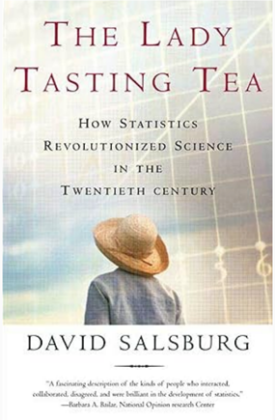
When likelihood goes wrong

Nancy Reid
University of Toronto

October 2 2024

Pfizer Colloquium

UCONN UNIVERSITY OF CONNECTICUT



"by the end of the twentieth century, almost all of science had shifted to using statistical models"

Salsburg, 2001

Outline

1. Models and science: a haphazard selection
2. Models and parameters
3. What's special about likelihood?
4. Some approaches to misspecification
5. Example
6. Conclusion

Models and science: a haphazard selection

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]

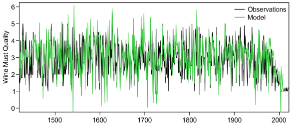


Figure 4. Observed series of wine quality (average; black) from 1420 to 2019 and series obtained with a statistical model calibrated in 1781–1800 (green). The model is explained in Sect. 3.

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

Climate of the Past
Open Access

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]

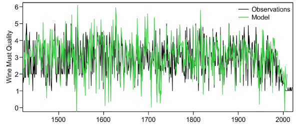


Figure 4. Observed series of wine quality (average; black) from 1420 to 2019 and series obtained with a statistical model calibrated in 1781–1800 (green). The model is explained in Sect. 3.

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

observational data

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]

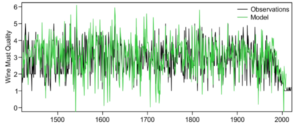


Figure 4. Observed series of wine quality (average; black) from 1420 to 2019 and series obtained with a statistical model calibrated in 1781–1800 (green). The model is explained in Sect. 3.

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

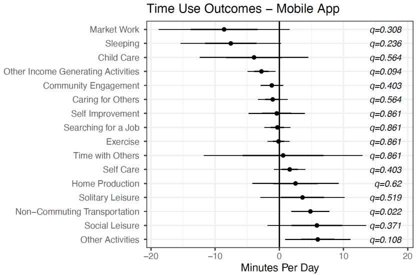
NBER WORKING PAPER SERIES

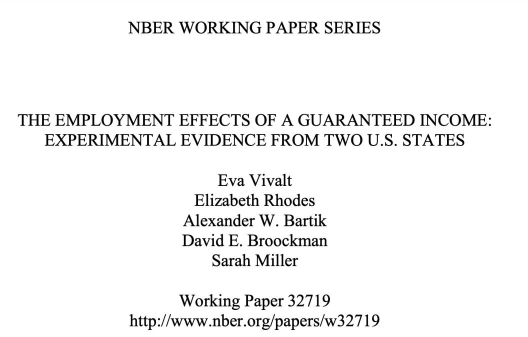
THE EMPLOYMENT EFFECTS OF A GUARANTEED INCOME: EXPERIMENTAL EVIDENCE FROM TWO U.S. STATES

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



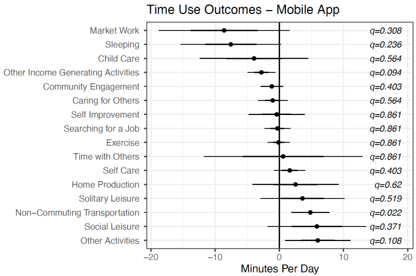

NBER WORKING PAPER SERIES

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 questions: Does guaranteed income supplement affect labor market measures?

randomized controlled trial

Income supplements

Vivalt et al 2024

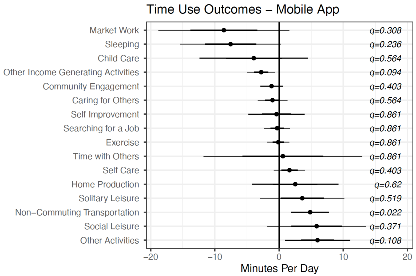
NBER WORKING PAPER SERIES

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 questions: 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.

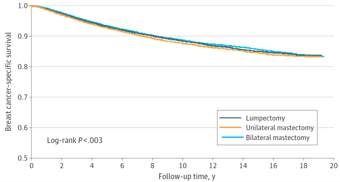
DESIGN, SETTINGS, AND PARTICIPANTS This cohort study used the Surveillance, Epidemiology, and End Results (SEER) Program registry database to identify women with unilateral breast cancer (invasive and ductal carcinoma in situ) who were diagnosed from 2000 to 2019. Three closely matched cohorts of equal 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

C 20-y Breast cancer-specific survival



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

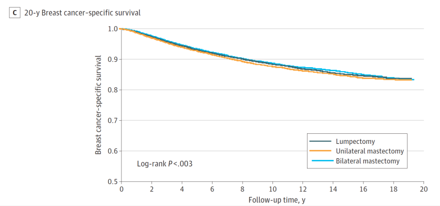
Editorial
Supplemental content

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).



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

matched case-cohort study

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.

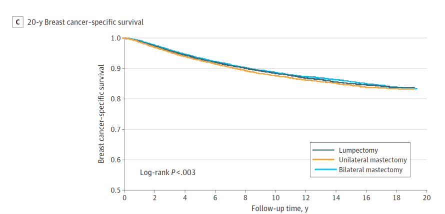
OBJECTIVES To estimate the 20-year cumulative risk of breast cancer mortality among women with stage 0 to stage III unilateral breast cancer according to the type of initial surgery performed.

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”

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[5], 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]

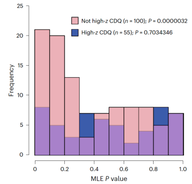


Fig. 3 | Histogram of the single-region P values from the directional test, not adjusted for multiple comparison. In pink, the subset of sources that

Yanbo Tang

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[5], 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]

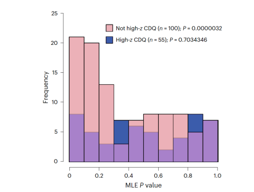


Fig. 3 | Histogram of the single-region P values from the directional test, not adjusted for multiple comparison. In pink, the subset of sources that

Yanbo Tang

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

observational data

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[5], 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]

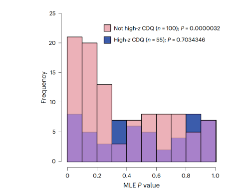


Fig. 3 | Histogram of the single-region P values from the directional test, not adjusted for multiple comparison. In pink, the subset of sources that

Yanbo Tang

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

observational data

Statistical model: compare background and source counts using Poisson distribution:

$$x_i \sim Po(a_i\beta_i), \quad y_i \sim Po(b_i\beta_i + b_i f_i \mu_i), \quad H_0 : \mu_i \equiv 0$$

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[5], 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]

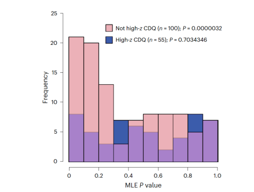


Fig. 3 | Histogram of the single-region *P* values from the directional test, not adjusted for multiple comparison. In pink, the subset of sources that

Yanbo Tang

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

observational data

Statistical model: compare background and source counts using Poisson distribution:

$$x_i \sim Po(a_i\beta_i), \quad y_i \sim Po(b_i\beta_i + b_if_i\mu_i), \quad H_0 : \mu_i \equiv 0$$

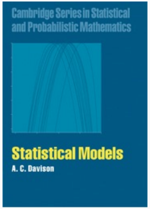
"variability in the *X*-ray emission is not compatible with current theory"

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
  - ...
- motivated by theory: economic, physical, ... — *X*-ray jets

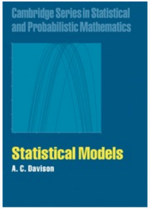
Some guidance from the experts

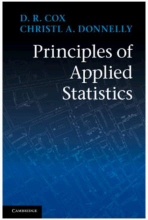


- the key feature of a statistical model is that variability is represented using probability distributions

- the art of modelling lies in finding a balance that enables the questions at hand to be answered or new ones posed

Some guidance from the experts





- 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

Statistical Science
1990, Vol. 5, No. 2, 160–168

Model Specification: The Views of Fisher and Neyman, and Later Developments

E. L. Lehmann

Role of Models in Statistical Analysis

D. R. Cox

Abstract. A number of distinct roles are identified for probability models used in the analysis of data. Examples are outlined. Some general issues arising in the formulation of such models are discussed.

empirical, or predictive models, contrasted with explanatory models

indirect models

Statistical Science
1990, Vol. 5, No. 2, 160–168

Model Specification: The Views of Fisher and Neyman, and Later Developments

E. L. Lehmann

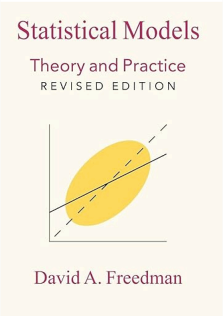
Role of Models in Statistical Analysis

D. R. Cox

Abstract. A number of distinct roles are identified for probability models used in the analysis of data. Examples are outlined. Some general issues arising in the formulation of such models are discussed.

empirical, or predictive models, contrasted with explanatory models

indirect models



The emphasis throughout is on the connection
– or lack of connection –
between the models and the real phenomena.

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

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 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

- this sounds simpler than it is

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

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

"Probability theory alone is insufficient to describe statistical methods" – Salsburg

What’s special about likelihood?

The likelihood function

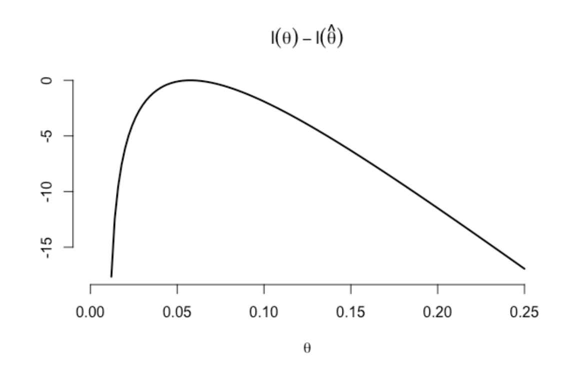
- it puts the emphasis on the model: $L(\theta; y) \propto f(y; \theta)$ — inverse problem
- provides a convenient way to compare parameter values — e.g. $L(\theta)/L(\hat{\theta})$

The likelihood function

- it puts the emphasis on the model: $L(\theta; y) \propto f(y; \theta)$
- provides a convenient way to compare parameter values

inverse problem

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

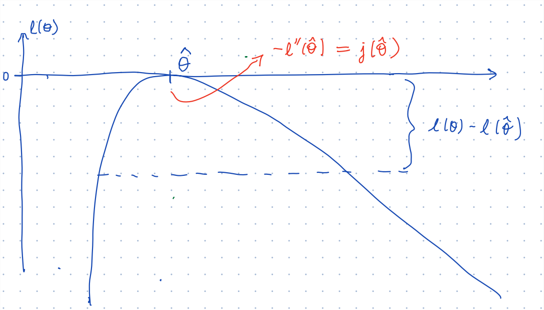


Pfizer vaccine

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

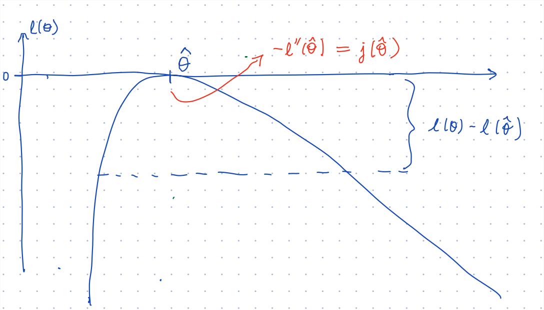
The likelihood function

- it puts the emphasis on the model: $L(\theta; y) \propto f(y; \theta)$ — inverse problem
- provides a convenient way to compare parameter values — e.g. $L(\theta)/L(\hat{\theta})$
- provides reliable summary measures



The likelihood function

- it puts the emphasis on the model: $L(\theta; y) \propto f(y; \theta)$ — inverse problem
- provides a convenient way to compare parameter values — e.g. $L(\theta)/L(\hat{\theta})$
- provides reliable summary measures



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

Inference

- Data and model: $y = y_1, \ldots, y_n$ independent; $\quad Y \sim f(y; \theta), \quad \theta \in \mathbb{R}^p$ $\qquad y \in \mathbb{R}^n$
- Likelihood function: $L(\theta; y) \propto f(y; \theta) = \prod_{i=1}^n f(y_i; \theta)$
- Log-likelihood function: $\ell(\theta; y) = \log L(\theta; y) = \sum_{i=1}^n \log f(y_i; \theta)$

Inference

- Data and model: $y = y_1, \ldots, y_n$ independent; $\quad Y \sim f(y; \theta), \quad \theta \in \mathbb{R}^p$ $\qquad y \in \mathbb{R}^n$

- Likelihood function: $L(\theta; y) \propto f(y; \theta) = \prod_{i=1}^n f(y_i; \theta)$

- Log-likelihood function: $\ell(\theta; y) = \log L(\theta; y) = \sum_{i=1}^n \log f(y_i; \theta)$

- Maximum likelihood estimator: $\hat{\theta} = \hat{\theta}(y) = \arg\sup_\theta \ell(\theta; y)$ $\qquad \ell'(\hat{\theta}) = 0$

- Score function $\ell'(\theta; y) = \Sigma_{i=1}^n \partial \ell(\theta; y_i)/\partial \theta^{\mathrm{T}}$ $\quad i(\theta) = \mathrm{var}\{\partial \ell(\theta; Y)/\partial \theta^{\mathrm{T}}\}$ $\qquad$ CLT

- observed Fisher information $\qquad j(\theta) = -\partial^2 \ell(\theta; y)/\partial \theta \partial \theta^{\mathrm{T}}$ $\qquad \mathrm{E}\{j(\theta)\} = i(\theta)$

Inference

- Data and model: $y = y_1, \ldots, y_n$ independent; $\quad Y \sim f(y;\theta), \quad \theta \in \mathbb{R}^p$ $\qquad y \in \mathbb{R}^n$
- Likelihood function: $L(\theta; y) \propto f(y;\theta) = \prod_{i=1}^n f(y_i;\theta)$
- Log-likelihood function: $\ell(\theta; y) = \log L(\theta; y) = \sum_{i=1}^n \log f(y_i;\theta)$
- Maximum likelihood estimator: $\hat\theta = \hat\theta(y) = \arg\sup_\theta \ell(\theta; y)$ $\qquad \ell'(\hat\theta) = 0$
- Score function $\ell'(\theta; y) = \Sigma_{i=1}^n \partial\ell(\theta; y_i)/\partial\theta^{\mathrm{T}}$ $\quad i(\theta) = \mathrm{var}\{\partial\ell(\theta; Y)/\partial\theta^{\mathrm{T}}\}$ $\qquad$ CLT
- observed Fisher information $\qquad j(\theta) = -\partial^2\ell(\theta; y)/\partial\theta\partial\theta^{\mathrm{T}}$ $\qquad \mathrm{E}\{j(\theta)\} = i(\theta)$
- in “large samples”

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

Too much notation!

CLT:

$$\frac{1}{\sqrt{n}}\ell'(\theta) \xrightarrow{d} N\{0, i_1(\theta)\}$$

large-sample approx:

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

Too much notation!

CLT:

$$\frac{1}{\sqrt{n}}\ell'(\theta) \xrightarrow{d} N\{0, i_1(\theta)\}$$

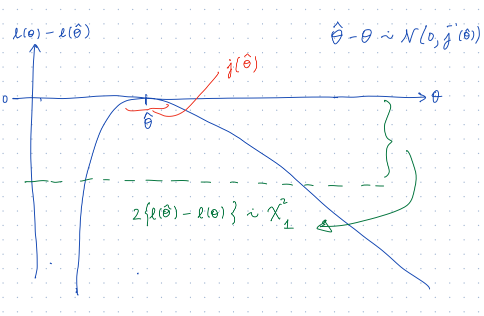
large-sample approx:

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

```
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
```



Aside: A bit too simple

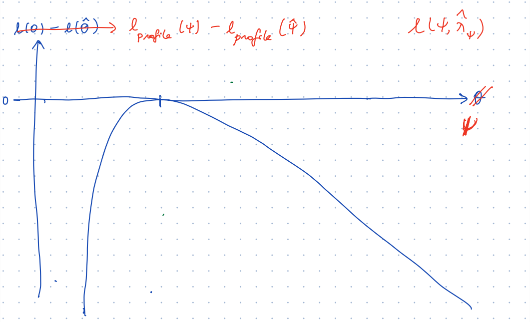
- model $f(y;\theta), \quad \theta \in \mathbb{R}^p$

Aside: A bit too simple

- model $f(y;\theta)$, $\quad \theta \in \mathbb{R}^p$
- $\theta = (\psi, \lambda)$ parameters of interest nuisance parameters

Aside: 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 $\quad \ell_{\text{profile}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$



What can go wrong?

- too many parameters

What can go wrong?

- too many parameters $\qquad p \sim n, \quad p/n \to C, \quad p/n \to \infty$

What can go wrong?

- too many parameters $\qquad p \sim n, \quad p/n \to C, \quad p/n \to \infty$

- weird parameter space

What can go wrong?

- too many parameters $\quad p \sim n, \quad p/n \to C, \quad p/n \to \infty$

- weird parameter space $\quad pf(y;\theta_1) + (1-p)f(y;\theta_2), \quad 0 \le p \le 1$

What can go wrong?

- too many parameters $\quad p \sim n, \quad p/n \to c, \quad p/n \to \infty$
- weird parameter space $\quad pf(y;\theta_1) + (1-p)f(y;\theta_2), \quad 0 \le p \le 1$
- computational intractability

What can go wrong?

- too many parameters $p \sim n, \quad p/n \to C, \quad p/n \to \infty$
- weird parameter space $pf(y;\theta_1) + (1-p)f(y;\theta_2), \quad 0 \leq p \leq 1$
- computational intractability $L(\theta,\tau;y) = \int_{\mathbb{R}^k} f(y \mid z;\theta) f(z;\tau) dz$

What can go wrong?

- too many parameters $\quad p \sim n, \quad p/n \to c, \quad p/n \to \infty$
- weird parameter space $\quad pf(y;\theta_1) + (1-p)f(y;\theta_2), \quad 0 \le p \le 1$
- computational intractability $\quad L(\theta,\tau;y) = \int_{\mathbb{R}^k} f(y \mid z;\theta)f(z;\tau)dz$
- $\boxed{\text{model is misspecified}}$

What can go wrong?

- too many parameters

$$p \sim n, \quad p/n \to C, \quad p/n \to \infty$$

- weird parameter space

$$pf(y;\theta_1) + (1-p)f(y;\theta_2), \quad 0 \le p \le 1$$

- computational intractability

$$L(\theta,\tau;y) = \int_{\mathbb{R}^k} f(y \mid z;\theta) f(z;\tau) dz$$

- $\boxed{\text{model is misspecified}}$

$$\boxed{\text{true } Y \sim m(y), \quad \nexists \theta_* \text{ w. } f(\cdot;\theta_*) = m(\cdot)}$$

What can go wrong?

- too many parameters

$$p \sim n, \quad p/n \to C, \quad p/n \to \infty$$

- weird parameter space

$$pf(y;\theta_1) + (1-p)f(y;\theta_2), \quad 0 \le p \le 1$$

- computational intractability

$$L(\theta,\tau;y) = \int_{\mathbb{R}^k} f(y \mid z;\theta) f(z;\tau) dz$$

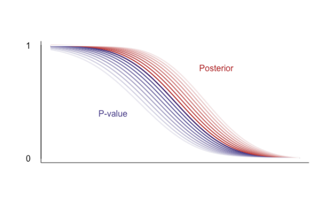
- $\boxed{\text{model is misspecified}}$

$$\boxed{\text{true } Y \sim m(y), \quad \nexists \theta_* \text{ w. } f(\cdot;\theta_*) = m(\cdot)}$$

- likelihood is not a probability

What can go wrong?

- too many parameters — $p \sim n, \quad p/n \to C, \quad p/n \to \infty$
- weird 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 $Y \sim m(y), \quad \nexists\theta_*$ w. $f(\cdot;\theta_*) = m(\cdot)$
- likelihood is not a probability



What can go wrong?

- too many parameters $p \sim n, \quad p/n \to C, \quad p/n \to \infty$
- weird 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 $\boxed{\text{true } Y \sim m(y), \quad \nexists \theta_* \text{ w. } f(\cdot;\theta_*) = m(\cdot)}$
- likelihood is not a probability

Some approaches to misspecification

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

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

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

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

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

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

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

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

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

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

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

KL-divergence

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

1. Classical

Cox 1961,2; Huber 1967; White 1982

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

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

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

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

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

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

KL-divergence

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

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

"sandwich variance"

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

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

1. Classical

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

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

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

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

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

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

KL-divergence

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

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

"sandwich variance"

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

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

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

- true model $m(\boldsymbol{y})$ fitted model $f(\boldsymbol{y};\theta)$ $\boldsymbol{y} = (y_1,\dots,y_n)$

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

- maximum likelihood estimator $\widehat{\theta}$ $\widehat{\theta} \equiv \arg\sup_\theta \ell(\theta;\boldsymbol{y})$

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

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

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

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

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

- 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

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 $\boldsymbol{y} = (\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)$

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

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 $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$

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

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

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

interpretation of θ

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

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

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

full distribution unspecified

$$\sum_{i=1}^{n} \frac{\partial\mu_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \frac{(y_i - \mu_i)}{V(\mu_i)} = \mathbf{0}$$

column vector

… More flexible inference functions

Quasi-likelihood and generalized estimating equations

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

- estimating equation for β

full distribution unspecified

$$\sum_{i=1}^{n} \frac{\partial \mu_i(\beta)}{\partial \beta} \frac{(y_i - \mu_i)}{V(\mu_i)} = \mathbf{0}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

- Example: semi-parametric regression

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

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

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

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

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

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

- Example: semi-parametric regression

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

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

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

- Possible model $\mathrm{E}\,(y \mid T, x) = \psi T + \omega(x)$ binary treatment T

- Define an estimand of interest limit of E-estimator, Robins et al 92

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

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

- Possible model $\mathrm{E}\,(y \mid T, x) = \psi T + \omega(x)$ binary treatment T

- Define an estimand of interest limit of E-estimator, Robins et al 92

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

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

- reduces to ψ under this model
- is a meaningful quantity when the model is incorrect e.g. interaction between T and x

- Possible model $\mathrm{E}\,(y \mid T, x) = \psi T + \omega(x)$ binary treatment T

- Define an estimand of interest limit of E-estimator, Robins et al 92

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

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

- reduces to ψ under this model
- is a meaningful quantity when the model is incorrect e.g. interaction between T and x

- linear model $\longrightarrow$ generalized linear model $g\{\mathrm{E}(y \mid T, x)\} = \psi T + \omega(x)$

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

- Possible model $\mathrm{E}\,(y \mid T, x) = \psi T + \omega(x)$ binary treatment T

- Define an estimand of interest limit of E-estimator, Robins et al 92

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

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

- reduces to ψ under this model
- is a meaningful quantity when the model is incorrect e.g. interaction between T and x

- linear model $\longrightarrow$ generalized linear model $g\{\mathrm{E}(y \mid T, x)\} = \psi T + \omega(x)$

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

Example

Exponential Matched Pairs

- survival times for n matched pairs (y_{1i}, y_{2i})

- random assignment of pair members to treatment/control
- nuisance parameters describing the pairs $\lambda_1, \ldots, \lambda_n$
- parameter of interest is the treatment effect

Exponential Matched Pairs

- survival times for n matched pairs (y_{1i}, y_{2i})

- random assignment of pair members to treatment/control
- nuisance parameters describing the pairs $\lambda_1, \ldots, \lambda_n$
- parameter of interest is the treatment effect

- model y_{1i} exponential with rate λ_i/ψ
  y_{2i} exponential with rate $\lambda_i \psi$
- ψ common parameter of interest; λ_i pair-specific nuisance parameters

Exponential Matched Pairs

- survival times for n matched pairs (y_{1i}, y_{2i})

- random assignment of pair members to treatment/control
- nuisance parameters describing the pairs $\lambda_1, \ldots, \lambda_n$
- parameter of interest is the treatment effect

- model y_{1i} exponential with rate λ_i/ψ
  y_{2i} exponential with rate $\lambda_i\,\psi$
- ψ common parameter of interest; λ_i pair-specific nuisance parameters

- consider modelling nuisance parameters λ_i as random effects

- model y_{1i} exponential with rate λ_i/ψ
  y_{2i} exponential with rate $\lambda_i\,\psi$
- random effects model: $\lambda_i \sim \text{Gamma}(\alpha, \beta)$ shape, rate
- integrated likelihood

$$L(\psi, \alpha, \beta; \boldsymbol{y}) = \int f(\boldsymbol{y}; \psi, \boldsymbol{\lambda}) g(\boldsymbol{\lambda}; \alpha, \beta) d\boldsymbol{\lambda}$$

- model y_{1i} exponential with rate λ_i/ψ
  y_{2i} exponential with rate $\lambda_i\,\psi$
- random effects model: $\lambda_i \sim \text{Gamma}(\alpha, \beta)$ shape, rate
- integrated likelihood

$$L(\psi, \alpha, \beta; \boldsymbol{y}) = \int f(\boldsymbol{y}; \psi, \boldsymbol{\lambda}) g(\boldsymbol{\lambda}; \alpha, \beta) d\boldsymbol{\lambda}$$

- in the integrated model, ψ is orthogonal to (α, β) $\hat{\psi}$ asy. ind't of $(\hat{\alpha}, \hat{\beta})$

... Exponential matched pairs

- model y_{1i} exponential with rate λ_i/ψ
  y_{2i} exponential with rate $\lambda_i\,\psi$

- random effects model: $\lambda_i \sim \text{Gamma}(\alpha, \beta)$ shape, rate

- integrated likelihood

$$L(\psi, \alpha, \beta; \boldsymbol{y}) = \int f(\boldsymbol{y}; \psi, \boldsymbol{\lambda}) g(\boldsymbol{\lambda}; \alpha, \beta) d\boldsymbol{\lambda}$$

- in the integrated model, ψ is orthogonal to (α, β) $\hat{\psi}$ asy. ind't of $(\hat{\alpha}, \hat{\beta})$
- even better: this is the case for any random effects distribution not just Gamma

- model y_{1i} exponential with rate λ_i/ψ
  y_{2i} exponential with rate $\lambda_i\,\psi$

- random effects model: $\lambda_i \sim \text{Gamma}(\alpha, \beta)$ shape, rate

- integrated likelihood

$$L(\psi, \alpha, \beta; \boldsymbol{y}) = \int f(\boldsymbol{y}; \psi, \boldsymbol{\lambda}) g(\boldsymbol{\lambda}; \alpha, \beta) d\boldsymbol{\lambda}$$

- in the integrated model, ψ is orthogonal to (α, β) $\hat{\psi}$ asy. ind't of $(\hat{\alpha}, \hat{\beta})$
- even better: this is the case for any random effects distribution not just Gamma

- conclude MLE $\widehat{\psi} \xrightarrow{p} \psi$ even if random effects model is misspecified

could be inefficient

... Exponential matched pairs

- model y_{1i} exponential with rate λ_i/ψ
  y_{2i} exponential with rate $\lambda_i\,\psi$

- random effects model: $\lambda_i \sim \text{Gamma}(\alpha, \beta)$ — shape, rate

- integrated likelihood

$$L(\psi, \alpha, \beta; \boldsymbol{y}) = \int f(\boldsymbol{y}; \psi, \boldsymbol{\lambda}) g(\boldsymbol{\lambda}; \alpha, \beta) d\boldsymbol{\lambda}$$

- in the integrated model, ψ is orthogonal to (α, β) — $\hat{\psi}$ asy. ind't of $(\hat{\alpha}, \hat{\beta})$
- even better: this is the case for any random effects distribution — not just Gamma

- conclude MLE $\widehat{\psi} \xrightarrow{p} \psi$ even if random effects model is misspecified

could be inefficient

can this be generalized?

1. for pair (y_{1i}, y_{2i})

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

2. orthog

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

3.

$$\mathrm{E}_m \left\{ \frac{y_{2i} - y_{1i}/\psi^2}{y_{2i}\psi + y_{1i}/\psi + \beta} \right\} = 0$$

4. for Gamma random effects, but also for any random effects distribution

interpretation of α, β when not Gamma

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

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
- 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 $m(\cdot)$ misspecified

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
- 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 $m(\cdot)$ misspecified
- Does $\widehat{\psi} \xrightarrow{p} \psi_*$?

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
- maximum likelihood estimates $(\hat{\psi},\hat{\lambda}) \xrightarrow{p} (\psi_m^0,\lambda_m^0)$ KL divergence
- assume no value of $\lambda \in \Lambda$ gives back $m(\cdot)$ misspecified
- Does $\widehat{\psi} \xrightarrow{p} \psi_*$? need $\mathrm{E}_m\{\partial\ell(\psi_*,\lambda_m^0)/\partial\psi\} = 0$ (1) λ_m^0 unknown

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
- maximum likelihood estimates $(\hat{\psi}, \hat{\lambda}) \xrightarrow{p} (\psi_m^0, \lambda_m^0)$ KL divergence
- assume no value of $\lambda \in \Lambda$ gives back $m(\cdot)$ misspecified
- Does $\widehat{\psi} \xrightarrow{p} \psi_*$? need $\mathrm{E}_m\{\partial \ell(\psi_*, \lambda_m^0)/\partial \psi\} = 0$ (1) λ_m^0 unknown
- can be easier to show $\mathrm{E}_m\{\partial \ell(\psi_*, \lambda)/\partial \psi\} = 0 \quad \forall \lambda$ (2)

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
- maximum likelihood estimates $(\hat{\psi},\hat{\lambda}) \xrightarrow{p} (\psi_m^0,\lambda_m^0)$ KL divergence
- assume no value of $\lambda \in \Lambda$ gives back $m(\cdot)$ misspecified
- Does $\widehat{\psi} \xrightarrow{p} \psi_*$? need $\mathrm{E}_m\{\partial\ell(\psi_*,\lambda_m^0)/\partial\psi\} = 0$ (1) λ_m^0 unknown
- can be easier to show $\mathrm{E}_m\{\partial\ell(\psi_*,\lambda)/\partial\psi\} = 0 \quad \forall\lambda$ (2)
- **Result :** (1) $\equiv$ (2) $\iff$ ψ_* is *m*-orthogonal to Λ: $\forall\lambda \quad \mathrm{E}_m\left\{\frac{\partial^2\ell(\psi,\lambda)}{\partial\psi\partial\lambda}\right\} = 0$

... towards a formalisation

Battey & R 2024

- true model $m(\boldsymbol{y})$ with parameter ψ and true value ψ_*
- fitted model $f(\boldsymbol{y}; \psi, \lambda)$
- Result : orthogonal parameters lead to consistent estimate — m-orthogonal
- Some weaker conditions will suffice for m-orthogonality, hence consistency
- Result : if the parametrization of f has a particular symmetry, then ψ_* is m-orthogonal to nuisance, hence consistency
- Example: generalized scale model — $f(y; \sigma) = (1/\sigma) f_0(y/\sigma)$

$$f(y; \psi\lambda) = \frac{1}{\psi} f_0\left(\frac{y}{\psi}; \lambda\right), \quad f(y; \lambda/\psi) = \psi f_0(\psi y; \lambda)$$

Tentative conclusions, further work

- Results above only establish consistency
- asymptotic variance is much more difficult

although estimating it might be okay

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

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 $\mathrm{E}_\theta\{-\partial^2\ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}\} = 0$ — $\theta = (\psi, \lambda)$
- *m*-orthogonality under true model $\mathrm{E}_m\{-\partial^2\ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}\} = 0$
- 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

Conclusion

What can go wrong?

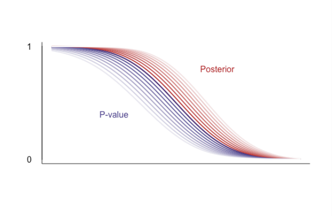
- too many parameters
- weird parameter space
- computational intractability
- $\boxed{\text{model is misspecified}}$
- likelihood is not a probability

$p \sim n, \quad p/n \to C, \quad p/n \to \infty$

$pf(y;\theta_1) + (1-p)f(y;\theta_2), \quad 0 \le p \le 1$

$L(\theta,\tau;y) = \int_{\mathbb{R}^k} f(y \mid z;\theta) f(z;\tau) dz$

$\boxed{\text{true } Y \sim m(y), \quad \nexists\theta_* \text{ w. } f(\cdot;\theta_*) = m(\cdot)}$

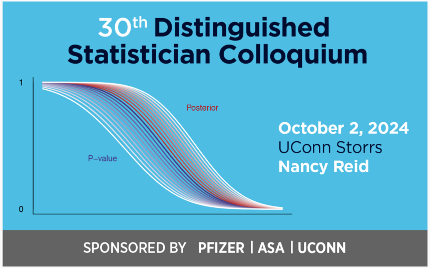


... what can go wrong?

• too many parameters	new asymptotic theory ($p \sim n$) regularization ($p > n$)	Sur & Candès 19; Zhao et al 22 Lasso, SCAD, MCP
• weird 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
• likelihood is not a probability	priors can be very influential	

... and so much more!

Thank you!


30th Distinguished Statistician Colloquium
1
0
Posterior
P–value
October 2, 2024
UConn Storrs
Nancy Reid
SPONSORED BY PFIZER | ASA | UCONN

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.*

References iii

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

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.

References iv

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

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.