JOHANNES KEPLER
UNIVERSITÄT LINZ | JKU

The Problem of Modeling Rare Events in ML-based Logistic Regression

Assessing Potential Remedies via MC Simulations

Heinz Leitgöb

University of Linz, Austria

Problem

- In logistic regression, MLEs are consistent but only asymptotically unbiased -> MLEs may be heavily biased away from 0
- McCullagh and Nelder (1989) determine the bias as

$$\mathbf{b} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\xi} \qquad (1)$$

- Rare events (= a very small #e) exacerbate **b**
- This phenomenon is well known in the statistical literature (for an overview see Gao and Shen (2007)) but—thus far—not adequately communicated to applied researchers by the available textbooks on logistic regression

Conventional ML-based logistic regression

Logistic regression model

$$y \sim B(1,\pi) \text{ with } \pi = \frac{exp(\eta)}{1+exp(\eta)} \quad with \quad \eta = \sum_{j=1}^{k} x_j \beta_j \qquad (2)$$

(log-)Likelihood functions and score vector

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i{}^{y_i} (1-\pi_i)^{1-y_i} \qquad (3)$$

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log \pi(\mathbf{x}_i, \boldsymbol{\beta}) + (1-y_i) \log\left(1-\pi(\mathbf{x}_i, \boldsymbol{\beta})\right) \right] \qquad (4)$$

$$\mathbf{q} = \left(\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) = \mathbf{0} \qquad (5)$$

Potential remedies

- Exact logistic regression (Stata command: *exlogistic*)
- Bias correction method proposed by King and Zeng (2001a, 2001b) (Stata command: *relogit*)
- Penalized maximum likelihood estimation (PMLE) proposed by Firth (1993) (Stata command: *firthlogit*)

Exact logistic regression

Principle: exact computation of parameter estimates -> foregoes asymptotic properties of estimates as in MLE

First result:

Exact logistic regression is only applicable when

- n is (very) small (<200)
- covariates are discrete (best: dichotomous)
- # of covariates is small

otherwise: working memory will

Bias correction method proposed by King and Zeng (2001a, 2001b)

Principle: 2-step correction procedure

- **Step 1:** correction of finite sample bias -> generation of an unbiased vector of parameter estimates $\widetilde{\boldsymbol{\beta}}$ (for bias$\left(\widehat{\boldsymbol{\beta}}\right)$ see Eq. (1))

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \widehat{\text{bias}}\left(\widehat{\boldsymbol{\beta}}\right) \qquad (6)$$

- **Step 2:** changes in $\widetilde{\boldsymbol{\beta}}$ usually do not affect $\tilde{\pi}$ symmetrically -> approximate correction can be realized by

$$\Pr(y_i = 1|\mathbf{x}_i) \approx \tilde{\pi}_i + C_i \qquad (7)$$

with

$$C_i = (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)\mathbf{x}_0 V\left(\widetilde{\boldsymbol{\beta}}\right)\mathbf{x}'_0 \qquad (8)$$

PMLE proposed by Firth (1993)

Principle: extending the logL_{ML}-function—and thus the elements of the score vector—by a penalization term which is sensitive to decreasing n and #e

$$L_{PML}(\boldsymbol{\beta}) = L_{ML}(\boldsymbol{\beta})|\mathbf{i}(\boldsymbol{\beta})|^{1/2} \qquad (9)$$

$$\log L_{PML}(\boldsymbol{\beta}) = \log L_{ML}(\boldsymbol{\beta}) + 1/2 \log|\mathbf{i}(\boldsymbol{\beta})| \quad (10)$$

$$\mathbf{q}_{PML} = \mathbf{q}_{ML} + 1/2\, tr\left[\mathbf{i}^{-1}\left(\frac{\partial \mathbf{i}}{\partial \boldsymbol{\beta}}\right)\right] \qquad (11)$$

$|\mathbf{i}(\boldsymbol{\beta})|^{1/2}$ … Jeffreys (1946) invariant prior

Research questions

- How biased are MLEs in small samples with rare events?
- How do the alternative estimation procedures perform under these conditions (focusing on unbiasedness)

Design of Monte Carlo (MC) simulation

Table 1: Simulation design (#e)

	n				
p	5,000	1,000	500	250	100
0.5	2,500	500	250	125	50
0.1	500	100	50	25	10
0.05	250	50	25	≈13	5
0.01	50	10	5	–	–

Linear predictors (η_p):

$\eta_{0.5} = 2x_1;\ x_1 \sim N(0,1)$

$\eta_{0.1} = -3.3 + 2x_1;\ x_1 \sim N(0,1)$

$\eta_{0.05} = -4.3 + 2x_1;\ x_1 \sim N(0,1)$

$\eta_{0.01} = -6.6 + 2x_1;\ x_1 \sim N(0,1)$

10,000 replications

Variation in n and p (-> in β_0) while $\beta_1 = 2$ and the # of covariates are kept constant

MC simulation results — conventional MLE

Table 2a: Conventional MLE — mean intercepts

p (β_0)	5,000	1,000	500	250	100
0.5 (0)	***-0.0001084***	***0.0004301***	***0.0018214***	***0.0001666***	***0.0028939***
0.1 (-3.3)	***-3.3041540***	***-3.3235550***	***-3.3491030***	***-3.4005300***	***-3.6099880***
0.05 (-4.3)	***-4.3109700***	***-4.3456310***	***-4.3998010***	***-4.5071530***	***-5.0832930***
0.01 (-6.6)	***-6.6438070***	***-6.9020720***	***-7.4014120***	—	—

Table 2b: Conventional MLE — mean slopes

p (β_1=2)	5,000	1,000	500	250	100
0.5	***2.0016170***	***2.0102410***	***2.0197690***	***2.0364510***	***2.1058500***
0.1	***2.0027970***	***2.0175180***	***2.0381940***	***2.0789060***	***2.2306880***
0.05	***2.0069420***	***2.0261880***	***2.0563780***	***2.1174040***	***2.4500680***
0.01	***2.0149170***	***2.1007360***	***2.2920260***	—	—

Italic implies, that the 95%ci does not contain the true score

MC simulation results — King/Zeng correction

Table 3a: King/Zeng correction— mean intercepts

p (β_0)	5,000	1,000	500	250	100
0.5 (0)	-0,0001084	0,0004286	0,0018100	0,0001658	0,0027876
0.1 (-3.3)	-3,2996170	-3,2995520	-3,3017820	-3,3001730	***-3,2840120***
0.05 (-4.3)	-4,3016320	-4,2998200	-4,3032600	-4,2905530	***-4,0773970***
0.01 (-6.6)	-6,5957210	***-6,5828270***	***-6,3526430***	—	—

Table 3b: King/Zeng correction — mean slopes

p (β_1=2)	5,000	1,000	500	250	100
0.5	1,9997340	2,0007440	2,0005830	1,9973300	2,0005910
0.1	1,9992970	1,9969960	2,0025320	2,0043180	1,9892110
0.05	2,0017520	2,0001370	2,0016380	1,9953640	***1,8951280***
0.01	1,9992300	1,9949910	***1,9447270***	—	—

Italic implies, that the 95%ci does not contain the true score

MC simulation results — Firth's PMLE

Table 4a: Firth's PMLE — mean intercepts

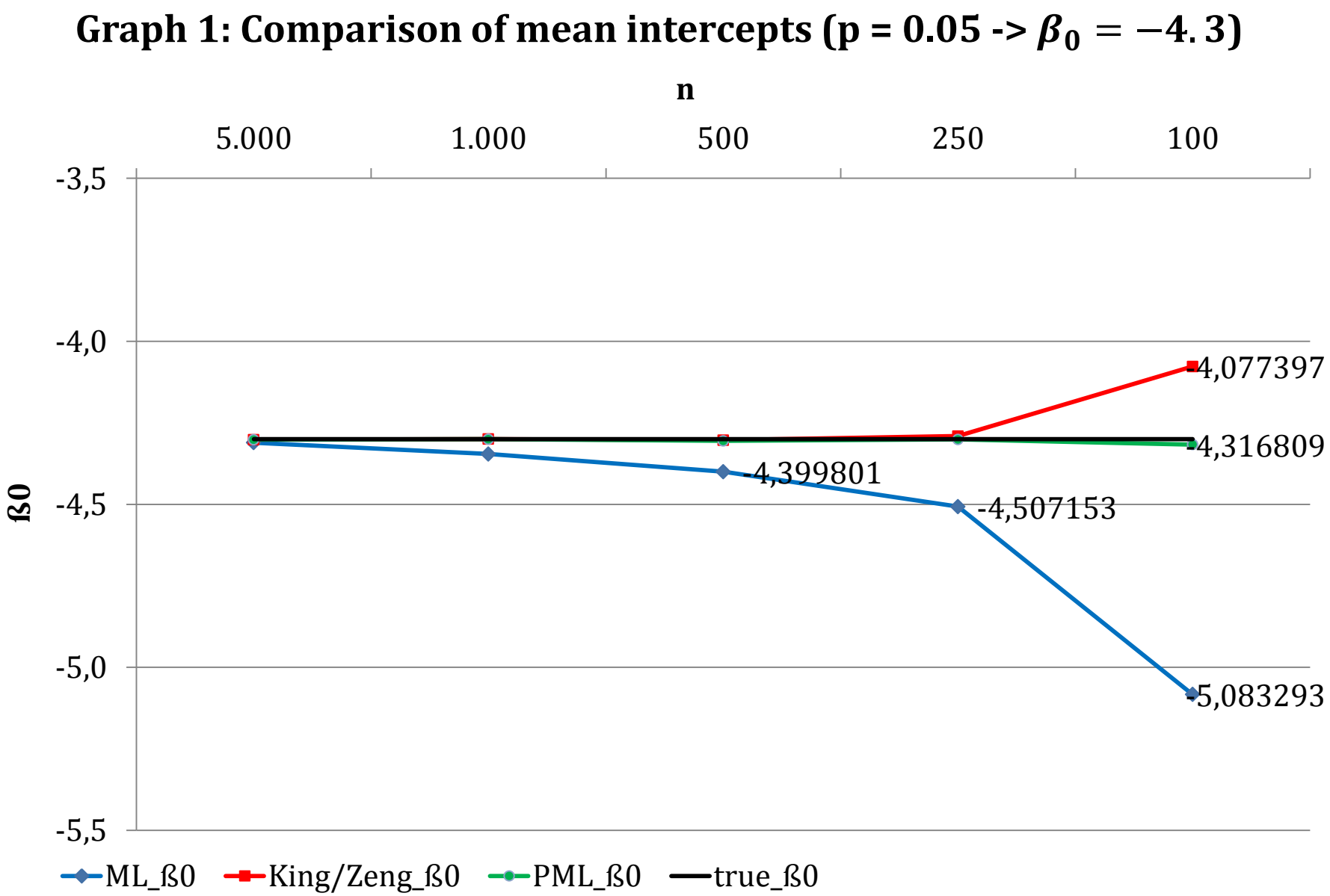
p (β_0)	5,000	1,000	500	250	100
0.5 (0)	-0,0001084	0,0004286	0,0018100	0,0001657	0,0027907
0.1 (-3.3)	-3,2996230	-3,2996900	-3,3023610	-3,3027510	-3,3129370
0.05 (-4.3)	-4,3016500	-4,3002850	-4,3053100	-4,3010580	-4,3168090
0.01 (-6.6)	-6,5961770	-6,6058880	-6,5950600	—	—

Table 4b: Firth's PMLE — mean slopes

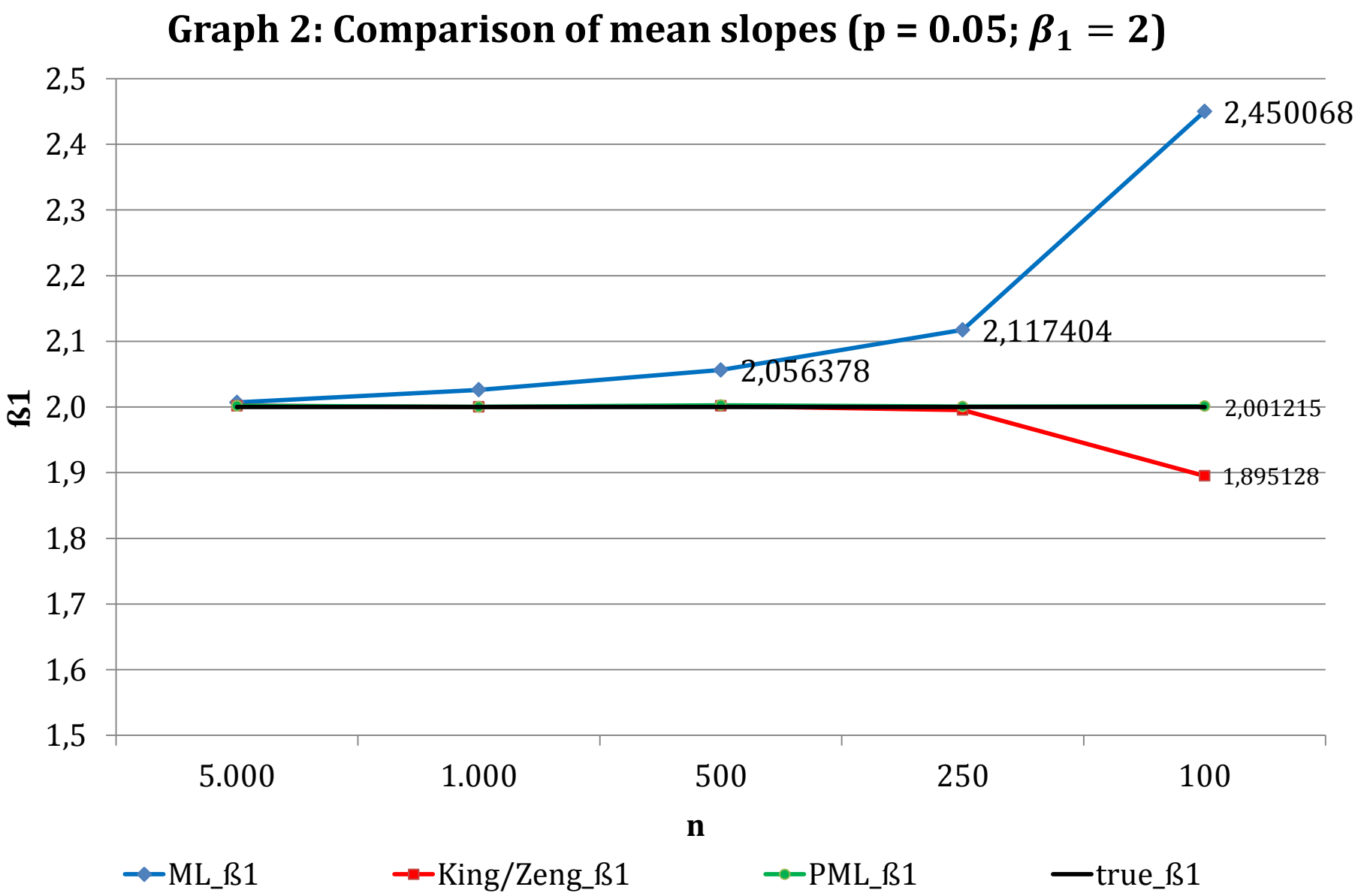
p (β_1=2)	5,000	1,000	500	250	100
0.5	1,9997350	2,0007690	2,0006860	1,9977570	2,0035800
0.1	1,9993010	1,9970950	2,0029460	2,0061450	2,0088150
0.05	2,0017610	2,0003810	2,0027020	2,0006890	2,0012150
0.01	1,9993380	1,9998850	***1,9713260***	—	—

Italic implies, that the 95%ci does not contain the true score

Comparison of mean intercepts



Comparison of mean slopes



Comparison of probabilities

Table 5: Comparison of $\Pr(y = 1|x1 = 1)$ for n = 100 and p = 0.05

	β_0	β_1	η	Pr	Pr*100
true	-4.3	2	-2.3	0.091122	9.112296
MLE	-5.083293	2.450068	-2.633225	0.067030	6.703049
King/Zeng	-4.077397	1.895128	-2.182269	0.101354	10.135408
PMLE	-4.316809	2.001215	-2.315594	0.089839	8.983968

Summary

- MLEs are systematically biased away from 0 as n and #e are getting small -> underestimation of the "true" $\Pr(y = 1|\mathbf{x})$

- In samples with n > 200 and/or in cases with "many" covariates and/or non-discrete covariates exact logistic regression will blow up working memory

- The correction method proposed by King/Zeng is somewhat overcorrecting bias in MLEs as n is getting small (<200)

- PMLEs seem unbiased, even in cases with small n and very few #e. **Further advantages:** PMLE is always converging and solves the "problem of separation" (Heinze/Schemper 2002)

Recommendations: Try to **keep n large** and **apply PMLE** when estimating logistic regression models (with rare events data)!

Future research

- Additional investigation of other desirable properties of estimates (e.g. consistency, efficiency)
- Testing the respective models for systematic bias in standard errors
- Testing the performance of the models for non-normal and discrete covariates (e.g. Poisson distributed covariates)
- Testing for a decreasing number of events per variable by including more than one covariate into the model (Peduzzi et al. 1996)

Literature

Firth, D. (1993): Bias reduction of maximum likelihood estimates. In: Biometrika 80: 27-38.

Gao, S./Shen, J. (2007): Asymptotic properties of a double penalized maximum likelihood estimator in logistic regression. In: Statistics and Probability Letters 77: 925-930.

Heinze, G./Schemper, M. (2002): A solution to the problem of separation in logistic regression. In: Statistics in Medicine 21: 2409-2419.

Jeffreys, H. (1946): An invariant form for the prior probability in estimation problems.

King, G./Zeng, L. (2001a): Logistic Regression in Rare Events Data. In: Political Analysis 9: 137-163.

King, G./Zeng, L. (2001b): Explaining Rare Events in international Relations. In: International Organization 55: 693-715.

McCullagh, P./Nelder, J. A. (1989): Generalized Linear Models. Chapman & Hall: Boca Raton.

Peduzzi, P./Concato, J./Kemper, E./Holford T. R./Feinstein, A. R. (1996): A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. In: Journal of Clinical Epidemiology 49: 1373-1379.

Thanks for your attention!

heinz.leitgoeb@jku.at