


Location-adjusted Wald statistics

Ioannis Kosmidis

 IKosmidis_

 ioannis.kosmidis@warwick.ac.uk

 <http://ucl.ac.uk/~ucakiko>

Reader in Data Science

University of Warwick & The Alan Turing Institute

in collaboration with Claudia Di Caterina
University of Padova

04 May 2018

Institute for Statistics and Mathematics
WU Wien

Outline

- 1 Wald inference
- 2 Location-adjusted Wald statistic
- 3 Wald inference and bias-corrected estimators
- 4 Logistic regression with structural parameters
- 5 Brain lesions
- 6 Recap

Outline

- 1 Wald inference
- 2 Location-adjusted Wald statistic
- 3 Wald inference and bias-corrected estimators
- 4 Logistic regression with structural parameters
- 5 Brain lesions
- 6 Recap

Wald statistic for scalar parameters

Data

$$(y_i, x_i^\top) \quad (i = 1, \dots, n)$$

$x_i = (x_{i1}, \dots, x_{ik})^\top \in \mathfrak{R}^k$ is a vector of explanatory variables for y_i

Model

Independent random variables Y_1, \dots, Y_n with pdf/pmf $p_Y(y_i|x_i; \theta)$

Parameter $\theta \in \Theta \subset \mathfrak{R}^p$ with

$\theta = (\psi, \lambda^\top)^\top$, where $\psi \in \mathfrak{R}$ is of interest

Task

Draw inference about ψ

Wald statistic

Log-likelihood¹

$$l(\theta) = \sum_{i=1}^n \log p_Y(y_i | x_i; \theta)$$

Wald statistic for testing $\psi = \psi_0$

$$t = \frac{\hat{\psi} - \psi_0}{\kappa(\hat{\theta})} \underset{\text{appr}}{\sim} N(0, 1)$$

Maximum likelihood estimator (MLE)

$$\hat{\theta} = (\hat{\psi}, \hat{\lambda}^\top)^\top = \arg \max_{\theta \in \Theta} l(\theta)$$

Standard error

$\kappa(\theta)$ is the square root of the (ψ, ψ) element of the variance-covariance $\{i(\theta)\}^{-1}$ of the (asymptotic) null distribution of $\hat{\theta}$

$i(\theta)$ is typically taken to be the expected information $E\{\nabla l(\theta)\nabla l(\theta)^\top\}$ or some “robust” variant

¹subject to usual regularity conditions; see, Pace and Salvan (1997, §4.3)

Wald statistic

Asymptotically equivalent alternatives

Signed root of the likelihood ratio statistic

$$r = \text{sign}(\hat{\psi} - \psi_0) \{l(\hat{\psi}, \hat{\lambda}) - l(\psi_0, \hat{\lambda}_{\psi_0})\}^{1/2} \overset{\text{appr}}{\sim} N(0, 1)$$

Signed root of the score statistic

$$s = \text{sign}(\hat{\psi} - \psi_0) \frac{\partial l(\psi_0, \hat{\lambda}_{\psi_0})}{\partial \psi} \kappa(\psi_0, \hat{\lambda}_{\psi_0}) \overset{\text{appr}}{\sim} N(0, 1)$$

where $\hat{\lambda}_{\psi_0} = \arg \max_{\lambda} l(\psi_0, \lambda)$ is the constrained MLE for λ

Pros of t

Computational convenience

Cons of t

Inferential performance depends on the properties of $\hat{\theta}$ (bias, efficiency, etc)

z test of coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.37540	0.68957	-0.5444	0.58617
lullyes	1.43237	0.73414	1.9511	0.05105
day2	-0.11394	1.04442	-0.1091	0.91313
day3	-0.58487	1.13343	-0.5160	0.60584
day4	-1.71670	1.31233	-1.3081	0.19083
day5	1.82912	1.30168	1.4052	0.15996
day6	0.24783	0.94155	0.2632	0.79238
day7	0.94994	0.99256	0.9571	0.33854
day8	0.46505	0.96850	0.4802	0.63111
day9	0.88646	1.11872	0.7924	0.42813
day10	1.66815	1.05172	1.5861	0.11271

Lack of reparameterization invariance

Reading accuracy IQ and dyslexia

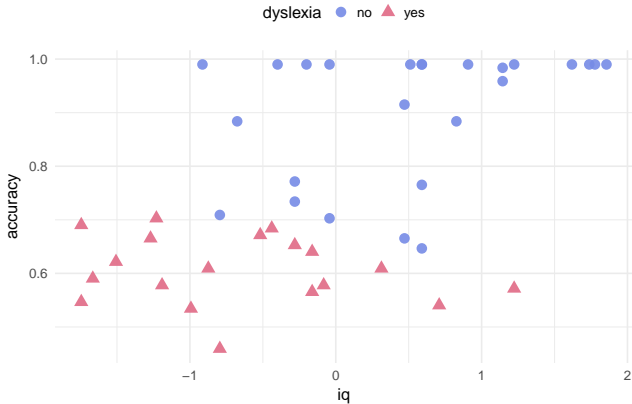
Data

Reading accuracy for 44 nondyslexic and dyslexic Australian children²
Ages between 8 years+5 months and 12 years+3 months

Variables

accuracy	the score on a reading accuracy test
iq	standardized score on a nonverbal intelligent quotient test
dyslexia	whether the child is dyslexic or not

²data from Smithson and Verkuilen (2006)



Aim

Investigate the relative contribution of nonverbal IQ to the distribution the reading scores, controlling for the presence of diagnosed dyslexia

Reading accuracy IQ and dyslexia

Model

Score of the i -th child is from a Beta distribution with mean μ_i and variance $\mu_i(1 - \mu_i)/(1 + \phi_i)$ with

$$\log \frac{\mu_i}{1 - \mu_i} = \beta_1 + \sum_{j=2}^4 \beta_j x_{ij} \quad \text{and} \quad \log \phi_i = \gamma_1 + \sum_{j=2}^3 \gamma_j x_{ij}$$

- x_{i2} takes value -1 if the i th child is dyslexic and 1 if not
- x_{i3} is the nonverbal IQ score, and
- $x_{i4} = x_{i2}x_{i3}$ is the interaction between dyslexia and iq

```
Call:
betareg(formula = accuracy ~ dyslexia * iq | dyslexia + iq, data = ReadingSkills,
         type = "ML")
```

```
Standardized weighted residuals 2:
```

	Min	1Q	Median	3Q	Max
	-2.3900	-0.6416	0.1572	0.8524	1.6446

```
Coefficients (mean model with logit link):
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	1.1232	0.1428	7.864	3.73e-15	***
dyslexia	-0.7416	0.1428	-5.195	2.04e-07	***
iq	0.4864	0.1331	3.653	0.000259	***
dyslexia:iq	-0.5813	0.1327	-4.381	1.18e-05	***

```
Phi coefficients (precision model with log link):
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	3.3044	0.2227	14.835	< 2e-16	***
dyslexia	1.7466	0.2623	6.658	2.77e-11	***
iq	1.2291	0.2672	4.600	4.23e-06	***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Type of estimator: ML (maximum likelihood)
```

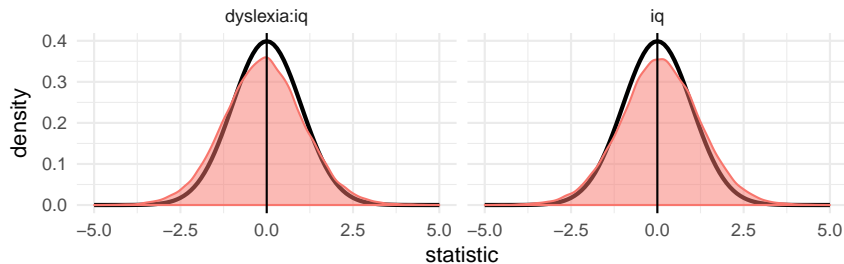
```
Log-likelihood: 65.9 on 7 Df
```

```
Pseudo R-squared: 0.5756
```

```
Number of iterations: 25 (BFGS) + 1 (Fisher scoring)
```

³see Grün, Kosmidis, and Zeileis (2012) for a range of modelling strategies and learning methods based on beta regression using the `betareg` R package

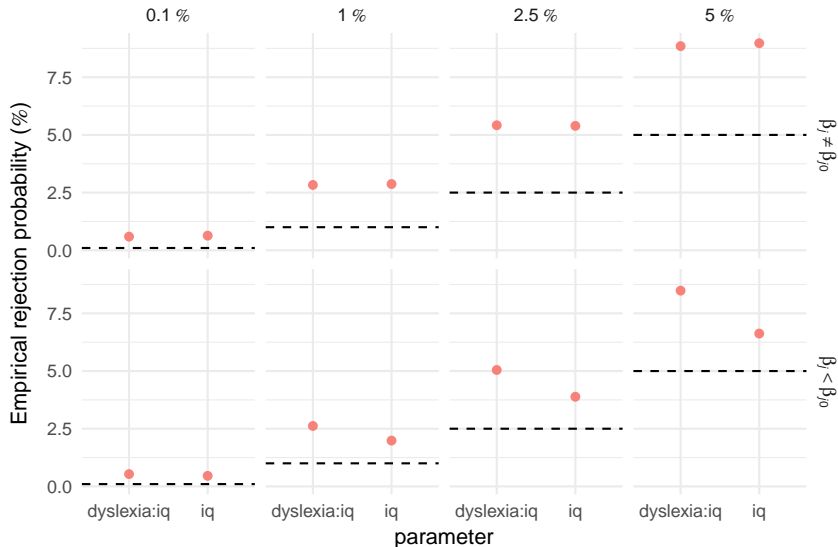
Null distribution of Wald statistic for $\beta_j = \beta_{0j}$



parameter	mean	sd
dyslexia:iq	-0.09	1.15
iq	0.08	1.15

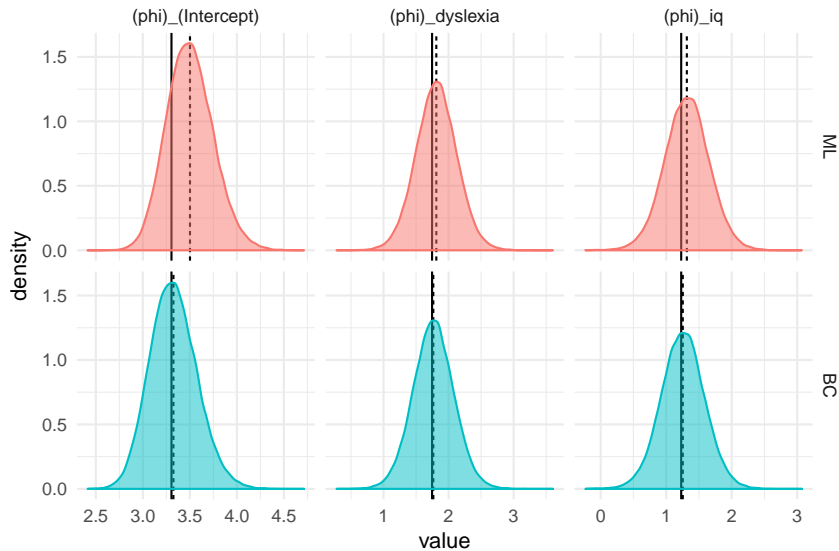
⁴figures based on 50 000 simulated samples under the maximum likelihood fit

Empirical null rejection probabilities



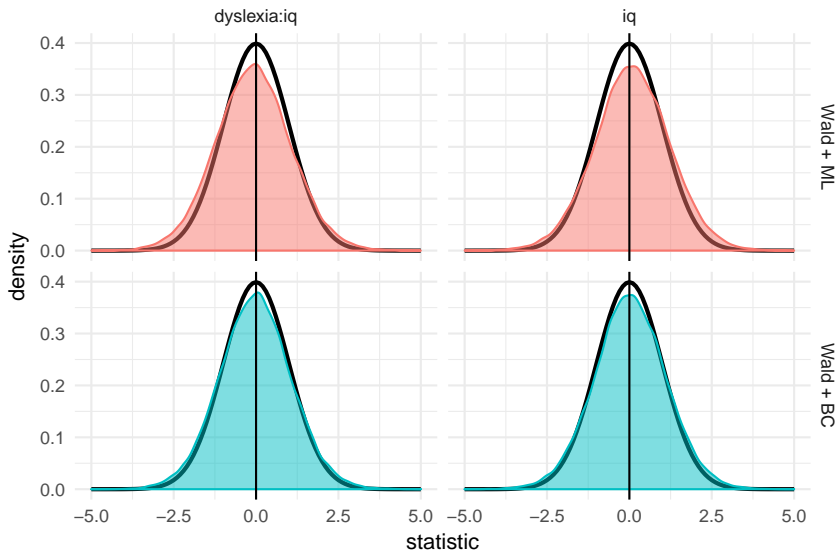
Empirical rejection probabilities are almost double the nominal level

MLE and bias corrected estimator⁵



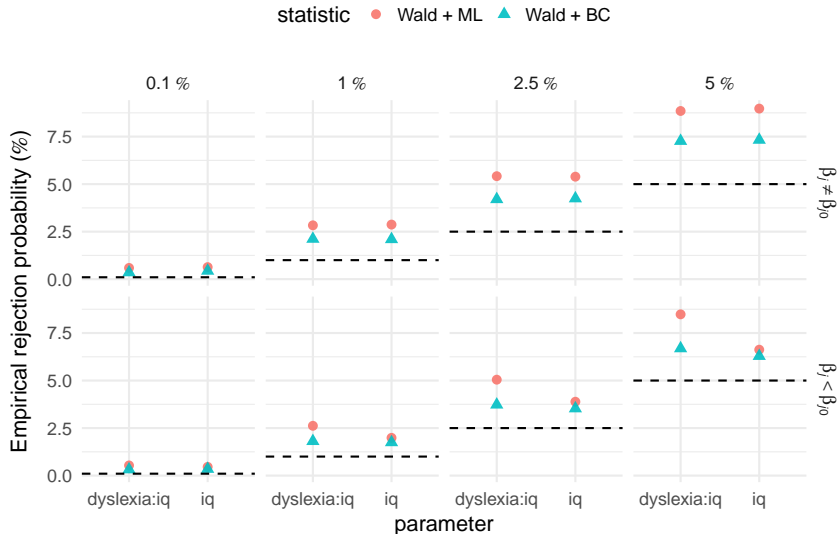
⁵with `type = BC` in the `betareg` call
see, Grün, Kosmidis, and Zeileis (2012) for details on bias correction

Null distribution of Wald statistic using BC estimators⁶



⁶Proposed in Kosmidis and Firth (2010)

Empirical null rejection probabilities



⁷figures based on 50 000 simulated samples under the maximum likelihood fit

Recap on Wald statistics with improved estimators

Use of improved estimators when forming Wald statistics can improve $N(0, 1)$ approximation and hence inferential performance⁸

But

Merely an observation, and in a few models

Rather indirect way to improving Wald inference

Better estimators in $t \not\Rightarrow$ null distribution of t closer to $N(0, 1)$

⁸see, e.g., Kosmidis and Firth (2010)

Outline

- 1 Wald inference
- 2 Location-adjusted Wald statistic**
- 3 Wald inference and bias-corrected estimators
- 4 Logistic regression with structural parameters
- 5 Brain lesions
- 6 Recap

Wald statistic as an estimator

Wald Transform

$$T(\theta; \psi_0) = \frac{\psi - \psi_0}{\kappa(\theta)}$$



The Wald statistic
 $t = T(\hat{\theta}; \psi_0)$
is the MLE of $T(\theta; \psi_0)$

Core idea

Bias reduction techniques to bring **asymptotic mean** of t “closer” to 0

Bias of t

Under regularity conditions⁹ it can be shown that

$$E\{T(\hat{\theta}; \psi_0) - T(\theta; \psi_0)\} = B(\theta; \psi_0) + O(n^{-3/2})$$

where

First-order bias of t

$$B(\theta; \psi_0) = b(\theta)^\top \nabla T(\theta; \psi_0) + \frac{1}{2} \text{trace} [\{i(\theta)\}^{-1} \nabla \nabla^\top T(\theta; \psi_0)]$$

First-order bias of $\hat{\theta}$

$b(\theta)$ such that $E(\hat{\theta} - \theta) = b(\theta) + o(n^{-1})$

⁹to guarantee that $T(\theta, \psi_0)$ is > 3 times differentiable wrt θ and $\hat{\theta}$ is consistent

Location-adjusted Wald statistic

Key result

The location-adjusted Wald statistic

$$t^* = T(\hat{\theta}; \psi_0) - B(\hat{\theta}; \psi_0)$$

has null expectation of order $O(n^{-3/2})$

Quantities in the bias of t

$i(\theta)$ and $b^u(\theta)$ are readily available for a wide range of models, including generalized linear and nonlinear models¹⁰

Gradient and Hessian of the Wald transform

$$\nabla T(\theta; \psi_0) = \{1_p - T(\theta; \psi_0)\nabla\kappa(\theta)\} / \kappa(\theta)$$

$$\nabla\nabla^\top T(\theta; \psi_0) = - \left[\nabla\kappa(\theta) \{\nabla T(\theta; \psi_0)\}^\top + \nabla T(\theta; \psi_0) \{\nabla\kappa(\theta)\}^\top + T(\theta; \psi_0)\nabla\nabla^\top\kappa(\theta) \right] / \kappa(\theta)$$

$\nabla\kappa(\theta)$ and $\nabla\nabla^\top\kappa(\theta)$ can be computed either analytically, or using automatic or numerical differentiation

¹⁰see, for example, Cook et al. (1986); Cordeiro and McCullagh (1991); Cordeiro and Vasconcellos (1997); Cordeiro and Toyama Udo (2008); Kosmidis and Firth (2009); Simas et al. (2010); Grün et al. (2012) etc Ioannis Kosmidis - Location-adjusted Wald statistics 21/45

Example: Exponential with mean $e^{-\theta}$

Cornish-Fisher expansions (Hall, 1992, § 2.5) of the α -level quantiles q_α and q_α^* of the distribution of t and t^* in terms of the corresponding standard normal quantiles z_α are

$$q_\alpha = z_\alpha + n^{-1/2} \frac{z_\alpha^2 + 2}{6} - n^{-1} \frac{11z_\alpha^3 - 65z_\alpha}{144} + O(n^{-3/2}),$$
$$q_\alpha^* = z_\alpha + n^{-1/2} \frac{z_\alpha^2 - 1}{6} - n^{-1} \frac{11z_\alpha^3 - 65z_\alpha}{144} + O(n^{-3/2}),$$

provided that $\epsilon < \alpha < 1 - \epsilon$ for any $0 < \epsilon < 1/2$

→ Quantiles of t^* are closer to those of $N(0, 1)$ than t

Computational complexity and implementation

No extra matrix inversions (beyond $\{i(\theta)\}^{-1}$) or optimisation when computing t^* ; only extra matrix multiplications

In its analytical form, t^* has the computational complexity $O(p^4)$, whence t has $O(p^3)$

Time complexity can be reduced drastically by exploiting sparsity in $i(\theta)$ in specific models and vectorising operations

Evaluation of t^* for each of the model parameters can be done post-fit and **in parallel**

Implementation with numerical derivatives of $\kappa(\theta)$

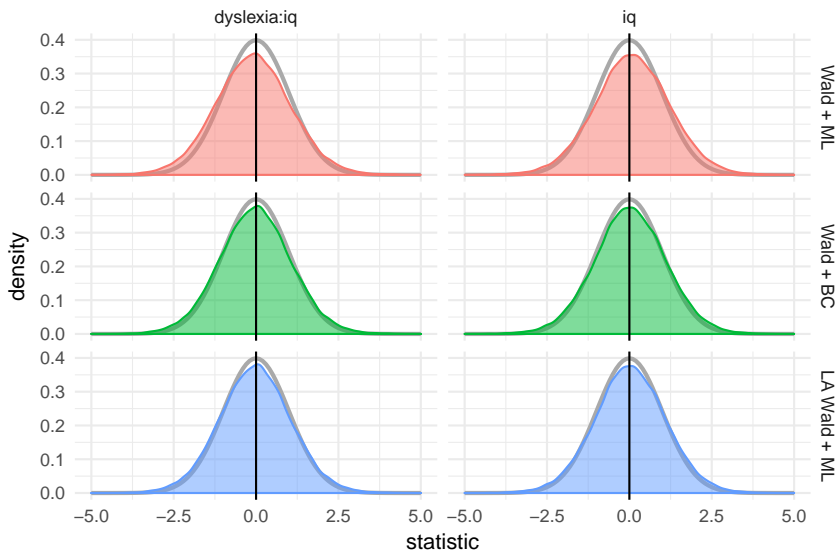
As implemented in the **waldi** R package

<https://github.com/ikosmidis/waldi>¹¹

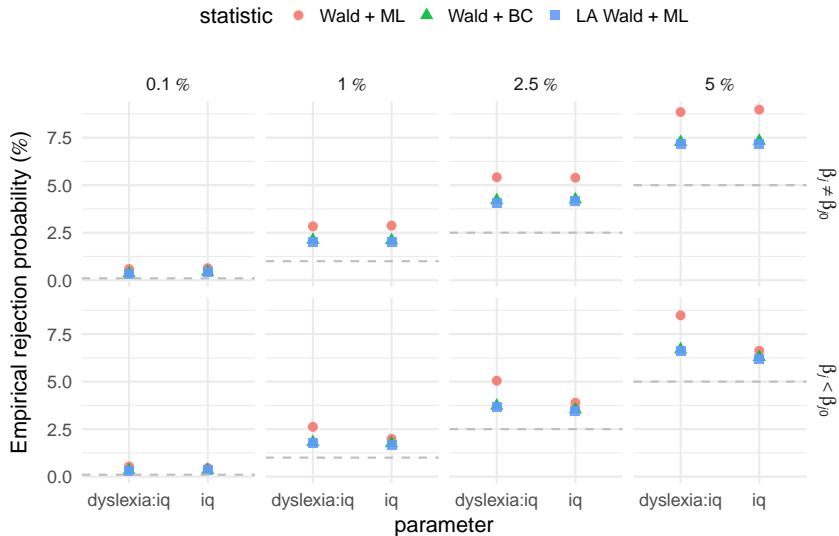
```
R> bias <- enrichwith::get_bias_function(object)
R> info <- enrichwith::get_information_function(object)
R>
R> t <- coef(summary(object))[, "z value"]
R> theta_hat <- coef(object)
R> b <- bias(theta_hat)
R> inverse_i_hat <- solve(info(theta_hat))
R>
R> kappa <- function(theta, j) {
+   inverse_i <- solve(info(theta))
+   sqrt(inverse_i[j, j])
+ }
R>
R> adjusted_t <- function(j) {
+   u <- numDeriv::grad(kappa, theta_hat, j = j)
+   V <- numDeriv::hessian(kappa, theta_hat, j = j)
+   a <- -t[j] * u
+   a[j] <- 1 + a[j]
+   t[j] - sum(a * b)/ses[j] +
+     (sum(inverse_i_hat * (tcrossprod(a, u)))/ses[j] +
+      0.5 * t[j] * sum(inverse_i_hat * V))/ses[j]
+ }
```

¹¹Using R packages `enrichwith` (Kosmidis, 2017) and `numDeriv` (Gilbert and Varadhan, 2016)

Beta regression: Reading accuracy and dyslexia



Empirical null rejection probabilities

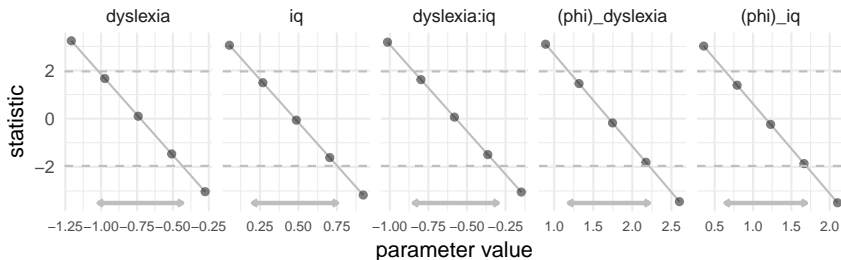


Confidence intervals based on t^*

$100(1 - \alpha)\%$ confidence intervals based on t^* can be obtained by finding all ψ such that

$$|T(\hat{\theta}; \psi) - B(\hat{\theta}; \psi)| \leq z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of $N(0, 1)$



Outline

- 1 Wald inference
- 2 Location-adjusted Wald statistic
- 3 Wald inference and bias-corrected estimators**
- 4 Logistic regression with structural parameters
- 5 Brain lesions
- 6 Recap

Wald statistics with bias-corrected estimators

$$\tilde{t} = T(\tilde{\theta}; \psi_0)$$

with $\tilde{\theta}$ being a bias-corrected estimator with

$$E(\tilde{\theta} - \theta) = o(n^{-1})$$

Bias of \tilde{t}

$E\{T(\tilde{\theta}; \psi_0) - T(\theta; \psi_0)\} = \tilde{B}(\theta; \psi_0) + o(n^{-1/2})$ with

$$\tilde{B}(\theta; \psi_0) = b(\theta)^\top \nabla T(\theta; \psi_0) + \frac{1}{2} \text{trace} [\{i(\theta)\}^{-1} \nabla \nabla^\top T(\theta; \psi_0)]$$

Use of bias-corrected estimators **eliminates a term**, but bias of Wald statistic is still $O(n^{-1/2})$

Models with categorical responses

Location-adjustment of \tilde{t} is still fruitful

Categorical response models, where bias-correction leads to estimates that are **always finite** even in case where the MLE is infinite.

Outline

- 1 Wald inference
- 2 Location-adjusted Wald statistic
- 3 Wald inference and bias-corrected estimators
- 4 Logistic regression with structural parameters**
- 5 Brain lesions
- 6 Recap

Lulling babies

Data

18 **matched pairs** of binomial observations on the effect of lulling on the crying of babies

Matching is per day and each day pair consists of the number of babies not crying out of a fixed number of control babies, and the outcome of lulling on a single child

Experiment involves 143 babies

Variables

crying crying status of the baby (1 not crying; 0 crying)

day the day of the experiment

lull has the baby been lulled?

Aim: Test the effect of lulling on the crying of children

Logistic regression: lulling babies

Model

Y_{ij} is a Bernoulli random variable for the crying status of baby j in day i with probability μ_{ij} of not crying

$$\log \frac{\mu_{ij}}{1 - \mu_{ij}} = \beta_i + \gamma z_{ij}$$

- z_{ij} is 1 if the j th child on day i was lulled, and 0 otherwise

Task

Test $\gamma = 0$ accounting for heterogeneity between days

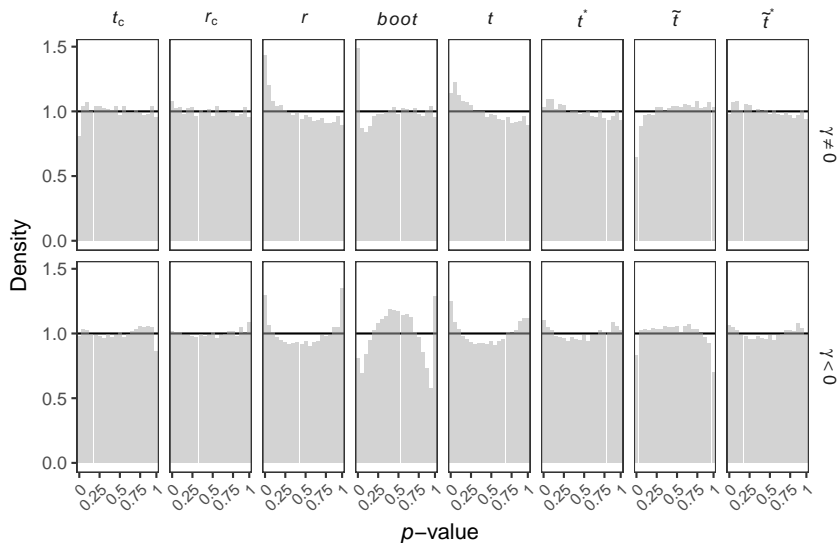
Testing for $\gamma = 0$

	t_c	r_c	r	t	t^*	\tilde{t}	\tilde{t}^*
statistic	1.8307	2.0214	2.1596	1.9511	1.9257	1.7362	1.9064
p -value	0.0671	0.0432	0.0308	0.0510	0.0541	0.0825	0.0566

t_c is the Wald statistic based on the maximum **conditional likelihood estimator**

r and r_c are the signed roots of the likelihood and conditional likelihood ratio statistics

Empirical p -value distribution

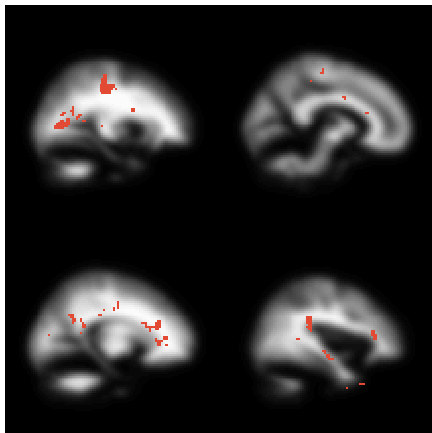


¹²based on 50 000 samples from the mdoel with $\beta_1, \dots, \beta_{18}$ set to their maximum likelihood estimates and $\gamma = 0$

Outline

- 1 Wald inference
- 2 Location-adjusted Wald statistic
- 3 Wald inference and bias-corrected estimators
- 4 Logistic regression with structural parameters
- 5 Brain lesions**
- 6 Recap

Mass univariate regression for brain lesions



resolution: $91 \times 109 \times 91$ (902 629 voxels)

Aim: Construct significance maps, highlighting voxels according to the evidence against the null hypothesis of no covariate effect

Sample

lesion maps for 50 patients¹³

Patient characteristics

multiple sclerosis type
(MS)¹⁴

age

gender

disease duration (DD)

two disease severity measures
(PASAT and EDSS)

¹⁴from the supplementary material of Ge et al. (2014)

¹⁵0 for relapsing-remitting and 1 for secondary progressive multiple sclerosis

Voxel-wise probit regressions

Lesion occurrence in voxel j for patient i

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$$

Lesion probability

$$\Phi^{-1}(\pi_{ij}) = \beta_{j0} + \beta_{j1}MS_i + \beta_{j2}age_i + \beta_{j3}gender_i + \beta_{j4}DD_i + \beta_{j5}PASAT_i + \beta_{j6}EDSS_i$$

Results

Occurrence of infinite estimates

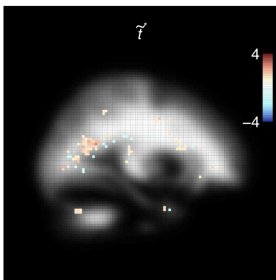
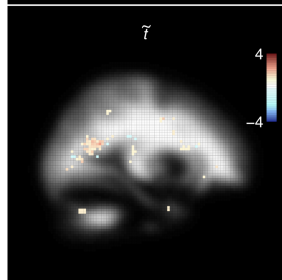
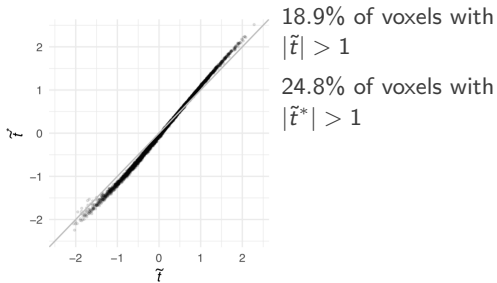
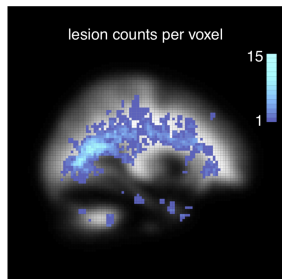
Covariate	Occurrence
MS	75.5%
age	63.7%
gender	78.3%
DD	63.7%
PASAT	63.6%
EDSS	63.2%

Failures in evaluation of r

Covariate	Occurrence
MS	19.2%
age	20.5%
sex	22.4%
DD	18.1%
PASAT	16.8%
EDSS	10.3%

¹⁵summaries based on voxels with lesion occurrence for at least one lesion across patients

Significance map for disease duration



Outline

- 1 Wald inference
- 2 Location-adjusted Wald statistic
- 3 Wald inference and bias-corrected estimators
- 4 Logistic regression with structural parameters
- 5 Brain lesions
- 6 Recap**

Recap

Location-adjustment can deliver **substantial improvements** to Wald inference

Extra computational overhead is mainly due to matrix multiplications

Location adjustment with “robust”¹⁶ variance-covariance matrices

Location adjustment with alternative estimators of estimator bias, including bootstrap and jackknife; particularly useful, e.g., for generalized linear mixed effects models

Extensions to other pivotal quantities, including Wald statistics for composite hypotheses, score statistics, or even directly p -values

¹⁶see, for example, MacKinnon and White (1985) Ioannis Kosmidis - Location-adjusted Wald statistics 42/45

References I

- Cook, R. D., C.-L. Tsai, and B. C. Wei (1986). Bias in nonlinear regression. *Biometrika* 73, 615–623.
- Cordeiro, G. and M. Toyama Udo (2008). Bias correction in generalized nonlinear models with dispersion covariates. *Communications in Statistics: Theory and Methods* 37, 2219–225.
- Cordeiro, G. M. and P. McCullagh (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B: Methodological* 53, 629–643.
- Cordeiro, G. M. and K. L. P. Vasconcellos (1997). Bias correction for a class of multivariate nonlinear regression models. *Statistics & Probability Letters* 35, 155–164.
- Ge, T., N. Müller-Lenke, K. Bendfeldt, T. E. Nichols, and T. D. Johnson (2014). Analysis of multiple sclerosis lesions via spatially varying coefficients. *Annals of Applied Statistics* 8(2), 1095–1118.
- Gilbert, P. and R. Varadhan (2016). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.
- Grün, B., I. Kosmidis, and A. Zeileis (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software* 48, 1–25.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer.
- Kosmidis, I. (2017). *enrichwith: Methods to enrich list-like R objects with extra components*. R package version 0.1.
- Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika* 96, 793–804.
- Kosmidis, I. and D. Firth (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics* 4, 1097–1112.
- MacKinnon, J. G. and H. White (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics* 29, 305–325.

References II

- Pace, L. and A. Salvani (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. London: World Scientific.
- Simas, A. B., W. Barreto-Souza, and A. V. Rocha (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis* 54, 348–366.
- Smithson, M. and J. Verkuilen (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11, 54–71.

Location-adjusted Wald statistic



$$t^* = T(\hat{\theta}; \psi_0) - B(\hat{\theta}; \psi_0)$$

Preprint


Di Caterina C and Kosmidis I (2017). Location-adjusted Wald statistic for scalar parameters. *ArXiv e-prints*. arXiv:1710.11217

Software

waldi R package¹⁷ (soon in CRAN!) for computing t^* for standard model classes, including GLMs (`glm`, `brglm2`) and beta regression (`betareg`)

 IKosmidis_

 ioannis.kosmidis@warwick.ac.uk

 <http://ucl.ac.uk/~ucakiko>



¹⁷<https://github.com/ikosmidis/waldi>