

Asymptotic properties of the modified median estimator

Jana Novotná

FNSPE, Czech Technical University in Prague

22.6.2018



Outline

- 1 Regression models
 - Generalized linear model
 - Logistic regression model
- 2 Parameter estimations
- 3 Asymptotic properties of the modified median estimator
 - Main theorem
- 4 Simulations for the modified median estimator
 - Robustness
 - Asymptotic properties

Generalized linear model

- Y_1, \dots, Y_n - dependent variables from the exponential family of distributions
- $\mathbf{x}_1^T = (x_{11}, \dots, x_{1p}), \dots, \mathbf{x}_n^T = (x_{n1}, \dots, x_{np})$ - explanatory variables (regressors)
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ - unknown parameters
- Expected value is modelled

$$g(E(Y_i)) = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Logistic regression model

- $Y_1 \sim \text{Be}(\pi_1), \dots, Y_n \sim \text{Be}(\pi_n)$
- The linking function is in the form

$$g(\pi_i) = \text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

- Which implies

$$\pi_i = \pi(\mathbf{x}_i^T \boldsymbol{\beta}) = \frac{\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}$$

Parameter estimations

- Maximum likelihood estimation

$$\hat{\beta}_{\text{MLE}} = \arg \min_{\beta} \sum_{i=1}^n d_i(\beta)$$

$$\begin{aligned}d_i(\beta) &= -\ln \left[\pi_i(\beta)^{Y_i} (1 - \pi_i(\beta))^{1-Y_i} \right] \\&= -Y_i \ln(\pi_i(\beta)) - (1 - Y_i) \ln(1 - \pi_i(\beta))\end{aligned}$$

- Modified median estimator

$$\hat{\beta}_{\text{MED}} = \arg \min_{\beta} \sum_{i=1}^n \int_0^1 |Y_i + u - m(\pi(x_i^T \beta))| du$$

$$m(p) = 1 + \frac{p - \frac{1}{2}}{\frac{1}{2} + |p - \frac{1}{2}|}$$

Asymptotic properties

Theorem

Under the assumption the modified median estimator $\hat{\beta}_{\text{MED}}$ is consistent estimator of β , its asymptotic distribution is given by

$$\sqrt{n}(\hat{\beta}_{\text{MED}} - \beta^0) \xrightarrow[n \rightarrow +\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}_p, \mathbf{Q}^{-1}(\beta^0)\boldsymbol{\Sigma}(\beta^0)\mathbf{Q}^{-1}(\beta^0)).$$

Asymptotic properties

$$\Sigma(\beta^0) = \frac{1}{4} \left(\int_{\Upsilon_1(\beta^0)} \left(\frac{\pi(x^T \beta)}{1 - \pi(x^T \beta)} \right)^3 x x^T dG(x) + \int_{\Upsilon_2(\beta^0)} \left(\frac{1 - \pi(x^T \beta)}{\pi(x^T \beta)} \right)^3 x x^T dG(x) \right),$$
$$Q(\beta^0) = -\frac{1}{2} \left(\int_{\Upsilon_1(\beta^0)} \frac{\pi^2(x^T \beta)}{1 - \pi(x^T \beta)} x x^T dG(x) + \int_{\Upsilon_2(\beta^0)} \frac{(1 - \pi(x^T \beta))^2}{\pi(x^T \beta)} x x^T dG(x) \right),$$

where

$$\Upsilon_1(\beta^0) = \left\{ x \in \mathcal{X} : 0 < \pi(x^T \beta^0) \leq \frac{1}{2} \right\} = \{x \in \mathcal{X} : x^T \beta^0 \leq 0\},$$

$$\Upsilon_2(\beta^0) = \left\{ x \in \mathcal{X} : \frac{1}{2} < \pi(x^T \beta^0) < 1 \right\} = \{x \in \mathcal{X} : x^T \beta^0 > 0\}.$$

Simulations

Experiment setting:

- The vector of regression coefficients equals

$$\beta = (\beta_1, \beta_2)^T = (-2,82; 2,82)^T$$

- Explanatory variables comply

$$x_{i1} = 1, \quad x_{i2} \sim \mathcal{N}(0, 1), \quad x_{i2} \text{ independent}, \quad i = 1, 2, \dots, n$$

- Dependent variables are generated by

$$Y_i \sim (1 - \varepsilon) \text{Be}(\pi_i) + \varepsilon \text{Be}(1 - \pi_i), \quad i = 1, 2, \dots, n$$

Simulations

Robustness

- Examined quantities

$$Z_j = \frac{1}{2}(|\hat{\beta}_1^{(j)} - \beta_1| + |\hat{\beta}_2^{(j)} - \beta_2|), \quad j = 1, \dots, N$$

$$\text{MAE} = \bar{Z}_N = \frac{1}{N} \sum_{j=1}^N Z_j$$

$$\text{s.e.} = \sqrt{\frac{1}{N-1} \sum_{j=1}^N (Z_j - \bar{Z}_N)^2}$$

Simulations

Robustness

n	MAE		s.e.	
	MED	MLE	MED	MLE
50	1,7576	1,1417	2,4277	1,6108
100	1,2791	0,6733	1,9236	0,7885
500	0,4163	0,2343	0,3763	0,1660
1000	0,2768	0,1636	0,2208	0,1145

Table: Non-contaminated data

Simulations

Robustness

n	MAE		s.e.	
	MED	MLE	MED	MLE
50	1,7074	1,0504	2,2985	0,6377
100	1,0333	1,0132	1,3249	0,4120
500	0,5223	1,0297	0,2821	0,1869
1000	0,4824	1,0372	0,2380	0,1319

Table: Data contamination 5 %

Simulations

Robustness

n	MAE		s.e.	
	MED	MLE	MED	MLE
50	1,6681	1,3972	1,9146	0,4548
100	1,4585	1,4930	1,6501	0,3414
500	0,8969	1,5130	0,3122	0,1344
1000	0,9308	1,5142	0,2248	0,0971

Table: Data contamination 10 %

Simulations

Asymptotic properties

- Theoretical asymptotic covariance matrixes

$$\mathbf{C}_{\text{MED}} = \begin{pmatrix} 92,748 & -89,957 \\ -89,957 & 110,205 \end{pmatrix}$$

$$\mathbf{C}_{\text{MLE}} = \begin{pmatrix} 34,704 & -31,097 \\ -31,097 & 42,586 \end{pmatrix}$$

- We examine

$$\text{ERR} = \frac{\sum_{i=1}^2 \sum_{j=1}^2 |\hat{c}_{ij} - c_{ij}|}{\sum_{i=1}^2 \sum_{j=1}^2 |c_{ij}|}$$

Simulations

Asymptotic properties

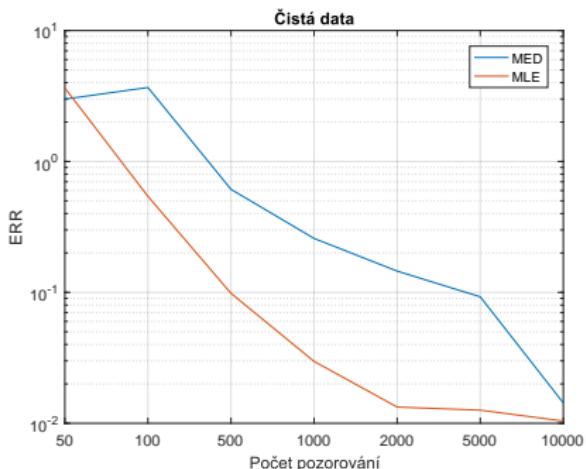


Figure: Non-contaminated data

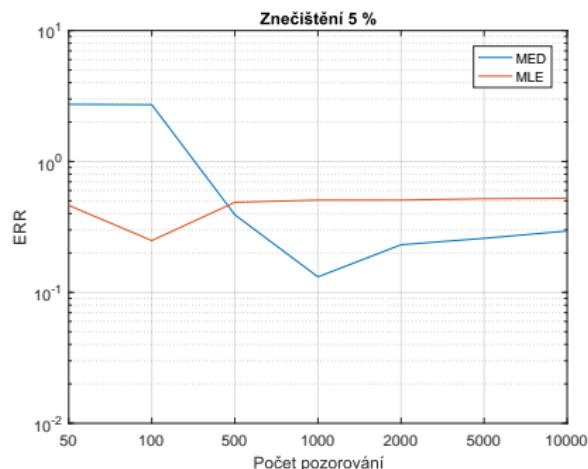


Figure: Data contamintion 5 %

Simulations

Asymptotic properties

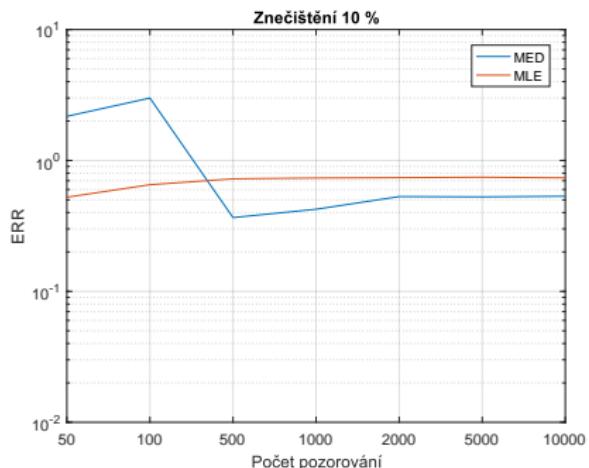


Figure: Data contamintion 10 %

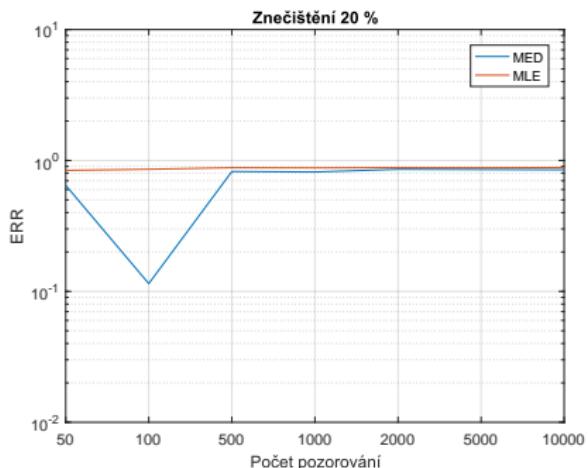


Figure: Data contamintion 20 %

Conclusion

- Asymptotic normality has not been rejected
- Results of the simulations correspond with the theorem