu_i \equiv 0$$

"variability in the *X*-ray emission is not compatible with proposed mechanism"

Models and parameters

Why these models?

- standard in the literature of that field

income

- standard in the publications of that lab

breast cancer

- follow some prescription:
  - binary response — use logistic regression
  - time to event — use PH model
  - time series — use ARMA

wine

  - repeated measures — use random effects

Alzheimer's twin study

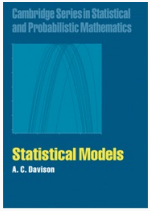
  - ...

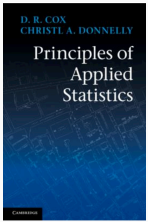
- motivated by theory: economic, physical, ...

X-ray jets

Some guidance from the experts

Davison; Cox & Donnelly





- the key feature of a statistical model is that variability is represented using probability distributions
- the art of modelling lies in finding a balance that enables the questions at hand to be answered or new ones posed
- probability models as an aid to the interpretation of data
- perturbations of no intrinsic interest distort an otherwise exact measurement
- substantial natural variability in the phenomenon under study

The role of parameters

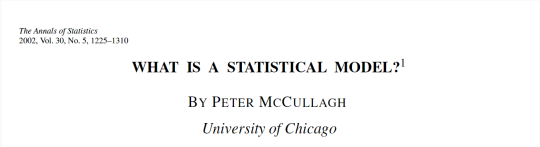
- probability models very likely be parameterized
- thus defining a class of models $\{f(y;\theta);\theta \in \Theta\}$
- parameters may be finite- or infinite-dimensional

parametric vs nonparametric

- ideally one or more parameters represent key aspects of the model

for the application at hand

- other parameters complete the specification
- the meaning of various parameters varies with the application


The Annals of Statistics
2002, Vol. 30, No. 5, 1225–1310

WHAT IS A STATISTICAL MODEL?[1]

BY PETER MCCULLAGH
University of Chicago

- this sounds simpler than it is

e.g. Box-Cox $y^{\lambda} = x^{\mathrm{T}}\beta + \epsilon$

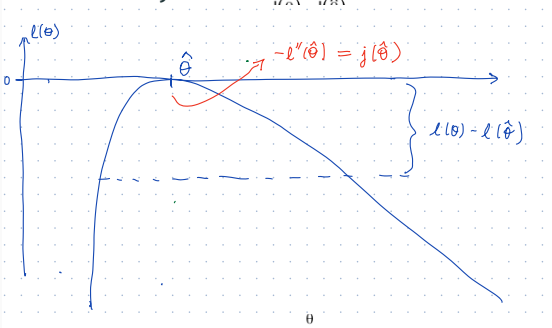
The likelihood function

- puts the emphasis on the model: $L(\theta; y) \propto f(\boldsymbol{y}; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$
- provides a convenient way to compare parameter values

inverse problem

e.g. $L(\theta)/L(\hat{\theta})$

- provides reliable summary measures



- can be converted to a probability, given a prior probability for θ

Pfizer vaccine

$Bin(162 + 8, \theta)$ via 2 Poissons

Limit theory

$$(i)\ \ \ell(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta \mid x_i), \quad (ii)\ \ \ell'(\theta) = \sum_{i=1}^{n} \nabla_\theta \log f(y_i; \theta \mid x_i), \quad (iii)\ \ \ell'(\hat{\theta}) = 0$$

Central Limit Theorem:

$$\frac{1}{\sqrt{n}} \ell'(\theta) \xrightarrow{d} N\{0, I_1(\theta)\}$$

Large-sample approximations:

$$\ell'(\theta) \dot{\sim} N_p\{0, I(\theta)\}, \quad \hat{\theta} \dot{\sim} N_p\{\theta, J^{-1}(\hat{\theta})\}, \quad 2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot{\sim} \chi^2_p$$

$J(\theta) = -\ell''(\theta), \quad I(\theta) = \mathbb{E}_\theta\{j(\theta)\}$

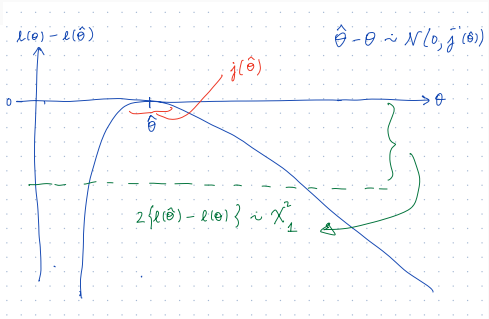
observed and expected Fisher information

… Limit theory

Large-sample approximation:

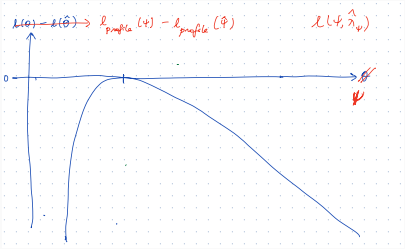
$$\ell'(\theta) \dot{\sim} N_p\{0, I(\theta)\}, \quad \hat{\theta} \dot{\sim} N_p\{\theta, J^{-1}(\hat{\theta})\}, \quad 2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot{\sim} \chi^2_p$$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.079      0.987   -3.12   0.0018 **
aged1         -0.292      0.754   -0.39   0.6988
stage1         1.373      0.784    1.75   0.0799 .
grade1         0.872      0.816    1.07   0.2850
xray1          1.801      0.810    2.22   0.0263 *
acid1          1.684      0.791    2.13   0.0334 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 40.710  on 22  degrees of freedom
Residual deviance: 18.069  on 17  degrees of freedom
```



A bit too simple

- model $f(y;\theta)$, $\quad \theta \in \mathbb{R}^p$
- $\theta = (\psi, \lambda)$ parameters of interest nuisance parameters
- results above used profile log-likelihood function $\ell_{\text{profile}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$



it helps if $\psi \perp \lambda$; $\quad I_{\psi\lambda}(\theta) = 0$

What can go wrong?

- the normal and/or χ^2 approximations might be poor — e.g. likelihood skewed
- too many parameters — $p \sim n^\alpha, \quad p/n \to C, \quad p/n \to \infty$
- irregular parameter space — $pf(y;\theta_1) + (1-p)f(y;\theta_2), \quad 0 \le p \le 1$
- computational intractability — $L(\theta, \tau; y) = \int_{\mathbb{R}^k} f(y \mid z; \theta) f(z; \tau) dz$
- model is misspecified — true $Y \sim m(y), \quad f(\cdot;\theta) \neq m(\cdot) \ \forall \theta$

Some approaches to misspecification

Cox 1961,2; Huber 1967; White 1982

- true model $m(\boldsymbol{y})$ fitted model $f(\boldsymbol{y};\theta)$

$\boldsymbol{y} = (y_1, \ldots, y_n)$

- maximum likelihood estimator $\widehat{\theta}$

$\ell(\theta;\boldsymbol{y}) \equiv \log f(\boldsymbol{y};\theta)$

$\widehat{\theta} \equiv \arg\sup_\theta \ell(\theta;\boldsymbol{y})$

- $\widehat{\theta}$ converges to the "closest true value"

KL-divergence

$$\theta_m^0 = \arg\min_\theta \int m(\boldsymbol{y}) \log\{\frac{m(\boldsymbol{y})}{f(\boldsymbol{y};\theta)}\} d\boldsymbol{y}$$

- $\widehat{\theta}$ has asymptotic normal distribution, but is not fully efficient

"sandwich variance"

$$\text{a.var.}\,(\widehat{\theta}) = G^{-1}(\theta_m^0), \qquad G(\theta) = J(\theta) I^{-1}(\theta) J(\theta)$$

$I = \text{var}_m(\ell')$, $J = \mathbb{E}_m(-\ell'')$

- change the inference goal, proceed more or less as usual

"we used robust standard errors "

2. More flexible inference functions

Composite likelihood

- true model $m(\boldsymbol{y}_i) = f(\boldsymbol{y}_i; \theta), \boldsymbol{y}_i \in \mathbb{R}^d$ fitted model $\prod_{A \in \mathcal{A}} f(y_{iA}; \theta)$ subsets A
- Example: pairwise likelihood $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$

$$L_{pair}(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} \prod_{s \neq t} f_2(y_{is}, y_{it}; \theta)$$

- Example $AR(1)$ likelihood $\boldsymbol{y} = (y_1, \ldots, y_n)$

$$L_{cond}(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i \mid y_{i-1}; \theta)$$

interpretation of θ

- Example pseudo-likelihood in spatial models condition on near neighbours; Besag 74

… More flexible inference functions

Quasi-likelihood and generalized estimating equations

$$g\{\mathbb{E}(y_i \mid \boldsymbol{x}_i)\} = g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \quad \mathrm{var}(y_i \mid \boldsymbol{x}_i) = \sigma^2 V(\mu_i)$$

- estimating equation for $\boldsymbol{\beta}$

full distribution unspecified

$$\sum_{i=1}^{n} \frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{(y_i - \mu_i)}{V(\mu_i)} = \mathbf{o}$$

column vector

Quadratic inference functions

Qu, Lindsay, Li 2000; Hector 2023

- replace $V^{-1}(\mu_i)$ above with an expansion in basis functions
- apply generalized method of moments

3. More flexible models

- identify one or more parameters of interest — here β
- use a highly flexible specification form for other aspects of the model

- Example: proportional hazards regression — instantaneous failure rate

$$h(t; x, \beta) = h_0(t)\exp(x^T\beta)$$

- Example: empirical likelihood — $T(F)$ to be specified; e.g. $\mathbb{E}_F(Y_i)$

$$\max_F L(F; \boldsymbol{y}), \text{ subject to } T(F) = \theta$$

$L(F; \boldsymbol{y}) = \prod_{i=1}^n F(y_i)$

- Example: semi-parametric regression

$$\mathbb{E}(y \mid T, x) = \psi T + \omega(x)$$

- does parameter of interest have a stable interpretation — model assumption

- Possible model $\mathbb{E}(y \mid T, x) = \psi T + \omega(x)$ binary treatment T
- Define an estimand of interest limit of E-estimator, Robins et al 92

$$\frac{\mathbb{E}[\pi(x)\{1-\pi(x)\}\{\mathbb{E}(y \mid T=1,x) - \mathbb{E}(y \mid T=0,x)\}]}{\mathbb{E}[\pi(x)\{1-\pi(x)\}]}$$

propensity score $\pi(x) = \mathrm{pr}(T = 1 \mid x)$

- reduces to ψ under possible model
- is a meaningful quantity when the model is incorrect e.g. interaction between T and x
- linear model $\longrightarrow$ generalized linear model $g\{\mathbb{E}(y \mid T, x)\} = \psi T + \omega(x)$

$$\frac{\mathbb{E}(\pi(x)\{1-\pi(x)\}[g\{\mathbb{E}(y \mid T=1,x)\} - g\{\mathbb{E}(y \mid T=0,x)\}])}{\mathbb{E}[\pi(x)\{1-\pi(x)\}]}$$

- when does parameter of interest have stable interpretation? orthogonality?

Example: exponential matched pairs

Battey & Cox 2020

- independent exponential pairs $(y_{1i}, y_{2i}), \quad i = 1, \ldots, n$ — $n + 1$ parameters
- rate parameters γ_i/ψ and $\gamma_i\psi$, respectively
- ψ common parameter of interest $\quad \gamma_i$ pair-specific nuisance parameters

- likelihood function

$$L(\psi, \boldsymbol{\gamma}; \boldsymbol{y}) \propto \prod_{i=1}^{n} \gamma_i^2 \exp\{-\gamma_i(\frac{y_{1i}}{\psi} + \psi y_{2i})\}$$

- possibilities for eliminating nuisance parameters
  - profile (concentrated) likelihood — maximize over γ
  - marginal likelihood: $f(\boldsymbol{t}; \psi) = \prod_{i=1}^{n} f(t_i; \psi)$ — $t_i = y_{1i}/y_{2i}$
  - random effects $\gamma_i \sim g(\cdot; \boldsymbol{\lambda})$ — more efficient, if ...

- independent exponential pairs (y_{1i}, y_{2i}), $i = 1, \ldots, n$ — $n + 1$ parameters
- rate parameters γ_i/ψ and $\gamma_i\psi$, respectively
- random effects: $\gamma_i \sim \text{Gamma}(\alpha, \beta)$ — λ = (shape, rate)
- likelihood function

$$L(\psi, \alpha, \beta; \boldsymbol{y}) \propto \prod_{i=1}^{n} \int \gamma_i^2 \exp\{-\gamma_i(\frac{y_{1i}}{\psi} + \psi y_{2i})\} g(\gamma_i; \alpha, \beta) d\gamma_i$$

- orthogonality:

$$\mathbb{E}_{gamma}\left\{-\frac{\partial^2 \log L(\psi, \alpha, \beta)}{\partial\psi\partial\alpha}\right\} = 0, \quad \mathbb{E}_{gamma}\left\{-\frac{\partial^2 \log L(\psi, \alpha, \beta)}{\partial\psi\partial\beta}\right\} = 0$$

- even better

$$\mathbb{E}_{m}\left\{-\frac{\partial^2 \log L(\psi, \alpha, \beta)}{\partial\psi\partial\alpha}\right\} = 0, \quad \mathbb{E}_{m}\left\{-\frac{\partial^2 \log L(\psi, \alpha, \beta)}{\partial\psi\partial\beta}\right\} = 0$$

any random effects distribution

- independent exponential pairs $(y_{1i}, y_{2i}), \quad i = 1, \ldots, n$ $n + 1$ parameters
- rate parameters γ_i/ψ and $\gamma_i\psi$, respectively
- random effects: $\gamma_i \sim \text{Gamma}(\alpha, \beta)$ λ = (shape, rate)
- likelihood function

$$L(\psi, \alpha, \beta; \mathbf{y}) \propto \prod_{i=1}^{n} \int \gamma_i^2 \exp\{-\gamma_i(\frac{y_{1i}}{\psi} + \psi y_{2i})\} g(\gamma_i; \alpha, \beta) d\gamma_i$$

- even better

$$\mathbb{E}_m \left\{ -\frac{\partial^2 \log L(\psi, \alpha, \beta)}{\partial\psi\partial\alpha} \right\} = 0, \quad \mathbb{E}_m \left\{ -\frac{\partial^2 \log L(\psi, \alpha, \beta)}{\partial\psi\partial\beta} \right\} = 0$$

- and

$$\hat{\psi} \xrightarrow{p} \psi$$

any random effects distribution

- random effects: $\gamma_i \sim \text{Gamma}(\alpha, \beta)$ λ = (shape, rate)
- density for one pair

$$f(y_{1i}, y_{2i}; \psi, \alpha, \beta) = \int f(y_{1i}, y_{2i}; \psi, \gamma_i) g(\gamma_i; \alpha, \beta) d\gamma_i$$
$$= \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(y_{1i}/\psi + \psi y_{2i} + \beta)^{\alpha+2}}$$

- Fisher information $\mathbb{E}\{-\partial^2 \ell(\boldsymbol{\theta})/\partial\theta\partial\theta^T\} = \mathbb{E}\{-\nabla^2 \ell(\boldsymbol{\theta})\}$ $\boldsymbol{\theta} = (\psi, \alpha, \beta)$

$$\ell_{i,\psi\alpha} = \frac{y_{2i} - y_{1i}/\psi^2}{(y_{1i}/\psi + \psi y_{2i} + \beta)}; \quad \ell_{i,\psi\beta} = \frac{-(\alpha+2)(y_{2i} - y_{1i}/\psi^2)}{(y_{1i}/\psi + \psi y_{2i} + \beta)^2}$$

$\ell_i(\cdot) = \log f(y_{1i}, y_{2i}; \cdot)$

$$\mathbb{E}_{gamma}(\ell_{i,\psi\alpha}) = 0 = \mathbb{E}_{gamma}(\ell_{i,\psi\beta})$$
$$\mathbb{E}_m(\ell_{i,\psi\alpha}) = 0 = \mathbb{E}_m(\ell_{i,\psi\beta})$$

PNAS RESEARCH ARTICLE | STATISTICS OPEN ACCESS

On the role of parameterization in models with a misspecified nuisance component

Heather S. Battey[a,2,1] and Nancy Reid[b,2,1]

Contributed by Nancy Reid; received February 8, 2024; accepted July 23, 2024; reviewed by Emmanuel J. Candès and Edward I. George

August 30, 2024 | 121 (36) e2402736121

Towards a formalisation

- true model $m(\boldsymbol{y})$ with parameter ψ and true value ψ_*
- fitted model $f(\boldsymbol{y}; \psi, \lambda)$ same parameter of interest, (many) nuisance parameters
  interpretation of ψ is stable
- we know maximum likelihood estimates $(\hat{\psi}, \hat{\lambda}) \xrightarrow{p} (\psi_m^o, \lambda_m^o)$ KL divergence
- assume no value of $\lambda \in \Lambda$ gives back $m(\cdot)$ 'truly' misspecified
- Does $\psi_m^o = \psi_*$? need $\mathbb{E}_m\{\partial \ell(\psi_*, \lambda_m^o)/\partial \psi\} = 0$ (1) λ_m^o unknown
- can be easier to show $\mathbb{E}_m\{\partial \ell(\psi_*, \lambda)/\partial \psi\} = 0 \quad \forall \lambda$ (2)
- Result 1: (1) $\equiv$ (2) $\iff$ ψ_* is m-orthogonal to Λ: $\forall \lambda \quad \mathbb{E}_m \left\{ \frac{\partial^2 \ell(\psi, \lambda)}{\partial \psi \partial \lambda} \right\} = 0$

BR: Prop 1.1

- true model $m(\boldsymbol{y})$ with parameter ψ and true value ψ_*
- fitted model $f(\boldsymbol{y}; \psi, \lambda)$, maximum likelihood estimate $\widehat{\psi}$
- Result 1: m-orthogonal parameters lead to consistent MLE $\psi_m^{\mathrm{o}} = \psi_*$

- But, $\widehat{\psi}$ can be consistent without this requirement

- Result 2: A weaker requirement: if BR: Prop 1.2

$$I^{\psi\psi}\mathbb{E}_m\left\{\frac{\partial \ell(\psi_*, \lambda)}{\partial \psi}\right\} + I^{\psi\lambda}\mathbb{E}_m\left\{\frac{\partial \ell(\psi_*, \lambda)}{\partial \lambda}\right\} = 0, \quad \forall\, \lambda, \text{ then } \psi_m^{\mathrm{o}} = \psi_*$$

$$I = I(\psi_*, \lambda) = \mathbb{E}_m\left\{-\frac{\partial^2 \ell(\psi_*, \lambda)}{\partial \theta \partial \theta^{\mathrm{T}}}\right\}, \qquad I^{-1} = \begin{pmatrix} I^{\psi\psi} & I^{\psi\lambda} \\ I^{\lambda\psi} & I^{\lambda\lambda} \end{pmatrix},$$

still too strong

Parameter orthogonality

- we can often establish parameter orthogonality in the assumed model $f(\boldsymbol{y}; \psi, \lambda)$
- all expectations with respect to this assumed model
- this is not usually the same as parameter orthogonality in the true model $m(\boldsymbol{y}; \psi)$
- Result 3: a special case BR: Prop 1.3

  If $\nabla^2_{\psi\lambda}\ell(\psi, \lambda; \boldsymbol{y})$ a function of $S = (S_1, \ldots, S_k)$, and is additive in S, and

  $$E_m(S_j) = \mathbb{E}_{(\psi,\lambda)}(S_j)$$

  then assumed-model orthogonality $\implies$ true-model orthogonality
- easier: information calculations under assumed model

Parameter Symmetry

- matched exponential pairs is a scale model: $\mathbb{E}(Y_{1i}) = \psi/\gamma_i; \quad \mathbb{E}(y_{2i}) = 1/(\psi\gamma_i)$
- the parameter of interest enters symmetrically
- the proof of consistency repeatedly uses the change of variables to y_{1i}/y_{2i} and $y_{1i}y_{2i}$
- how to generalize this observation?
- first, another matched pairs example, even simpler

... Parameter Symmetry

- $Y_{1i} \sim N(\gamma_i + \psi, 1), \quad Y_{2i} \sim N(\gamma_i - \psi, 1), \quad i = 1, \ldots, n$

- mixing distribution $h(\cdot; \lambda)$ on nuisance parameters
- likelihood and log-likelihood function

$$L(\psi; y_{1i}, y_{2i}, \lambda) = \int \exp\{-\frac{1}{2}\{(y_{1i} - \gamma_i - \psi)^2 + (y_{2i} - \gamma_i + \psi)^2\} h(\gamma_i; \lambda) d\gamma_i$$

$$\log L(\psi; \boldsymbol{y}, \lambda) = \ell(\psi; \boldsymbol{y}, \lambda) = -\frac{1}{2}[\sum \{y_{1i} - \psi)^2 + (y_{2i} + \psi)^2\} + n \log k(\lambda)]$$

$$= \qquad \ell_1(\psi) \qquad + \qquad \ell_2(\lambda)$$

- parameters ψ and λ separate completely in the likelihood function — param cut

- MLE $\widehat{\psi}$ is consistent , also asymptotically efficient, as long as $h(\cdot; \lambda)$ has finite variance — $\widehat{\psi} = (y_{1+} - y_{2+})/(2n)$; could be non-parametric

... Parameter Symmetry

- from earlier results, want extended orthogonal parametrization $\ell = \log L$

$$\mathbb{E}_m\{-\partial^2 \ell(\psi, \lambda)/\partial\psi\partial\lambda^{\mathrm{T}}\} = 0$$

or at least at ψ_*

- we don't know the true model m, so can't check this
- the matched pairs examples are group models
- their parametrization ensures cancellation of terms

- Result 4: If the joint distribution of Y_1, Y_2 is parametrized ψ-symmetrically, and this parametrization induces anti-symmetry on the ψ-score function, then

$$\psi^* \perp_m \Lambda, \quad \mathbb{E}_m\{\partial\ell(\psi_*, \lambda)/\partial\psi\} = 0, \text{ which implies}$$

$$\widehat{\psi} \text{ is consistent} \qquad \psi_m^{\circ} = \psi_*.$$

- Result 5: a version of Result 4 for two-group problems stratified not matched

BR: Props 2.1, 2.2

... Formalization and Parameter Symmetry

- Result 4: If the joint distribution of Y_1, Y_2 is parametrized ψ-symmetrically, and this parametrization induces anti-symmetry on the ψ-score function, then

$$\psi^* \perp_m \Lambda, \quad \mathbb{E}_m\{\partial \ell(\psi_*, \lambda)/\partial \psi\} = 0, \text{ which implies}$$

$$\widehat{\psi} \text{ is consistent} \qquad \psi_m^0 = \psi_*.$$

- Example: Location family

$g_\psi \in$ location group

$$f_{Y_1}(y_1; \lambda + \psi) = f_U(y_1 - \psi; \lambda),$$
$$f_{Y_2}(y_2; \lambda - \psi) = f_U(y_2 + \psi; \lambda)$$

- Example: Scale family

$g_\psi \in$ scale group

$$f_{Y_1}(y_1; \lambda\psi) = f_U(y_1/\psi; \lambda)(1/\psi),$$
$$f_{Y_2}(y_2; \lambda/\psi) = f_U(y_2\psi; \lambda)\psi,$$
$$U, \stackrel{d}{=} g_\psi^{-1}\, Y_1 \stackrel{d}{=} g_\psi Y_2$$

Aside: definitions

- joint distribution of Y_1, Y_2 parametrized ψ-symmetrically:
  - p_1 and p_2 are measures for a transformation model on G

$$p(gy; g\lambda)d(gy) = p(y; \lambda)dy, \qquad g \in G, y \in \mathcal{Y}, \lambda \in \Lambda$$

  - **and** group action g depends only on ψ
  - **and** p_1 and p_2 are on the same λ-orbit: $\forall u \in \mathcal{Y}, p_1(gu; g\lambda)d(gu) = p_2(g^{-1}u; g^{-1}\lambda)d(g^{-1}u)$

- if this parametrization induces anti-symmetry on the ψ-score function
  - log-likelihood function

$$\ell(\psi; \lambda, y_1, y_2) = \log f_1(y_1; g_\psi \lambda) + \log f_2(y_2; g_\psi^{-1}\lambda)$$

  - as a function of u:

$$\ell(\psi; \lambda, u_1, u_2), \qquad u_1 = g_\psi^{-1} y_1, \quad u_2 = g_\psi y_2$$

  - anti-symmetry:

$$\partial\ell(\psi; \lambda, u_1, u_2)/\partial\psi = -\partial\ell(\psi; \lambda, u_2, u_1)/\partial\psi$$

then, $\widehat{\psi} \xrightarrow{p} \psi_*$

Overview

- parameter of interest ψ is well-defined
- model with nuisance parameters may be misspecified — random effects
- when can we recover the true value of ψ
- does parameter orthogonality play a role?

- yes, it does, but may be difficult to verify directly — $\mathbb{E}_m$
- models based on groups satisfy this orthogonality
- with particular parameter structure

- most natural examples seem to involve misspecified random effects — GLM disp
- another example is marginal structural model in a 'frugal parameterization'
- propensity score is the nuisance; other aspects correspond to ψ — Evans & Didelez (2024)
- E&D model has a parameter space cut, hence orthogonal

Tentative conclusions, further work

- Results above only establish consistency
- asymptotic variance is much more difficult — although estimating it might be okay

- in the matched pairs examples, nuisance parameters treated as arbitrary constants can be eliminated by transformation to conditional or marginal distributions
- effectively assuming an arbitrary (nonparametric) mixing distribution
- less efficient when the random effects model is correct

- orthogonality under assumed model $\mathbb{E}_\theta\{-\partial^2\ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}\} = 0$ — $\theta = (\psi, \lambda)$
- *m*-orthogonality under true model $\mathbb{E}_m\{-\partial^2\ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}\} = 0$
- connection to Neyman orthogonality? — decorrelated score

$$\partial\ell(\psi,\lambda)/\partial\psi - w^{\mathrm{T}}\partial\ell(\psi,\lambda)/\partial\lambda, \quad w = I_{\psi\lambda}I_{\lambda\lambda}^{-1}$$

- extension to general estimating equations important in 2-debiased ML

Chernozhukov et al 2018, Ning et al 2017, Jorgensen & Knudsen 2004

Aside: Another example

Cox & Wong 2010

- two treatment classes, also stratified by some covariates
- observations $Y_{ij1}, i = 1, \ldots, r_{j1}$; $Y_{ij2}, i = 1, \ldots, r_{j2}$
- e.g Poisson counts with rates $\gamma_j\psi_*, \gamma_j/\psi_*$
- sufficient statistics $Y_{\cdot j1} \sim \mathit{Poisson}(r_{j1}\gamma_j\psi_*), \quad Y_{\cdot j2} \sim \mathit{Poisson}(r_{j2}\gamma_j/\psi_*)$

- assume $\gamma_j \sim \mathit{Gamma}(\alpha, \mu)$ $\lambda = (\alpha, \nu)$
- $\widehat{\psi} \xrightarrow{d} \psi_*$ under any random effects distribution with the same mean μ

- e.g. exponential, Weibull, ?log-normal?

- no group structure here because of discrete distribution
- proof uses linearity of log-likelihood in sufficient statistics Result 3

Conclusion

What can go wrong?

- the normal and/or χ^2 approximations might be poor — e.g. likelihood skewed
- too many parameters — $p \sim n^\alpha, \quad p/n \to C, \quad p/n \to \infty$
- irregular parameter space — $pf(y;\theta_1) + (1-p)f(y;\theta_2), \quad 0 \le p \le 1$
- computational intractability — $L(\theta,\tau;y) = \int_{\mathbb{R}^k} f(y \mid z;\theta) f(z;\tau) dz$
- model is misspecified — true $Y \sim m(y), \quad f(\cdot;\theta) \neq m(\cdot)\ \forall\theta$

What else can go wrong?

• the normal and/or χ^2 approximations might be poor	more accurate approximations	HOA
• too many parameters	new asymptotic theory ($p \sim n$) regularization ($p > n$)	Sur & Candès 19; Zhao et al 22 Lasso, SCAD, MCP
• irregular parameter space	different asymptotic theory, e.g. $\chi^2_D \to \Sigma\lambda_j\chi^2_{1j}$	Battey & McCullagh 24
• computational intractability	composite likelihood	Genton et al 15

... and so much more!

Examples

- climate change

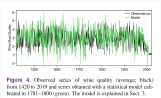
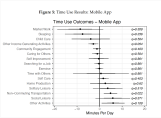


Figure 4. Observed series of wine quality (average; black) from 1420 to 2019 and series obtained with a statistical model calibrated in 1781–1800 (green). The model is explained in Sect. 3.

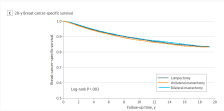
linear regression; time series

- guaranteed income



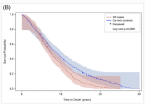
linear regression; treatment effect

- breast cancer survival



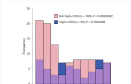
nonparametric methods

- Alzheimer's survival



proportional hazards model with random effects

- Xray jets



Poisson distribution

Thank you

UCLA
Statistics & Data Science

Department of Biostatistics

References i

Battey, H.S. & Cox, D.R. (2020). High dimensional nuisance parameters: an example from parametric survival analysis. *Information Geometry* **3** 119–148. matched pairs exp

Battey, H.S. & McCullagh, P. (2024). An anomaly arising in the analysis of processes with more than one source of variability. *Biometrika* **111**, 677–689.

Battey, H.S. & Reid, N. (2024). On the role of parametrization in models with a misspecified nuisance component. arxiv *PNAS*, to appear.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B **36**, 192–236.

Chernozhukov, V. et al. (2018). Double debiased machine learning. *Econometrics Journal* **21**, C1–C68. doi: 10.1111/ectj.12097 Neyman orthogonality

Cox, D.R. (1961). Tests of separate families of hypotheses. *4th Berkeley Symposium* **1**, 105–123.

References ii

Cox, D.R. (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B **24**, 406–424.

Cox, D.R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5**, 169–174.

Cox, D.R. and Donnelly, C.A. (2011). *Principles of Applied Statistics*. Cambridge University Press.

Davison, A.C. (2003). *Statistical Models*. Cambridge University Press.

Freedman, D.A. (2005). *Statistical Models*. Cambridge University Press.

Genton, M.G., Padoan, S.A. and Sang, H. (2015). Multivariate max-stable processes. *Biometrika* **102**, 215–230.

Hector, E. (2023). Fused mean structure learning in data integration with dependence. *Canad. J. Statist.*

Huber, P.J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions. *5th Berkeley Symposium* **1**, 221–233.

References iii

Jorgensen, B. & Knudsen, S.J. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scand. J. Statist.* **31**, 93–114.

Lehmann, E.L. (1990). Model specification: the views of Fisher and Neyman, and later developments. *Statist. Sci.* **5**, 160–168.

McCullagh, P. (2002). What is a statistical model? (with discussion). *Ann. Statist.* **30**, 1225–1310.

Ning, Y. & Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high-dimensional models. *Annals of Statistics* **45**, 158–195. decorrelated score

Qu, A.m Lindsay, B.G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836.

Sartori, N., Severini, T.A. & Marras, E. (2010). An alternative specification of generalized linear mixed models. *Comp. Stat. Data. Anal.* **54**, 575–584. https://arxiv.org/abs/2402.05708

References iv

Vansteelandt, S. & Dukes, O. (2022). Assumption-lean inference for generalized linear models (with discussion). *J. R. Statist. Soc. B* **84**, 657–685. focus on estimand

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Zhao, Q., Sur, P. and Candès, E. (2022). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli* **28**, 1835–1861.