Location-adjusted Wald statistics

Ioannis Kosmidis

Ioannis.Kosmidis

ioannis.kosmidis@warwick.ac.uk

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

Reader in Data Science

University of Warwick & The Alan Turing Institute

joint work 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
Data
\((y_i, x_i^\top) \quad (i = 1, \ldots, n)\)
\(x_i = (x_{i1}, \ldots, x_{ik})^\top \in \mathbb{R}^k\) is a vector of explanatory variables for \(y_i\)

Model
Independent random variables \(Y_1, \ldots, Y_n\) with pdf/pmf \(p_Y(y_i|x_i; \theta)\)
Parameter \(\theta \in \Theta \subset \mathbb{R}^p\) with
\(\theta = (\psi, \lambda^\top)^\top\), where \(\psi \in \mathbb{R}\) is of interest

Task
Draw inference about \(\psi\)
Wald statistic

Log-likelihood\(^1\)

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

Wald statistic for testing \( \psi = \psi_0 \)

\[ t = \frac{\hat{\psi} - \psi_0}{\kappa(\hat{\theta}) \text{ aprr} \sim N(0, 1)} \]

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 \( (\psi, \psi) \) 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

\(^1\)subject 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 \( \lambda \)

**Pros of \( t \)**

Computational convenience

**Cons of \( t \)**

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

Lack of reparameterization invariance
Reading accuracy IQ and dyslexia

Data
Reading accuracy for 44 nondyslexic and dyslexic Australian children\(^2\)
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

\(^2\)data from Smithson and Verkuilen (2006)
**Aim**

Investigate the relative contribution of nonverbal IQ to the distribution of 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 $\mu_i$ 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:
```r
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)

---

3 see 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 $\beta_j = \beta_{0j}$

![Graphs showing null distribution of Wald statistic for dyslexia:iq and iq parameters.]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>dyslexia:iq</td>
<td>-0.09</td>
<td>1.15</td>
</tr>
<tr>
<td>iq</td>
<td>0.08</td>
<td>1.15</td>
</tr>
</tbody>
</table>

*Figures 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 estimator\(^5\)

\[
\begin{align*}
\text{(phi)}_{\text{(Intercept)}} & \quad \text{(phi)}_{\text{dyslexia}} & \quad \text{(phi)}_{\text{iq}} \\
\text{ML} & \quad \text{BC} \\
2.5 & \quad 3.0 & \quad 3.5 & \quad 4.0 & \quad 4.5 \\
1 & \quad 2 & \quad 3 \\
0.0 & \quad 0.5 & \quad 1.0 & \quad 1.5 \\
\end{align*}
\]

value density

\(^5\) with type = BC in the betareg call
see, Grün, Kosmidis, and Zeileis (2012) for details on bias correction
Null distribution of Wald statistic using BC estimators

Proposed in Kosmidis and Firth (2010)
Empirical null rejection probabilities

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Empirical rejection probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dyslexia:iq</td>
<td>0.1%</td>
</tr>
<tr>
<td>iq</td>
<td>1%</td>
</tr>
<tr>
<td>dyslexia:iq</td>
<td>2.5%</td>
</tr>
<tr>
<td>iq</td>
<td>5%</td>
</tr>
<tr>
<td>dyslexia:iq</td>
<td>7.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Empirical rejection probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dyslexia:iq</td>
<td>0.0%</td>
</tr>
<tr>
<td>iq</td>
<td>2.5%</td>
</tr>
<tr>
<td>dyslexia:iq</td>
<td>5.0%</td>
</tr>
<tr>
<td>iq</td>
<td>7.5%</td>
</tr>
</tbody>
</table>

statistic
- **Wald + ML**
- **Wald + BC**

7 figures 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 performance.\(^8\)

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 $N(0, 1)$

\(^8\)see, 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**

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

\[ \downarrow \]

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 conditions\textsuperscript{9} 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 $t$**

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

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

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

\textsuperscript{9}to guarantee that $T(\theta, \psi_0)$ is > 3 times differentiable wrt $\theta$ and $\hat{\theta}$ is consistent
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 $b(\theta)$ are readily available for a wide range of models, including generalized linear and nonlinear models\(^{10}\)

**Gradient and Hessian of the Wald transform**

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

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

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

\(^{10}\)see, 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
Example: Exponential with mean $e^{-\theta}$

Cornish-Fisher expansions (Hall, 1992, § 2.5) of the $\alpha$-level quantiles $q_\alpha$ and $q^*_\alpha$ of the distribution of $t$ and $t^*$ in terms of the corresponding standard normal quantiles $z_\alpha$ 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 $0 < \epsilon < 1/2$

$\rightarrow$ 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\textsuperscript{11}

\begin{verbatim}
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]
+ }
\end{verbatim}

\textsuperscript{11}Using R packages enrichwith (Kosmidis, 2017) and numDeriv (Gilbert and Varadhan, 2016)
Beta regression: Reading accuracy and dyslexia

![Graph showing distribution of dyslexia:iq and iq with different statistical methods: Wald + ML, Wald + BC, and LA Wald + ML.](image)
Empirical null rejection probabilities

![Empirical null rejection probabilities](image)

Empirical rejection probability (\%) statistic

- Wald + ML
- Wald + BC
- LA Wald + ML

---

Ioannis Kosmidis - Location-adjusted Wald statistics 26/45
Confidence intervals based on $t^*$

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

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

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

<table>
<thead>
<tr>
<th>statistic</th>
<th>dyslexia</th>
<th>iq</th>
<th>dyslexia:iq</th>
<th>(phi)_dyslexia</th>
<th>(phi)_iq</th>
</tr>
</thead>
<tbody>
<tr>
<td>parameter value</td>
<td>−1.25−1.00−0.75−0.50−0.25</td>
<td>0.25 0.50 0.75</td>
<td>−1.00 −0.75 −0.50 −0.25</td>
<td>1.0 1.5 2.0 2.5</td>
<td>0.5 1.0 1.5 2.0</td>
</tr>
</tbody>
</table>

Ioannis Kosmidis - Location-adjusted Wald statistics  27/45
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

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

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 $\hat{\beta}$ 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 \( \mu_{ij} \) 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 $\gamma = 0$

<table>
<thead>
<tr>
<th></th>
<th>$t_c$</th>
<th>$r_c$</th>
<th>$r$</th>
<th>$t$</th>
<th>$t^*$</th>
<th>$\tilde{t}$</th>
<th>$\tilde{t}^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>statistic</td>
<td>1.8307</td>
<td>2.0214</td>
<td>2.1596</td>
<td>1.9511</td>
<td>1.9257</td>
<td>1.7362</td>
<td>1.9064</td>
</tr>
<tr>
<td>$p$-value</td>
<td>0.0671</td>
<td>0.0432</td>
<td>0.0308</td>
<td>0.0510</td>
<td>0.0541</td>
<td>0.0825</td>
<td>0.0566</td>
</tr>
</tbody>
</table>

$t_c$ 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

$\gamma \neq 0$

$\gamma > 0$

---

$^{12}$based on 50 000 samples from the model with $\beta_1, \ldots, \beta_{18}$ 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

Sample
lesion maps for 50 patients\textsuperscript{13}

Patient characteristics
multiple sclerosis type (MS)\textsuperscript{14}
age
gender
disease duration (DD)
two disease severity measures (PASAT and EDSS)

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

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

\textsuperscript{14}from the supplementary material of Ge et al. (2014)

\textsuperscript{15}0 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

$$\Phi^{-1}(\pi_{ij}) = \beta_{j0} + \beta_{j1} MS_i + \beta_{j2} age_i + \beta_{j3} gender_i + \beta_{j4} DD_i + \beta_{j5} \text{PASAT}_i + \beta_{j6} \text{EDSS}_i$$
## Results

### Occurrence of infinite estimates

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS</td>
<td>75.5%</td>
</tr>
<tr>
<td>age</td>
<td>63.7%</td>
</tr>
<tr>
<td>gender</td>
<td>78.3%</td>
</tr>
<tr>
<td>DD</td>
<td>63.7%</td>
</tr>
<tr>
<td>PASAT</td>
<td>63.6%</td>
</tr>
<tr>
<td>EDSS</td>
<td>63.2%</td>
</tr>
</tbody>
</table>

### Failures in evaluation of \( r \)

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS</td>
<td>19.2%</td>
</tr>
<tr>
<td>age</td>
<td>20.5%</td>
</tr>
<tr>
<td>sex</td>
<td>22.4%</td>
</tr>
<tr>
<td>DD</td>
<td>18.1%</td>
</tr>
<tr>
<td>PASAT</td>
<td>16.8%</td>
</tr>
<tr>
<td>EDSS</td>
<td>10.3%</td>
</tr>
</tbody>
</table>

\(^{15}\) summaries based on voxels with lesion occurrence for at least one lesion across patients
Significance map for disease duration

18.9% of voxels with $|\tilde{t}| > 1$

24.8% of voxels with $|\tilde{t}^*| > 1$
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
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.

\(^{16}\)see, for example, MacKinnon and White (1985)


Location-adjusted Wald statistic

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

Preprint


Software

waldi R package\(^{17}\) (soon in CRAN!) for computing \(t^*\) for well-used models, 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