Comparison of robust saddlepoint tests: the negative binomial regression case

Stephane Heritier

Monash University, Melbourne, Australia

ICORS 2017, Wollongong, Australia
3-7 July 2017

Email: stephane.heritier@monash.edu

Collaborators: William Aeberhard and Eva Cantoni
Motivation

- Adaptive clinical trial in MS: combine $p$-values from different SMALL stages (phase II and phase III)
- Endpoint = nb of active lesions in the brain, NB regression used
Motivation

- **Adaptive clinical trial** in MS: combine \( p \)-values from different SMALL stages (phase II and phase III)
- Endpoint = nb of active lesions in the brain, **NB regression** used

- **Wald/LRT tests** in count data models - liberal
- \( \chi^2 \) asymptotic **distribution poor** (under \( H_0 \))
- **Absolute error** = \( O(n^{-1/2}) \) - both classical & robust
Motivation

- Adaptive clinical trial in MS: combine \( p \)-values from different SMALL stages (phase II and phase III)
- Endpoint = nb of active lesions in the brain, NB regression used

- Wald/LRT tests in count data models - liberal
- \( \chi^2 \) asymptotic distribution poor (under \( H_0 \))
- Absolute error = \( O(n^{-1/2}) \) - both classical & robust

- Emerging saddlepoint tests (since 2003)
- Relative error = \( O(n^{-1}) \) - both classical & robust
- Performance of these tests in this setting - not too different from GLM
Framework and notation

- $n$ i.i.d. random variables $z_1, \ldots, z_n$ from $F_\theta$
- We want to test
  \[ H_0 : \theta_2 = 0 \]
  against two-sided alternative
  \[ H_1 : \theta_2 \neq 0, \]
  where $\theta = (\theta_1^T, \theta_2^T)^T$, with $\text{dim}(\theta_2) = d < p = \text{dim}(\theta)$. 
Framework and notation

- $n$ i.i.d. random variables $z_1, \ldots, z_n$ from $F_\theta$
- We want to test
  $$H_0 : \theta_2 = 0$$
  against two-sided alternative
  $$H_1 : \theta_2 \neq 0,$$
  where $\theta = (\theta_1^T, \theta_2^T)^T$, with $\text{dim}(\theta_2) = d < p = \text{dim}(\theta)$.

- NB model: $z_i = (x_i, y_i)$ where $y_i$ is the count response, $x_i$ the vector of covariates
  - $E(y_i) = \mu_i = \exp(x_i^T \beta)$, $\text{var}(y_i) = \mu_i + \sigma^2 \mu_i^2$
  - $\theta_1 = (\sigma, \beta_0, \ldots, \beta_{p-d-1})^T$, nuisance parameter
  - $\theta_2 = (\beta_{p-d}, \ldots, \beta_p)^T$, parameter of interest
$\hat{\theta}$ solution for $\theta$ to:

$$\sum_{i=1}^{n} \psi(z_i; \theta) = 0$$

where $\psi(z; \theta)$ is a (bounded) score function
Motivation

Theoretical saddlepoint test

Empirical saddlepoint tests

Simulations - Design

**M-estimator**

- \( \hat{\theta} \) solution for \( \theta \) to:

\[
\sum_{i=1}^{n} \psi(z_i; \theta) = 0
\]

where \( \psi(z; \theta) \) is a (bounded) score function

- In particular:
  - NB: model score & Tukey biweight (Aeberhard et al., 2014)
  - Any redescending \( M \)-estimator (e.g. deviance-based) could have been used
\[ \hat{\theta} \] solution for \( \theta \) to:

\[
\sum_{i=1}^{n} \psi(z_i; \theta) = 0
\]

where \( \psi(z; \theta) \) is a (bounded) score function

- In particular:
  - NB: model score & Tukey biweight (Aeberhard et al., 2014)
  - Any redescending \( M \)-estimator (e.g. deviance-based) could have been used

- Usual assumption: \( E_{\theta}(\psi(z; \theta)) = 0 \)
Robust Estimation of $\beta$ and $\sigma$

Robust $M$-estimator $\hat{\sigma}$ and $\hat{\beta}$ are the solutions to

\[
0 = \sum_{i=1}^{n} \psi(y_i, x_i; \sigma, \beta)
= \left( \frac{\sum_{i=1}^{n} (\tilde{w}(y_i, \mu_i) s_\sigma(y_i, x_i, \beta, \sigma) w(x_i) - a_i(\sigma, \tilde{w}))}{\sum_{i=1}^{n} (w^*(y_i, \mu_i) s_\beta(y_i, x_i, \beta, \sigma) w(x_i) - a_i(\beta, w^*))} \right)
\]

where $s_\sigma, s_\sigma$ are the model score functions,
\[
a_i(\sigma, \tilde{w}) = E \left[ \tilde{w}(y_i, \mu_i) \psi_\sigma(y_i, x_i, \beta, \sigma) \right] w(x_i) \text{ and }
a_i(\beta, w^*) = E \left[ w^*(y_i, \mu_i) \psi_\beta(y_i, x_i, \beta, \sigma) \right] w(x_i)
\]
ensure Fisher consistency.
Robust Estimation of $\beta$ and $\sigma$

- Robust $M$-estimator $\hat{\sigma}$ and $\hat{\beta}$ are the solutions to

$$0 = \sum_{i=1}^{n} \psi(y_i, x_i; \sigma, \beta)$$

$$= \left( \sum_{i=1}^{n} \left( \tilde{w}(y_i, \mu_i) s_{\sigma}(y_i, x_i, \beta, \sigma) w(x_i) - a_i(\sigma, \tilde{w}) \right) \right) / \left( \sum_{i=1}^{n} \left( w^*(y_i, \mu_i) s_{\beta}(y_i, x_i, \beta, \sigma) w(x_i) - a_i(\beta, w^*) \right) \right)$$

where $s_{\sigma}$, $s_{\sigma}$ are the model score functions,

$a_i(\sigma, \tilde{w}) = E \left[ \tilde{w}(y_i, \mu_i) \psi_{\sigma}(y_i, x_i, \beta, \sigma) \right] w(x_i)$ and

$a_i(\beta, w^*) = E \left[ w^*(y_i, \mu_i) \psi_{\beta}(y_i, x_i, \beta, \sigma) \right] w(x_i)$ ensure Fisher consistency.

- Choice of weight $\tilde{w}$, $w^*$: based on Tukey biweight of the Pearson residual
Robust Estimation of $\beta$ and $\sigma$

- Robust $M$-estimator $\hat{\sigma}$ and $\hat{\beta}$ are the solutions to

$$0 = \sum_{i=1}^{n} \psi(y_i, x_i; \sigma, \beta)$$

$$= \left( \frac{\sum_{i=1}^{n} (\tilde{w}(y_i, \mu_i)s_{\sigma}(y_i, x_i, \beta, \sigma)w(x_i) - a_i(\sigma, \tilde{w}))}{\sum_{i=1}^{n} (w^*(y_i, \mu_i)s_{\beta}(y_i, x_i, \beta, \sigma)w(x_i) - a_i(\beta, w^*))} \right)$$

where $s_{\sigma}, s_{\sigma}$ are the model score functions,

$a_i(\sigma, \tilde{w}) = E\left[\tilde{w}(y_i, \mu_i)\psi_{\sigma}(y_i, x_i, \beta, \sigma)\right]w(x_i)$ and

$a_i(\beta, w^*) = E\left[w^*(y_i, \mu_i)\psi_{\beta}(y_i, x_i, \beta, \sigma)\right]w(x_i)$ ensure Fisher consistency.

- Choice of weight $\tilde{w}, w^*$: based on Tukey biweight of the Pearson residual
- Mallow’s weight $w(x_i)$ set to 1 for simplicity
Robinson, Ronchetti & Young (2003) proposed the test statistic:

$$2nh(\hat{\theta}_2) = 2n \min_{\theta_1} \max_{\lambda} \left\{ -K_\psi(\lambda; \theta_1, \hat{\theta}_2) \right\},$$

where $K_\psi(\lambda; \theta_1, \theta_2) = \log E[\exp(\lambda^T \psi(z_i; \theta_1, \theta_2))]$ is the common CGF of the score $\psi(z_i; \theta_1, \theta_2)$. 
Theoretical Saddlepoint Test (SDPT)

- Robinson, Ronchetti & Young (2003) proposed the test statistic:

\[ 2nh(\hat{\theta}_2) = 2n \min_{\theta_1} \max_{\lambda} \left\{ -K_\psi(\lambda; \theta_1, \hat{\theta}_2) \right\}, \]

where \( K_\psi(\lambda; \theta_1, \theta_2) = \log E[\exp(\lambda^T \psi(z_i; \theta_1, \theta_2))] \) is the common CGF of the score \( \psi(z_i; \theta_1, \theta_2) \).

- Relative error property, under \( H_0 \):

\[ P(2nh(\hat{\theta}_2) \geq 2nh(\hat{\theta}_2^{obs})) = \left(1 - Q_d(2nh(\hat{\theta}_2^{obs}))\right)\left(1 + O(1/n)\right) \]

where \( Q_d(.) \) is the CDF of a \( \chi^2(d) \).
Robinson, Ronchetti & Young (2003) proposed the test statistic:

$$2nh(\hat{\theta}_2) = 2n \min_{\theta_1} \max_{\lambda} \left\{ -K_\psi(\lambda; \theta_1, \hat{\theta}_2) \right\},$$

where $K_\psi(\lambda; \theta_1, \theta_2) = \log E[\exp(\lambda^T \psi(z_i; \theta_1, \theta_2))]$ is the common CGF of the score $\psi(z_i; \theta_1, \theta_2)$.

Relative error property, under $H_0$:

$$P(2nh(\hat{\theta}_2) \geq 2nh(\hat{\theta}_2^{obs})) = \left(1 - Q_d(2nh(\hat{\theta}_2^{obs}))\right)\left(1 + O(1/n)\right)$$

where $Q_d(.)$ is the CDF of a $\chi^2(d)$.

Robust version for GLMs given in Lô and Ronchetti (2009)
Empirical SDPT test (ESDPT)

- Computation of the expectation in $K_\psi(\lambda; \theta)$ often untractable
- Replace expectation by a weighted sum:

$$K_w(\lambda; \theta_1, \hat{\theta}_2) = \log \sum_{i=1}^{n} \hat{\pi}_i \exp(\lambda^T \psi(z_i; \theta_1, \hat{\theta}_2)),$$

where $\hat{\pi}_i$'s are fixed ET weights computed under $H_0$

$$ESDPT = 2n \min_{\theta_1} \max_{\lambda} \{-K_w(\lambda; \theta_1, \hat{\theta}_2)\}.$$
Empirical SDPT test (ESDPT)

- Computation of the expectation in $K_\psi(\lambda; \theta)$ often untractable
- Replace expectation by a weighted sum:

$$K_w(\lambda; \theta_1, \hat{\theta}_2) = \log \sum_{i=1}^{n} \hat{\pi}_i \exp(\lambda^T \psi(z_i; \theta_1, \hat{\theta}_2)),$$

where $\hat{\pi}_i$'s are fixed ET weights computed under $H_0$

$$ESDPT = 2n \min_{\theta_1} \max_{\lambda} \{-K_w(\lambda; \theta_1, \hat{\theta}_2)\}.$$

- $\chi^2(d)$ approximation under the null still holds
- Relative error of $O(1/n)$ preserved - Ma & Ronchetti (2011)
Exponential tilting (ET) weights

- $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)^\top$ = tilted empirical distribution which satisfies the $M$-estimating equations under $H_0$.

- chosen to be the closest to the empirical distribution $1/n \forall i$ with respect to the backward Kullback-Leibler (KL) divergence

- In practice: solve a first min-max problem under the null (without weights)
Camponovo & Otsu (manuscript) recently proposed:

$$TETT = 2n \log \sum_{i=1}^{n} \hat{\pi}_i \exp(\tilde{\lambda}_0^T \psi(z_i; \tilde{\theta}_1, 0))$$

where $\tilde{\theta} = (\tilde{\theta}_1^T, 0^T)^T$ is the ET estimator of $\theta$ under $H_0$ (obtained through a min-max of the empirical version of $-K_\psi$).

$\hat{\pi}_i$’s are again the ET weights computed under $H_0$
Camponovo & Otsu (manuscript) recently proposed:

$$TETT = 2n \log \sum_{i=1}^{n} \hat{\pi}_i \exp(\tilde{\lambda}_0^T \psi(z_i; \tilde{\theta}_1, 0))$$

where $\tilde{\theta} = (\tilde{\theta}_1^T, 0^T)^T$ is the ET estimator of $\theta$ under $H_0$ (obtained through a min-max of the empirical version of $-K_\psi$).

$\hat{\pi}_i$’s are again the ET weights computed under $H_0$.

Extension of the GEL test statistic (Newey & Smith, 2004)
Camponovo & Otsu (manuscript) recently proposed:

\[
TETT = 2n \log \sum_{i=1}^{n} \hat{\pi}_i \exp(\tilde{\lambda}_0^T \psi(z_i; \tilde{\theta}_1, 0)
\]

where \( \tilde{\theta} = (\tilde{\theta}_1^T, 0^T)^T \) is the ET estimator of \( \theta \) under \( H_0 \) (obtained through a min-max of the empirical version of \(-K_\psi\)).

\( \hat{\pi}_i \)'s are again the ET weights computed under \( H_0 \)

Extension of the GEL test statistic (Newey & Smith, 2004)

Relative error of \( O(1/n) \) is preserved - manuscript, 2017
Camponovo & Otsu (manuscript) recently proposed:

\[ \text{TETT} = 2n \log \sum_{i=1}^{n} \hat{\pi}_i \exp(\tilde{\lambda}_0^T \psi(z_i; \tilde{\theta}_1, 0) \]

where \( \tilde{\theta} = (\tilde{\theta}_1^T, 0^T)^T \) is the ET estimator of \( \theta \) under \( H_0 \) (obtained through a min-max of the empirical version of \(-K_\psi\)).

\( \hat{\pi}_i \)'s are again the ET weights computed under \( H_0 \)

Extension of the GEL test statistic (Newey & Smith, 2004)

Relative error of \( O(1/n) \) is preserved - manuscript, 2017

TETT simpler than ESDPT; only requires one min-max under \( H_0 \).
Simulations NB - design

- $y_i \sim NB(\mu_i, \sigma)$, $\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})$
  - overdispersion $\sigma = 0.6$
- Covariates $x_{i1}, x_{i3}, x_{i4}$ are i.i.d $N(0, 1)$, $x_{i2}$ binary indicator
Simulations NB - design

- $y_i \sim NB(\mu_i, \sigma)$, $\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})$
  overdispersion $\sigma = 0.6$
- Covariates $x_{i1}, x_{i3}, x_{i4}$ are i.i.d $N(0, 1)$, $x_{i2}$ binary indicator
- $H_0$: $\theta_2 = (\beta_3, \beta_4)^T = 0$, null distribution $\chi^2(2)$
- $\theta_1 = (\sigma, \beta_0, \beta_1, \beta_2)^T = (0.7, 0.4, 0.7, -0.5)^T$, nuisance parameter
Simulations NB - design

- \( y_i \sim NB(\mu_i, \sigma) \), \( \mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}) \)
  overdispersion \( \sigma = 0.6 \)
- Covariates \( x_{i1}, x_{i3}, x_{i4} \) are i.i.d \( N(0, 1) \), \( x_{i2} \) binary indicator

- \( H_0: \theta_2 = (\beta_3, \beta_4)^T = 0 \), null distribution \( \chi^2(2) \)
- \( \theta_1 = (\sigma, \beta_0, \beta_1, \beta_2)^T = (0.7, 0.4, 0.7, -0.5)^T \), nuisance parameter

- **Level**: \( n = 20, 50, 100 \) and \( B = 50,000 \) runs - only \( n = 50 \) here!
- **Power**: let \( \theta_2 = (\beta_3, \beta_4)^T \) move away from the null; only \( B = 10,000 \) replications
Simulations NB - design (2)

- Contamination: add 20 to 5% of observations (3 in practice for $n = 50$)

- Robust: tuning constant calibrated to reach 90% efficiency at NB model

- Tests considered: standard asymptotic tests (LRT, Wald, score) + all saddlepoint tests

- Classical and robust versions
Simulations: level at NB model
Simulations: level, 5% contamination
Simulations: power at NB model
Simulations: power, 5% contamination
Conclusions

- **SPTs** generally more accurate than standard asymptotic classical or robust tests.
- Robust score test performed unexpectedly well!
- Theoretical SPT (if available) preferable to empirical SPTs. May require lower tuning constants to control the level.
Conclusions

- **SPTs** generally *more accurate* than standard asymptotic classical or robust tests
- Robust score test performed unexpectedly well!
- **Theoretical SPT** (if available) *preferable* to empirical SPTs. May require *lower tuning constants* to control the level

- **TETT** better than **ESDPT** in terms of level but power is lower (classical case)
- **Issue with robust ESDPT** (power) - unclear why
Conclusions

- **SPTs** generally more accurate than standard asymptotic classical or robust tests
- **Robust score test** performed unexpectedly well!
- **Theoretical SPT** (if available) preferable to empirical SPTs. May require lower tuning constants to control the level
- **TETT** better than ESDPT in terms of level but power is lower (classical case)
- **Issue with robust ESDPT** (power) - unclear why
- **Computation = challenge:** stable algorithms required
- **Collapsing power:** needs further investigation (possibly linked to redescending estimators)
References

- R-functions: https://github.com/williamaeberhard/robnb