Penalized likelihood logistic regression with rare events

Georg Heinze[1], Angelika Geroldinger[1], Rainer Puhr[2],
Mariana Nold[3], Lara Lusa[4]

[1] Medical University of Vienna, CeMSIIS, Section for Clinical Biometrics, Austria
[2] University of New South Wales, The Kirby Institute, Australia
[3] Universitätsklinikum Jena, Institute for Medical Statistics, Computer Sciences and Documentation, Germany
[4] University of Ljubljana, Institute for Biostatistics and Medical Informatics, Slovenia

Funded by the Austrian Science Fund (FWF) and the Slovenian Research Agency (ARRS)

Rare events: examples

Medicine:	
• Side effects of treatment	1/1000s to fairly common
• Hospital-acquired infections	9.8/1000 pd
• Epidemiologic studies of rare diseases	1/1000 to 1/200,000
Engineering:	
• Rare failures of systems	0.1-1/year
Economy:	
• E-commerce click rates	1-2/1000 impressions
Political science:	
• Wars, election surprises, vetos	1/dozens to 1/1000s
…	

Problems with rare events

- ‚Big' studies needed to observe enough events
- Difficult to attribute events to risk factors

- Low absolute number of events
- Low event rate

Our interest

- Models
  - for prediction of binary outcomes
  - should be interpretable,

    i.e., betas should have a meaning

    → explanatory models

Logistic regression

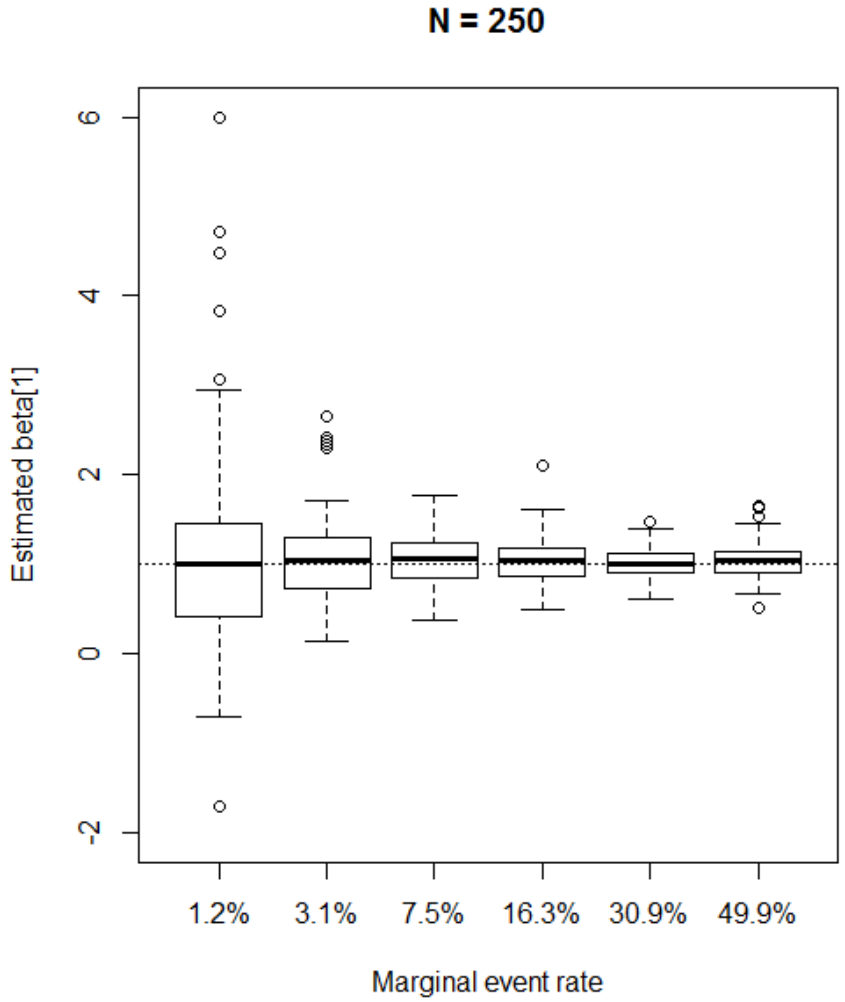
- $\Pr(Y = 1) = \pi = [1 + \exp(-X\beta)]^{-1}$

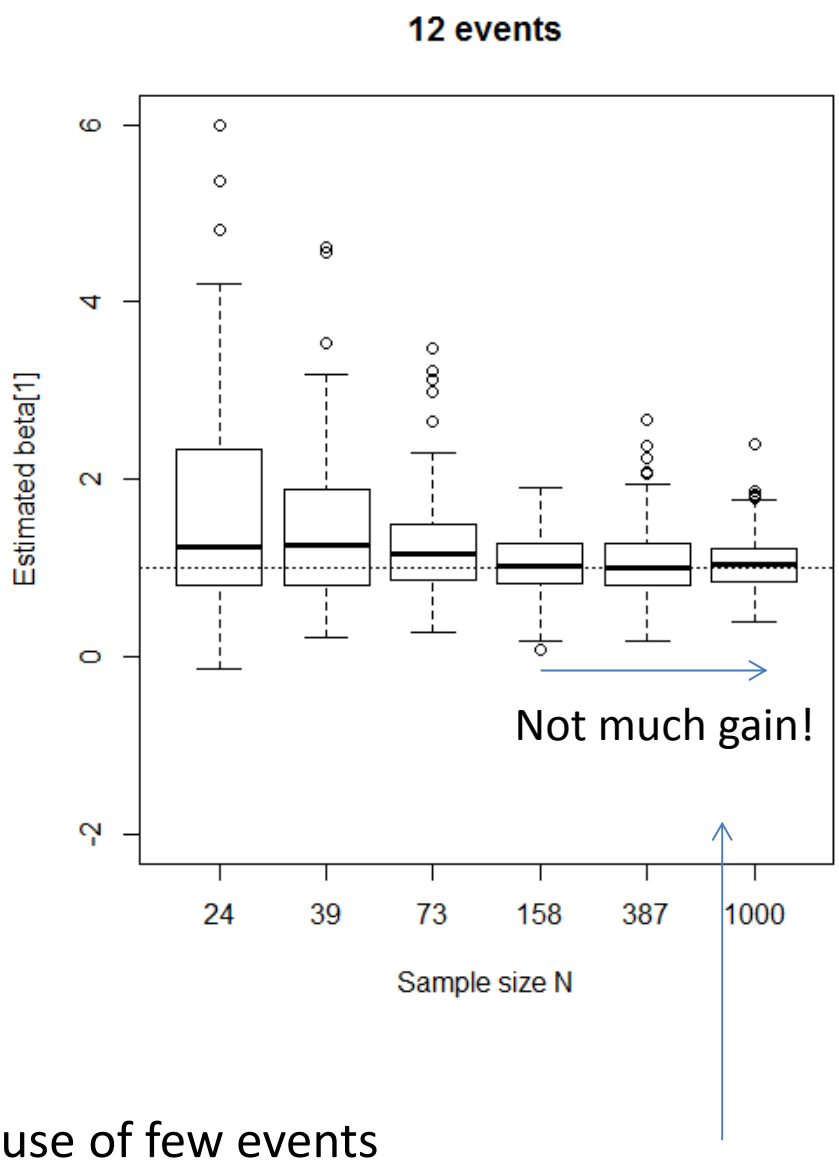
- Leads to odds ratio interpretation of $\exp(\beta)$:
- $\exp(\beta) = \frac{\Pr(Y = 1|X = x_0 + 1)/\Pr(Y=0|X=x_0+1)}{\Pr(Y = 1|X = x_0)/\Pr(Y=0|X=x_0)}$

- Likelihood: $\mathrm{L}(\beta|X) = \prod_{i=1}^{n} \hat{\pi}_i{}^{y_i}(1 - \hat{\pi}_i)^{1-y_i}$
- Its n^{th} root: Probability of correct prediction

- How well can we estimate β if events ($y_i = 1$) are rare?

Rare event problems…





Logistic regression with 5 variables:

- estimates are unstable (large MSE) because of few events
- removing some ‚non-events' does not affect precision

Penalized likelihood regression

$$\log L^*(\beta) = \log L\,(\beta) + A(\beta)$$

Imposes priors on model coefficients, e.g.

- $A(\beta) = -\lambda \sum \beta^2$ (ridge: normal prior)
- $A(\beta) = -\lambda \sum |\beta|$ (LASSO: double exponential)
- $A(\beta) = \frac{1}{2} \log \det(I(\beta))$ (Firth-type: Jeffreys prior)

in order to

- avoid extreme estimates and stabilize variance (ridge)
- perform variable selection (LASSO)
- correct small-sample bias in β (Firth-type)

Firth type penalization

In exponential family models with canonical parametrization the **Firth-type penalized likelihood** is given by

$$L^*(\theta) = L(\theta)\det(I(\theta))^{1/2},$$

Jeffreys invariant prior

where $I(\theta)$ is the Fisher information matrix.

This **removes the first-order bias** of the ML-estimates.

Software:

- logistic regression: R (logistf, brglm, pmlr), SAS, Stata…
- Cox regression: R (coxphf), SAS…

Firth, 1993; Heinze and Schemper, 2002; Heinze, 2006; Heinze, Beyea and Ploner, 2013

Firth type penalization

We are interested in logistic regression:

Here the penalized log likelihood is given by

$$\log L(\beta) + \frac{1}{2}\log\det(X^t W X)$$

with

$$W = \mathrm{diag}(\mathrm{expit}(X\beta)(1 - \mathrm{expit}(X\beta))').$$

- W is maximised at $\beta = 0$, i.e. the ML estimates are shrunken towards zero,
- for a 2×2 table (logistic regression with one binary regressor), Firth's bias correction amounts to adding $1/2$ to each cell.

Separation

(Complete) separation: a combination of the explanatory variables (nearly) perfectly predicts the outcome

- frequently encountered with small samples,
- “monotone likelihood”,
- some of the ML-estimates are infinite,
- but Firth estimates do exist!

Example:

complete separation

	A	B
0	0	10
1	10	0

quasi-complete separation

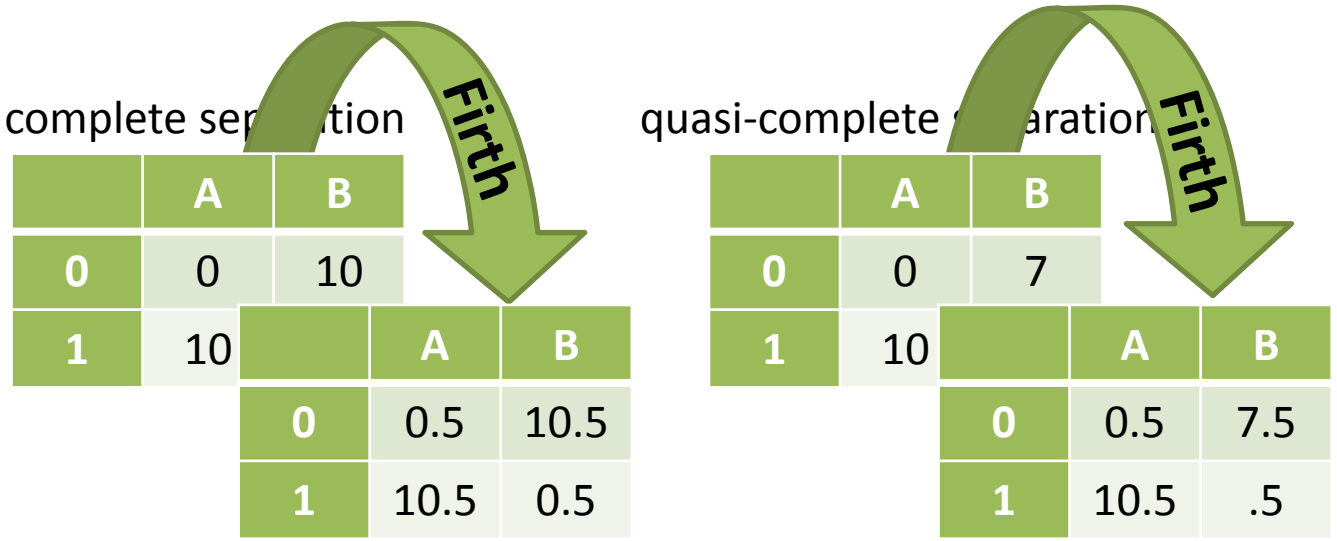
	A	B
0	0	7
1	10	3

Separation

(Complete) separation: a combination of the explanatory variables (nearly) perfectly predicts the outcome

- frequently encountered with small samples,
- “monotone likelihood”,
- some of the ML-estimates are infinite,
- but Firth estimates do exist!

Example:



	A	B
0	0	10
1	10	

	A	B
0	0.5	10.5
1	10.5	0.5

	A	B
0	0	7
1	10	

	A	B
0	0.5	7.5
1	10.5	.5

Impact of Firth correction on predictions

- Example from Greenland & Mansournia, 2015

no separation

	A	B
0	9	2966
1	1	16

quasi-complete separation

	A	B
0	10	2966
1	0	16

- ML predictions:

	A	B
1	10%	0.53%

	A	B
1	0%	0.53%

- Firth predictions:

	A	B
1	13.6%	0.55%

	A	B
1	4.5%	0.55%

Impact of Firth correction on predictions

- Example from Greenland & Mansournia, 2015

no separation

	A	B	
0	9	2966	
1	1	16	0.56%

quasi-complete separation

	A	B	
0	10	2966	
1	0	16	0.53%

- ML predictions:

	A	B	
1	10%	0.53%	0.56%

	A	B	
1	0%	0.53%	0.53%

- Firth predictions:

	A	B	
1	13.6%	0.55%	**0.59**%

	A	B	
1	4.5%	0.55%	**0.56**%

Example: Bias in logistic regression

Consider a model containing only intercept, no regressors:

$$\text{logit}\,(P(Y=1)) = \alpha.$$

With n observations, k events, the ML estimator of α is given by:

$$\hat{\alpha} = \text{logit}\,(\text{k/n}).$$

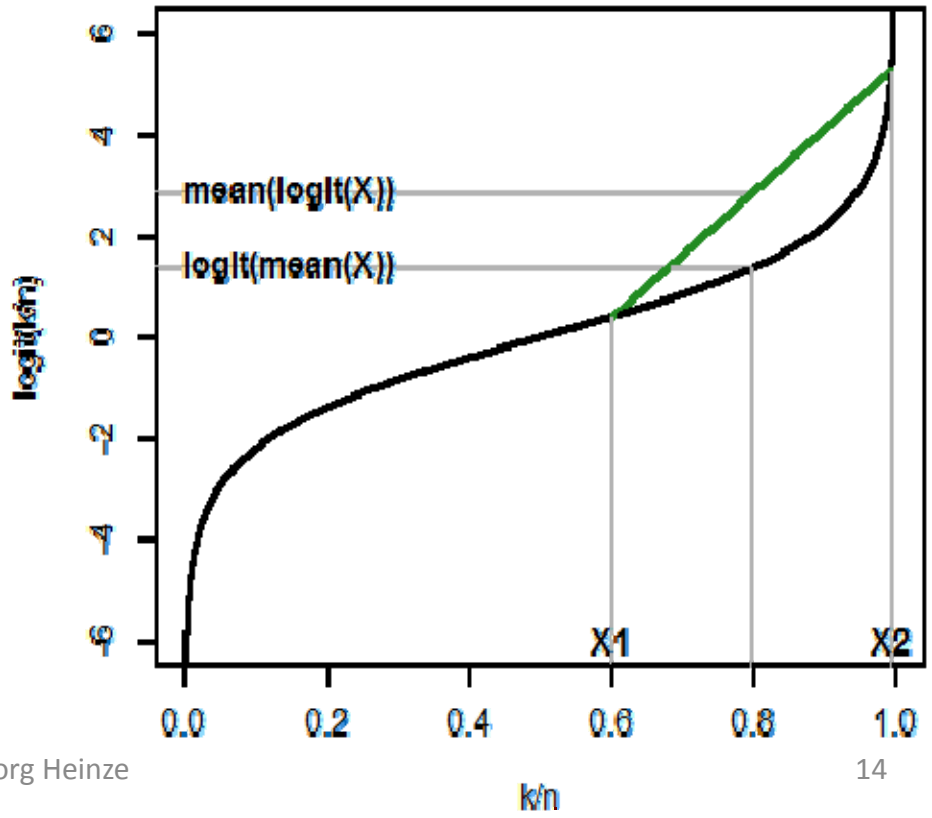
Since k/n is unbiased,

$\hat{\alpha}$ is biased!

(If $\hat{\alpha}$ was unbiased,

$\text{expit}(\hat{\alpha})$ would be biased!)

Jensen's inequality



Penalized logistic regression: ridge

The penalized likelihood is given by

$$\log L(\beta) - \lambda \|\beta\|_2^2$$

- where λ is an unknown tuning parameter,
- and the β's should be suitably standardized.

- Usually, X's are standardized to unit variance before application, and λ is tuned by cross-validation
- After application, $\hat{\beta}$ can be back-transformed to original scale

See, e.g., Friedman et al, 2010

Known und unknown features of ridge

- It reduces RMSE of predictions
- It reduces RMSE of beta coefficients
- It introduces bias in the beta coefficients
- The bias is towards the null

Known und unknown features of ridge

- It reduces RMSE of predictions
- It reduces RMSE of beta coefficients
- It introduces bias in the beta coefficients
- The bias is towards the null **?**

The ‚*we won‘t let you down*‘ effect of ridge

LogReg with 15 correlated covariates,

N=3000,

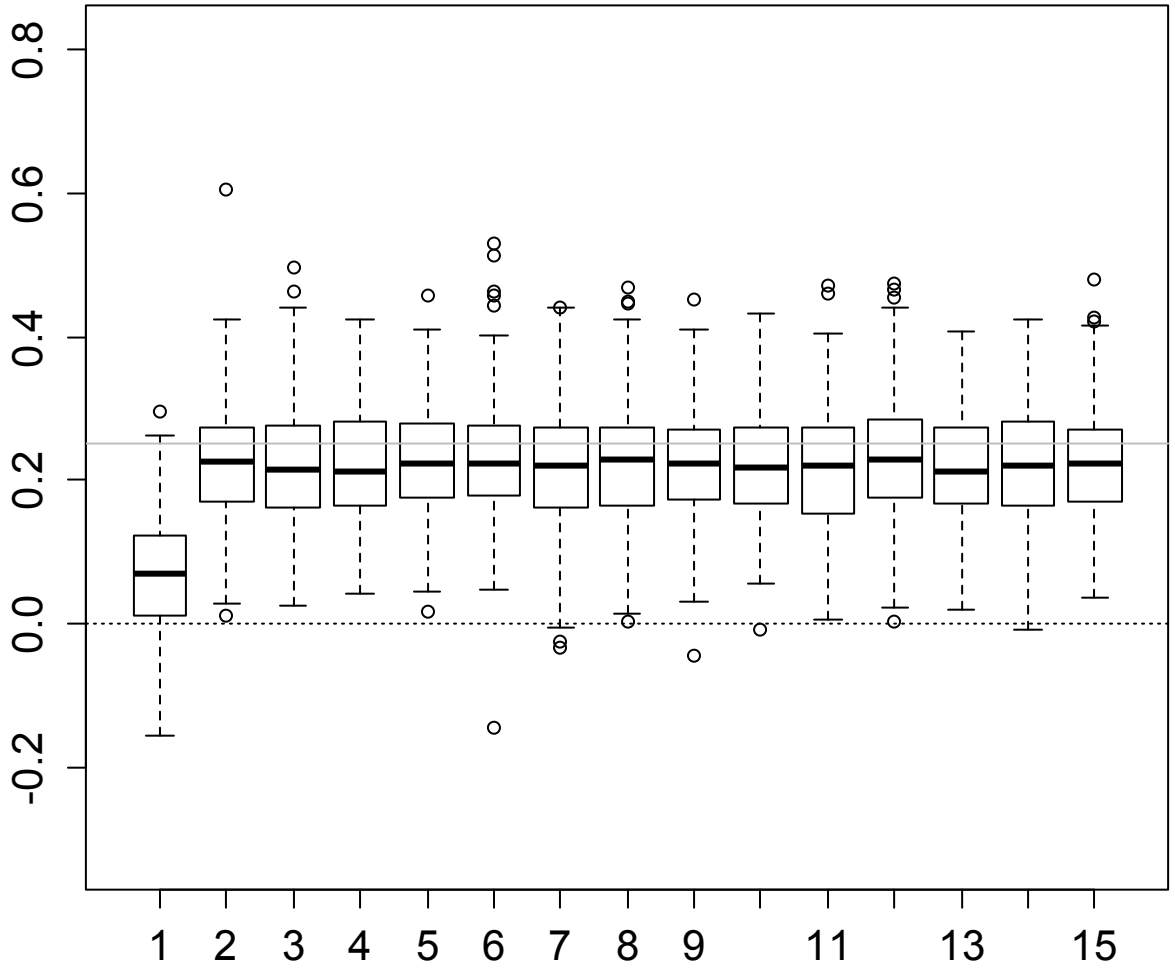
Marginal event rate 1%

True effects:

0 for X1

0.25 for X2-X15

Plot shows the betas (simulation)



For comparison: Firth

LogReg with 15 correlated covariates,

N=3000,

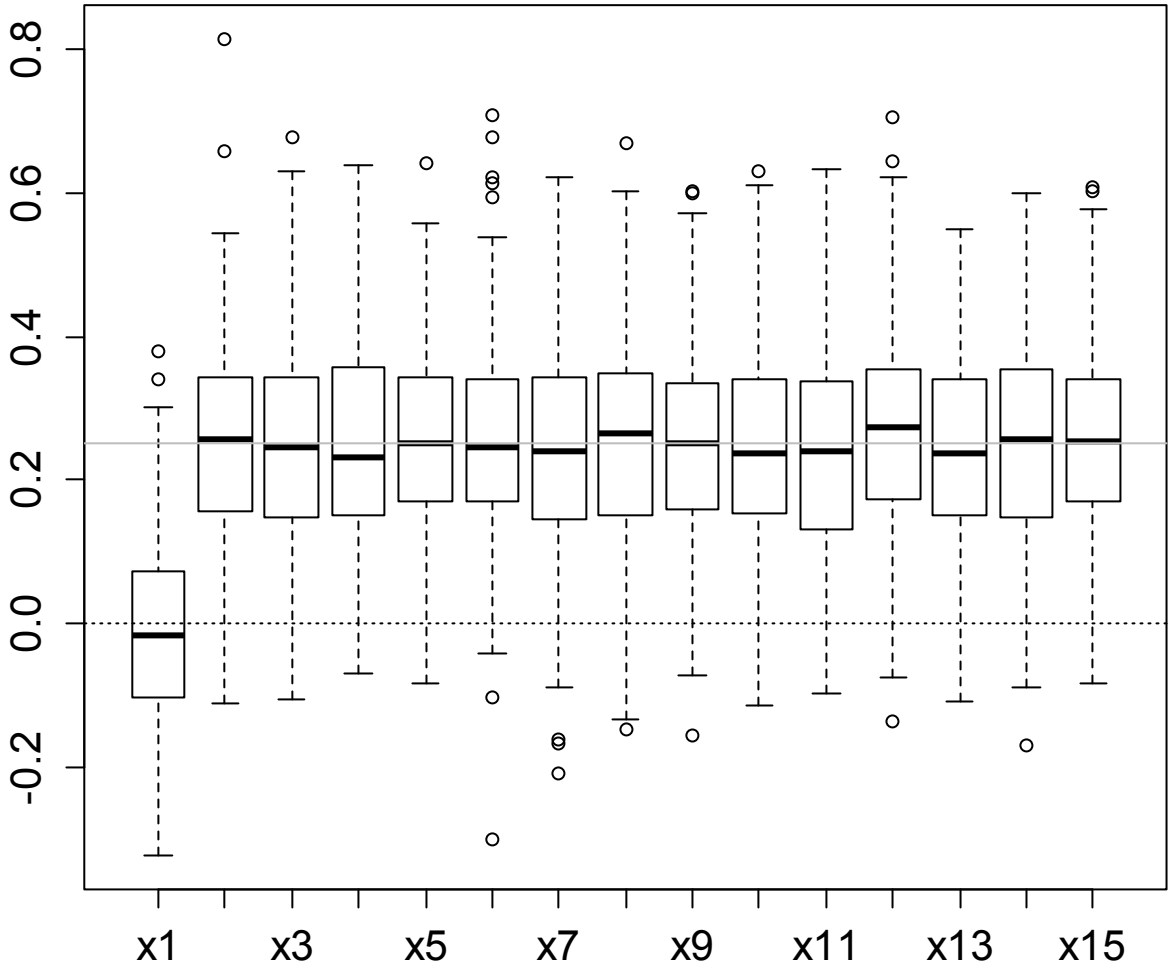
Marginal event rate 1%

True effects:

0 for X1

0.25 for X2-X15

Plot shows the betas (simulation)



Recent criticisms on Firth for prediction

- Elgmati et al (Lifetime Data Analysis 2015):

  upwards bias in predictions for rare events

- Greenland and Mansournia (Statistics in Medicine 2015):

  ‚Bayesian non-collapsibility‘ caused by including correlations in Jeffreys prior

Generalized Firth correction

- Elgmati et al studied a generalized Firth correction:

$$\log L(\beta) + \lambda \log \det(X^t W X)$$

with $\lambda \in [0,0.5]$

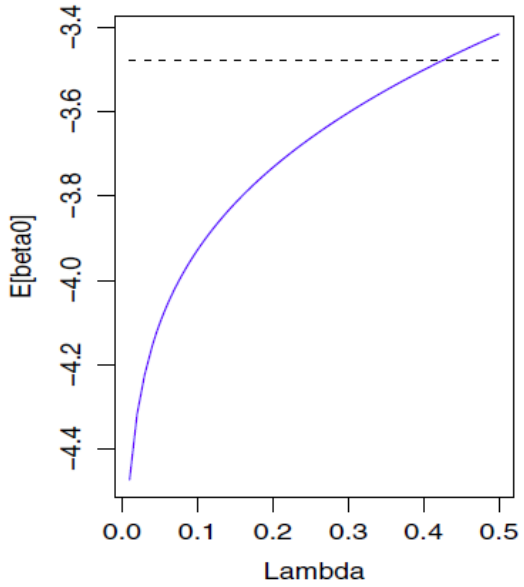
- (In the two-group case, this corresponds to adding λ to each cell.)
- They derived formulae for bias and MSE of predicted probabilities in the two-group case, and evaluated the impact of λ.

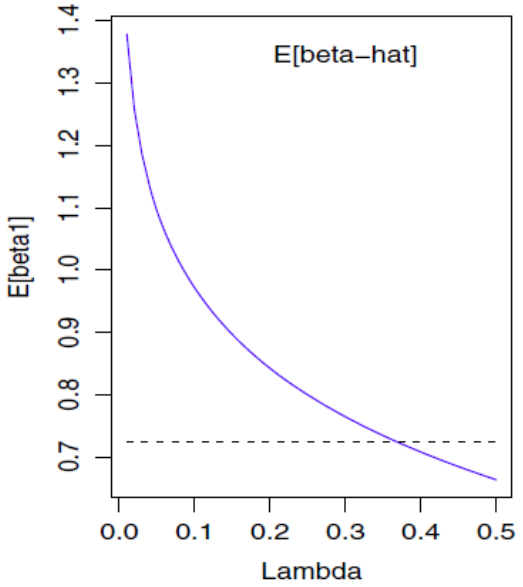
Generalized Firth correction

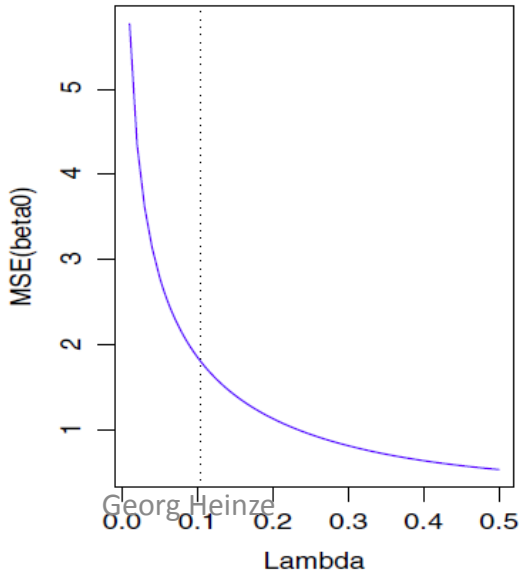
From Elgmati et al, 2015

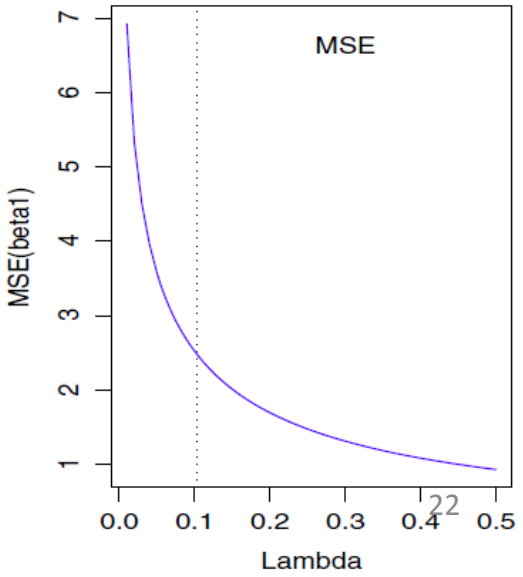
Two group case,

- $n_0 = n_1 = 50$,
- $\pi_0 = 0.03$,
- $\pi_1 = 0.06$









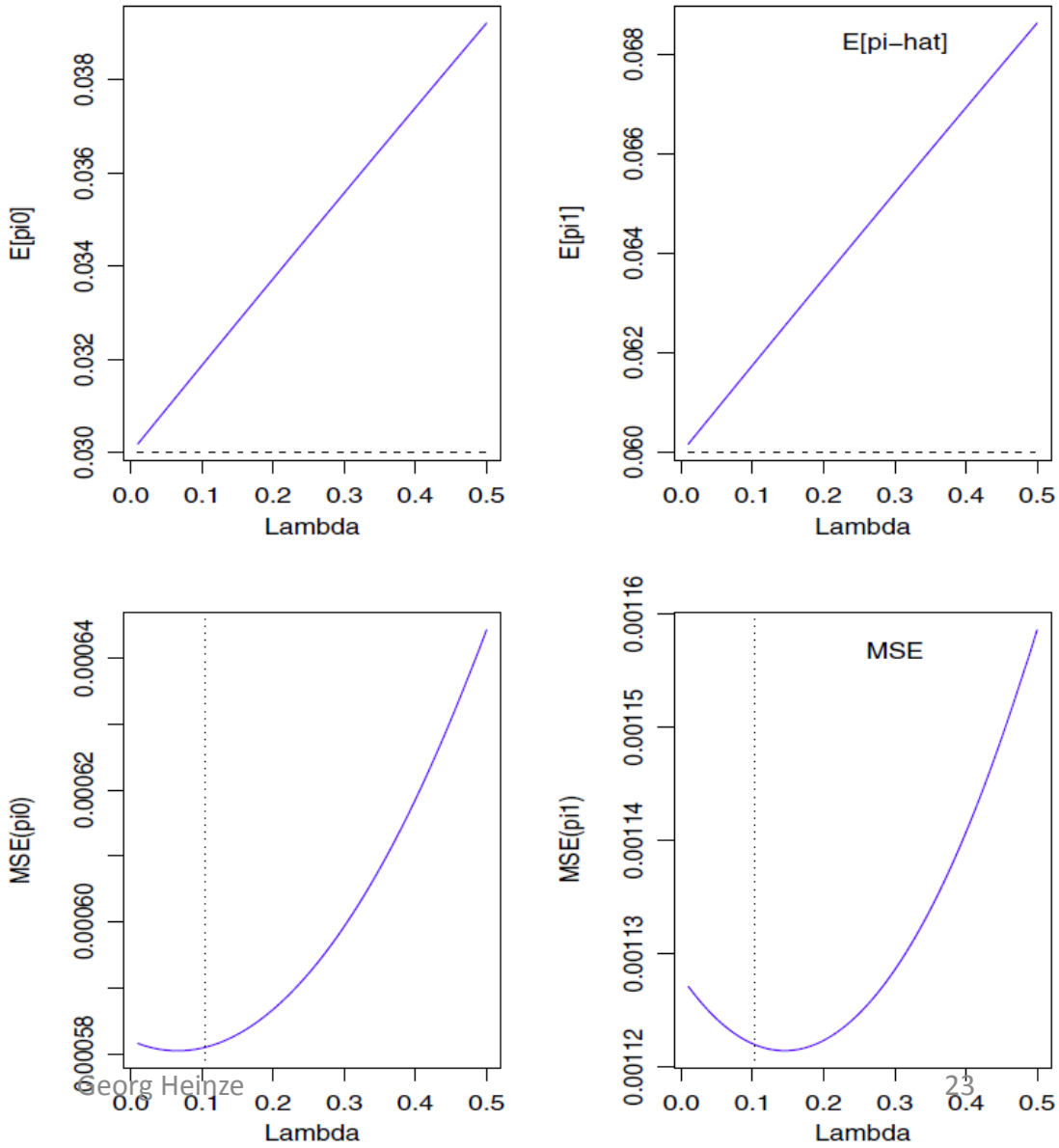
Generalized Firth correction

From Elgmati et al, 2015

Two group case,

- $n_0 = n_1 = 50$,
- $\pi_0 = 0.03$,
- $\pi_1 = 0.06$

→They suggest a weak Firth correction with $\lambda = 0.1$ to minimise MSE of predictions



Problems (of Bayesians) working with the Jeffreys prior

- Greenland and Mansournia (Statistics in Medicine 2015):

- „The Jeffreys prior (=Firth correction) is data-dependent and includes correlation between covariates“
- „This correlation is needed as also the MLE bias to be corrected has correlations“

- „The marginal prior for a given β can change in opaque ways as model covariates are added or deleted”
- „It may give surprising results in sparse data sets“

- „It is not clear in general how the penalty translates into prior probabilities for odds ratios“

Bayesian non-collapsibility

- In their data example, G&M show that the Firth-corrected estimate is further away from 0 than the ML estimate:

	X=1	X=0
Y=0	9	2966
Y=1	1	16

ML estimate of β_1:
3.03 (0.08, 4.8)

Firth estimate of β_1:
3.35 (1.07, 4.9)

The logF(1,1) prior

- Greenland and Mansournia (SiM 2015) suggest the logF(1,1) prior, leading to the penalized likelihood representation

$$\log L(\beta) + \sum_j \frac{\beta_j}{2} - \log(1 + \exp \beta_j)$$

- They show that this prior coincides with the Jeffreys prior in a one-parameter model (e.g., matched pairs case-control)
- They strongly argue against imposing a prior on the intercept

Implementation of the logF(1,1) is amazingly simple with standard software:

- Just augment the original data by adding two pseudo-observations per variable with a value of 1 for that variable, and 0‘s for all other variables (including the constant)
- The pseudo-observations have y=0 and y=1, with weights 0.5 and 0.5

e.g.	‚Constant‘	X	Y	Weight
	0	1	0	0.5
	0	1	1	0.5

Example, again

- In their data example, G&M show that the Firth-corrected estimate is further away from 0 than the ML estimate:

	X=1	X=0
Y=0	9	2966
Y=1	1	16

logF(1,1) estimate of β_1:
2.41 (-0.64, 4.4)

ML estimate of β_1:
3.03 (0.08, 4.8)

Firth estimate of β_1:
3.35 (1.07, 4.9)

Solving the issue by simulation

- By simulating 1000 2x2 tables with the expected cell frequencies (conditional on marginals of X), we obtain:

Method	Bias	RMSE
ML	n.a.	n.a.
Firth	0.18	0.81
logF(1,1)	1.10	1.63

True value = 3.03

Summary so far

- We observe that the original Firth penalty results in good bias and MSE properties for betas
- There is an upwards bias (towards 0.5) in predictions
- ‚weak' Firth penalty optimizes MSE of predictions, but induces bias in betas
- logF is simple to implement, yields unbiased mean predictions, but possibly too much correction for betas

- We would like to keep the good properties for the betas,
- but improve the performance for prediction

We called the flic

Austrian flic („Kibara“)
(unknown outside A)

French flic
(very popular in A)

FLIC: Firth Logistic regression with Intercept Correction

Consists of the Firth model,

but with adjusted intercept to fulfill $\sum y_i = \sum \hat{\pi}_i$

Two stage estimation:

- First fit Firth-corrected model
- Then hold betas fixed, but re-estimate intercept without penalization

- Corrects the bias in mean prediction

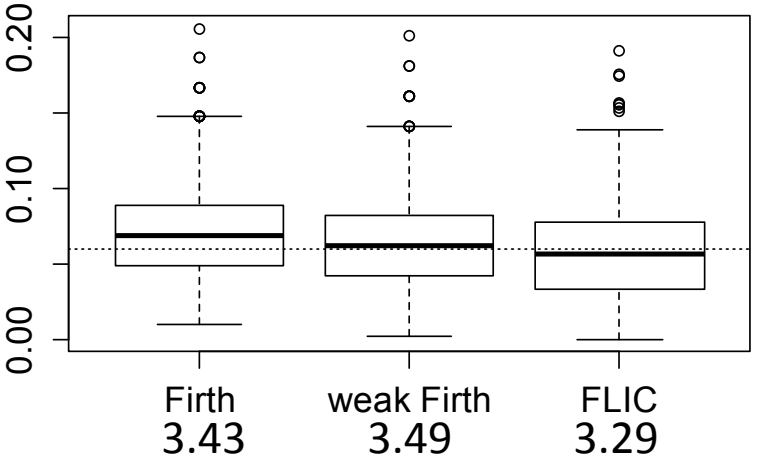
Re-simulating the example of Elgmati

From Elgmati et al, 2015

Two group case,

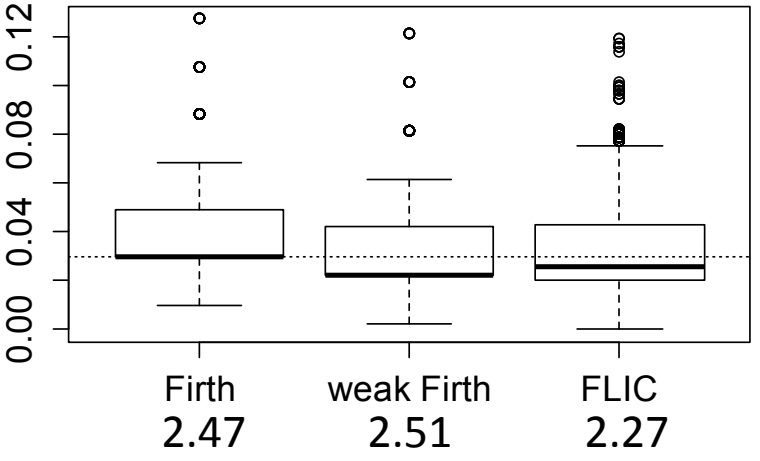
- $n_0 = n_1 = 50$

$\pi_1 = 0.06$



RMSEx100: Firth 3.43, weak Firth 3.49, FLIC 3.29

$\pi_0 = 0.03$



RMSEx100: Firth 2.47, weak Firth 2.51, FLIC 2.27

Other approaches for rare events

Considering uncertainty in estimated β coefficients, King and Zeng (2001) propose a correction of estimated probabilities:

$$\Pr(Y_i = 1) = \int \Pr(Y_i = 1|\beta^*) P(\beta^*) d\beta^*$$

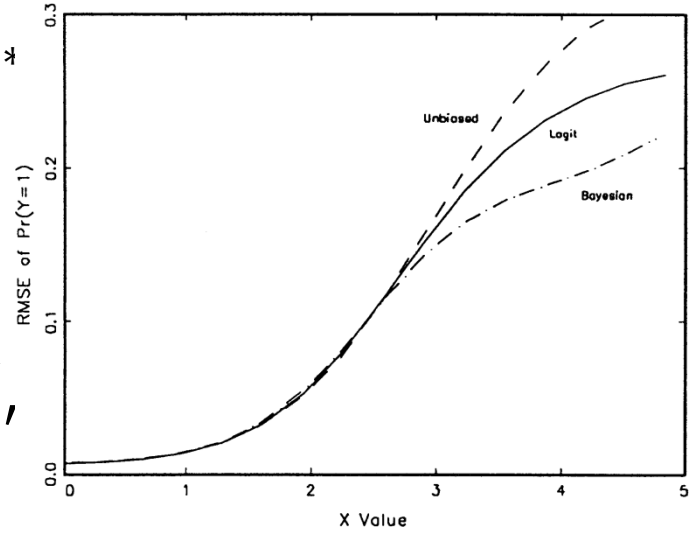
where β^* is a bias-corrected estimate of β,

and $P(\beta^*)$ is the posterior distribution of β^*

This can be approximated by

$$\Pr(Y_i = 1) \approx \widetilde{\pi}_i + C_i$$

where $C_i = (0.5 - \widetilde{\pi}_i)\widetilde{\pi}_i(1 - \widetilde{\pi}_i)x_i V(\tilde{\beta})x_i'$



A simulation study

X matrix

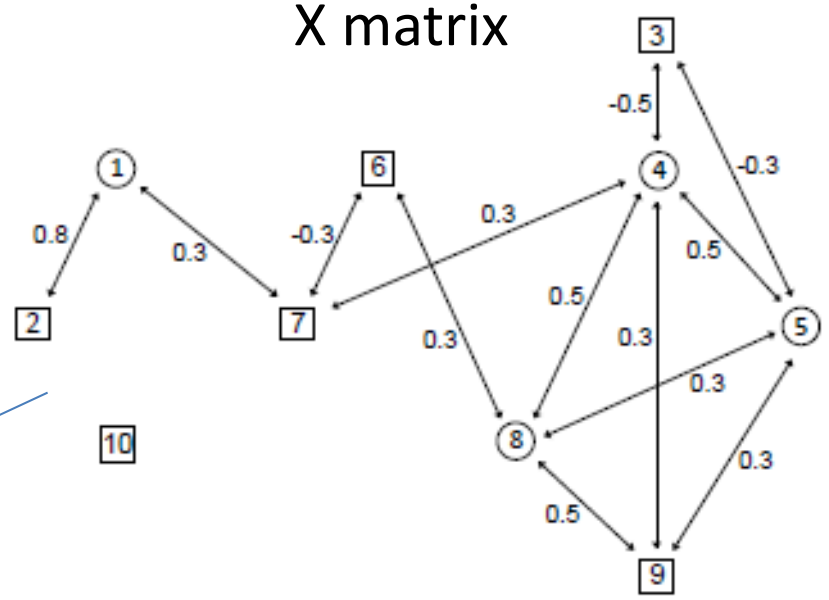


Figure 2: Partial correlation structure of the underlying variables, rectangles indicating underlying variables of metric variables, circles that of categorial.

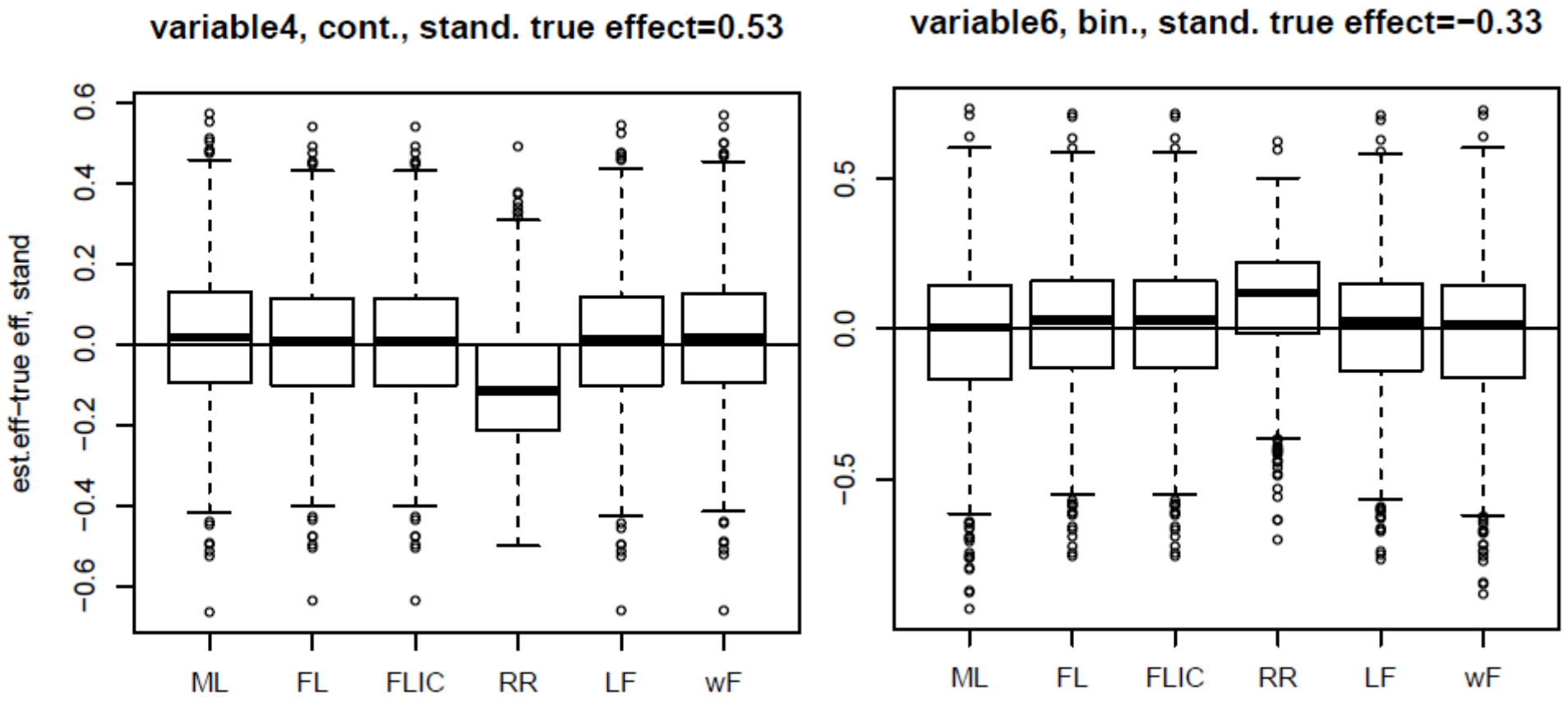
N (500, 1400, 3000)

Y (1-10%)

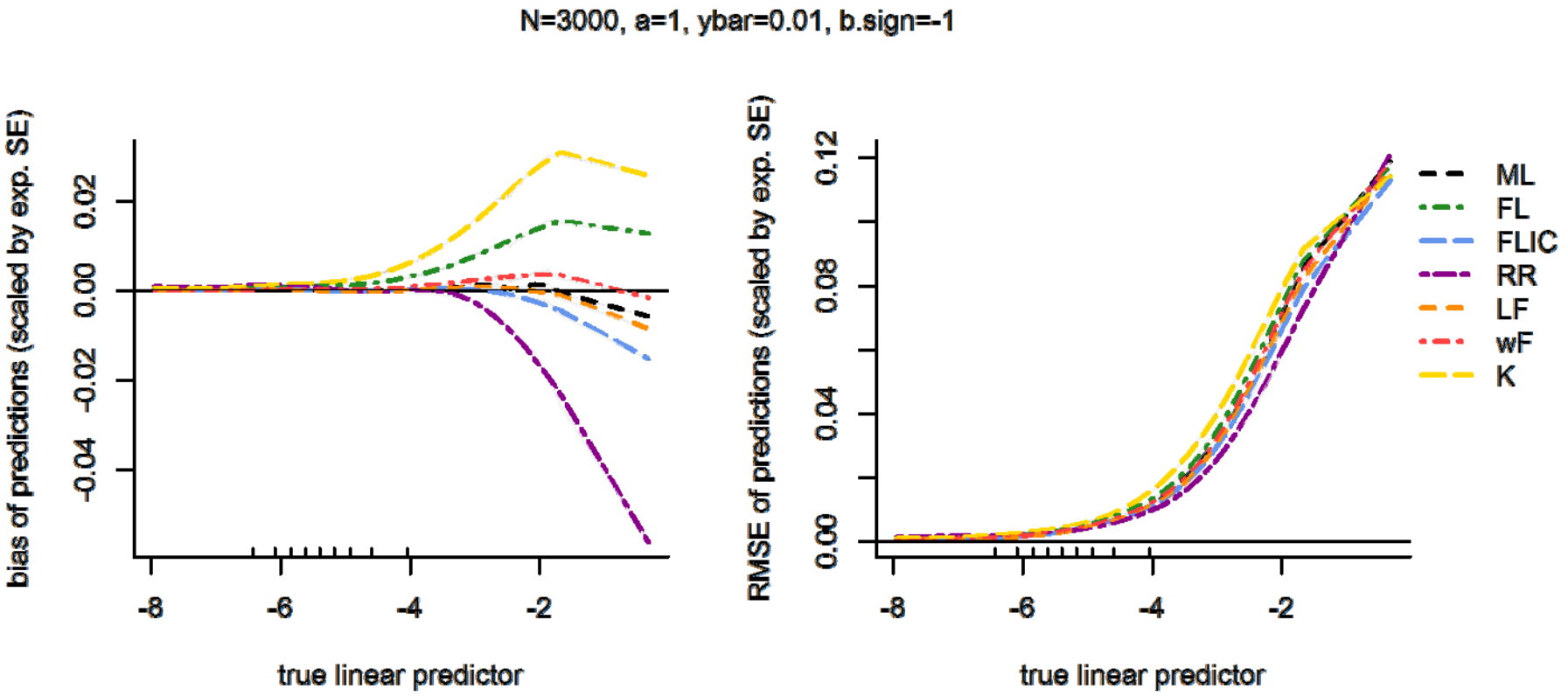
Puhr, 2015; Binder et al, 2011

Scenario with N=3000, 1% event rate

Bias of beta estimates

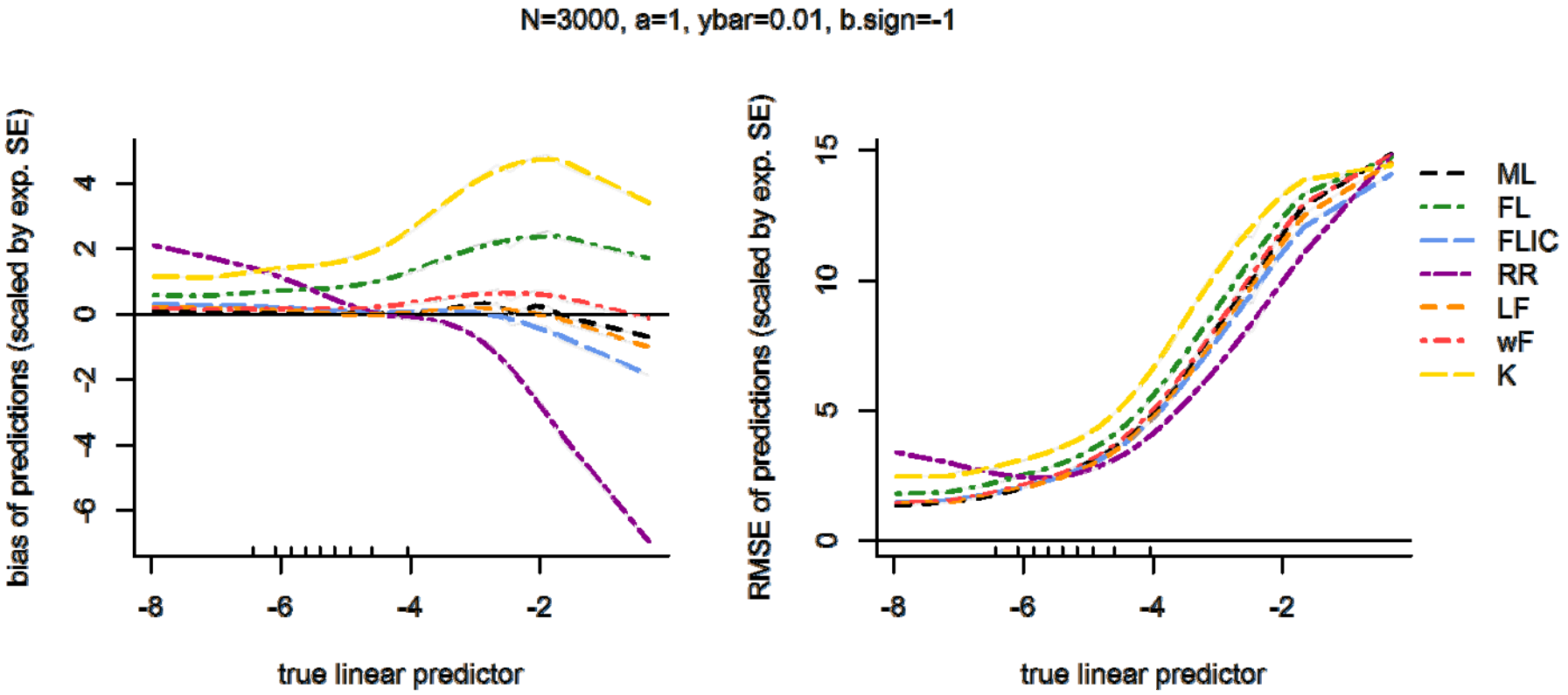


N=3000, 1% event rate, bias of predictions



N=3000, 1% event rate, bias of predictions

- Rescaled by expected standard error of predictions



Conclusions

- Prediction and effect estimation

- Ridge models perform best for prediction (RMSE but not bias), but should be seen as black boxes
- Always ‚on the safe side': sometimes overly pessimistic

- Among the less conservative methods, FLIC performed well
- It does not sacrifice the bias-preventive properties of Firth

References

- Binder H, Sauerbrei W, Royston P. Multivariable Model-Building with Continuous Covariates: Performance Measures and Simulation Design. Unpublished Manuscript, 2011.
- Elgmati E, Fiaccone RL, Henderson R, Matthews JNS. Penalised logistic regression and dynamic prediction for discrete-time recurrent event data. *Lifetime Data Analysis* 21:542-560, 2015
- Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 80:27-38, 1993
- Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33:1, 2010
- Greenland S, Mansournia M. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine* 34:3133-3143, 2015
- Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21:2409-2419, 2002
- Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine* 25:4216-4226, 2006
- Heinze G, Ploner M, Beyea J. Confidence intervals after multiple imputation: combining profile likelihood information from logistic regressions. *Statistics in Medicine* 32:5062-5076, 2013
- King G, Zeng L. Logistic Regression in Rare Events Data. *Political Analysis* 9:137-163, 2001.
- Puhr R. Vorhersage von seltenen Ereignissen mit pönalisierter logistischer Regression. Master‘s Thesis, University of Vienna, 2015.