Prediction and explanation in studies with rare events: problems and solutions

Georg Heinze

Medical University of Vienna

Supported by the Austrian Science Fund FWF (I2276-N33)

Freiburg, 26 November 2018

Georg.heinze@meduniwien.ac.at @Georg__Heinze http://prema.mf.uni-lj.si http://cemsiis.meduniwien.ac.at/en/kb

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

- Statistical models
  - for prediction of binary outcomes
  - should be interpretable,

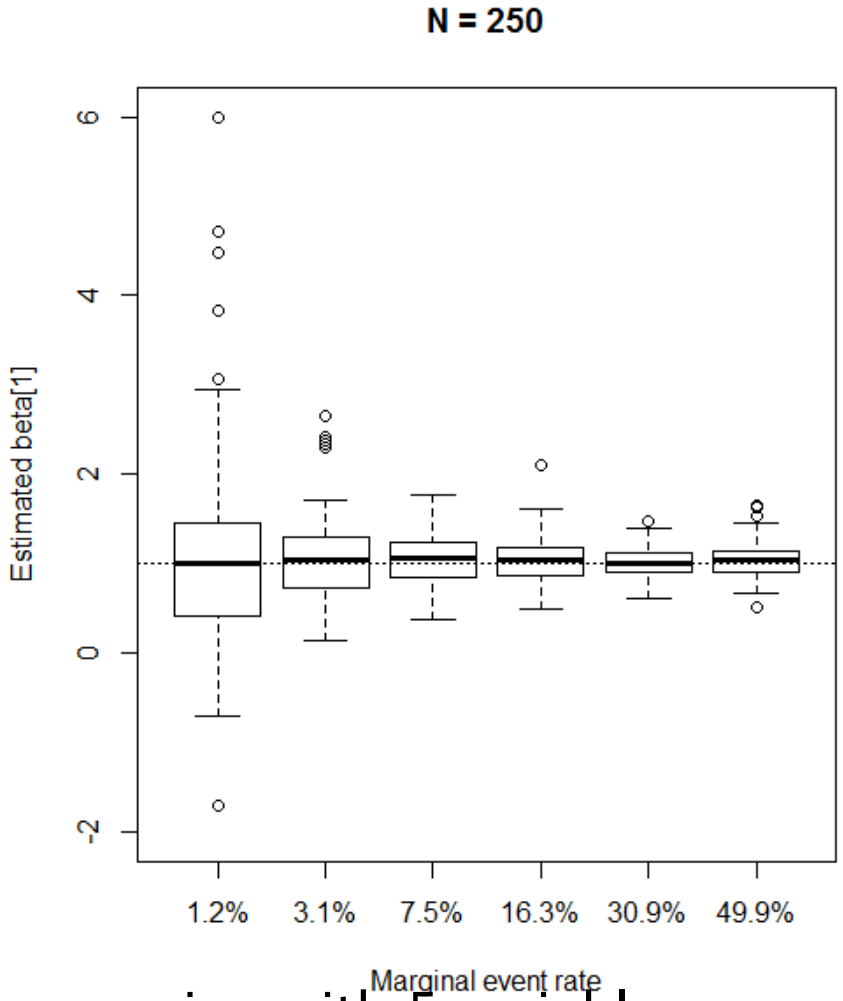
    i.e., ‚betas' should have a meaning

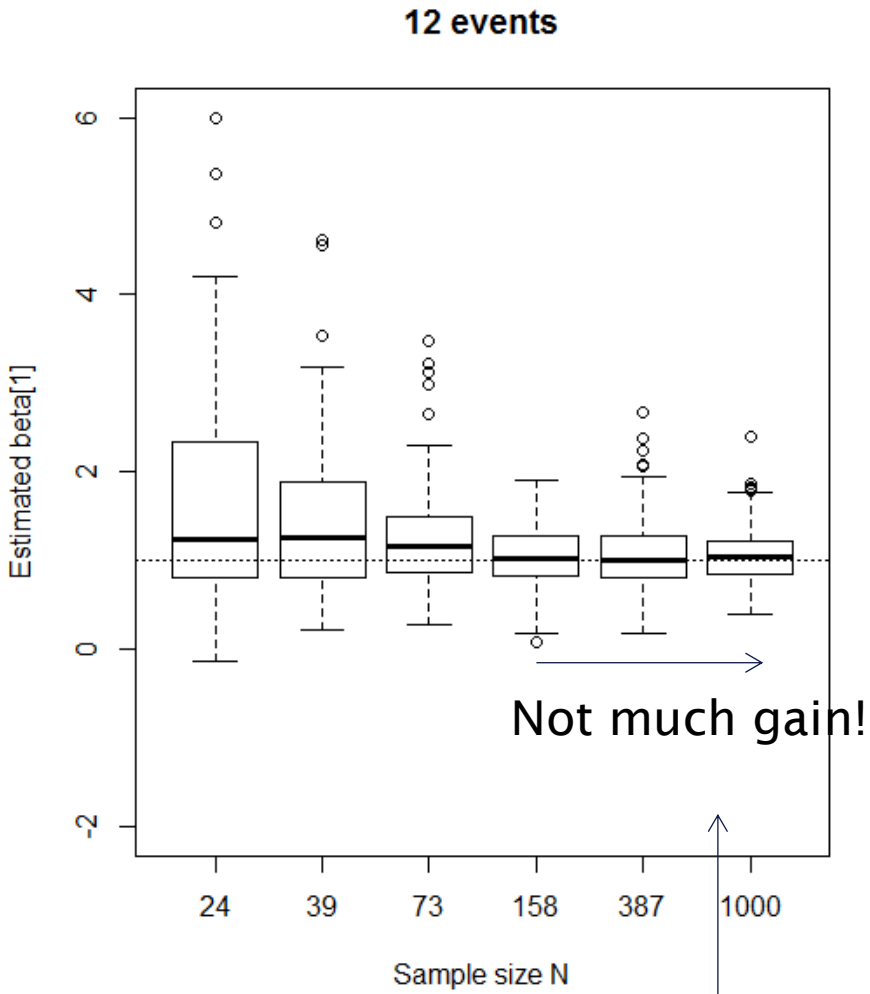
    → explanatory models based on logistic regression

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

- How well can we estimate β if events $(y_i = 1)$ are rare?
- How well can we predict Y if π is not ‚average'?

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

More rare events problems: separation

- From Mansournia et al, AmJEpi 2018:

Table 1. Diaphragm Use and Urinary Tract Infection in the Data Reported by Foxman et al.[a], 1997

Diaphragm Use	Urinary Tract Infection	
	Yes	No
Yes	7	0
No	140	290

[a] Foxman et al. (9).

- Odds ratio?

Separation and the ‚monotone likelihood'

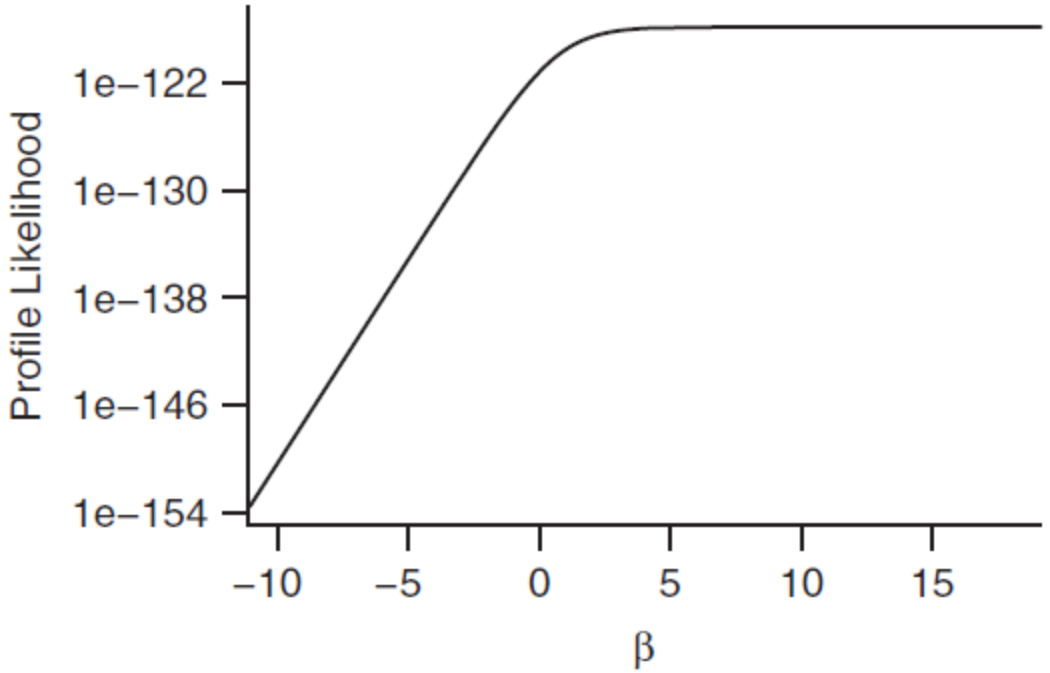


Figure 2. Profile likelihood (on logarithmic scale) for the log odds ratio β of diaphragm use in univariate analysis of the data from Foxman et al. (9), 1997. For each value of β, the profile likelihood is obtained by maximizing the likelihood as a function of the intercept given β.

Mansournia et al, AmJEpi 2018

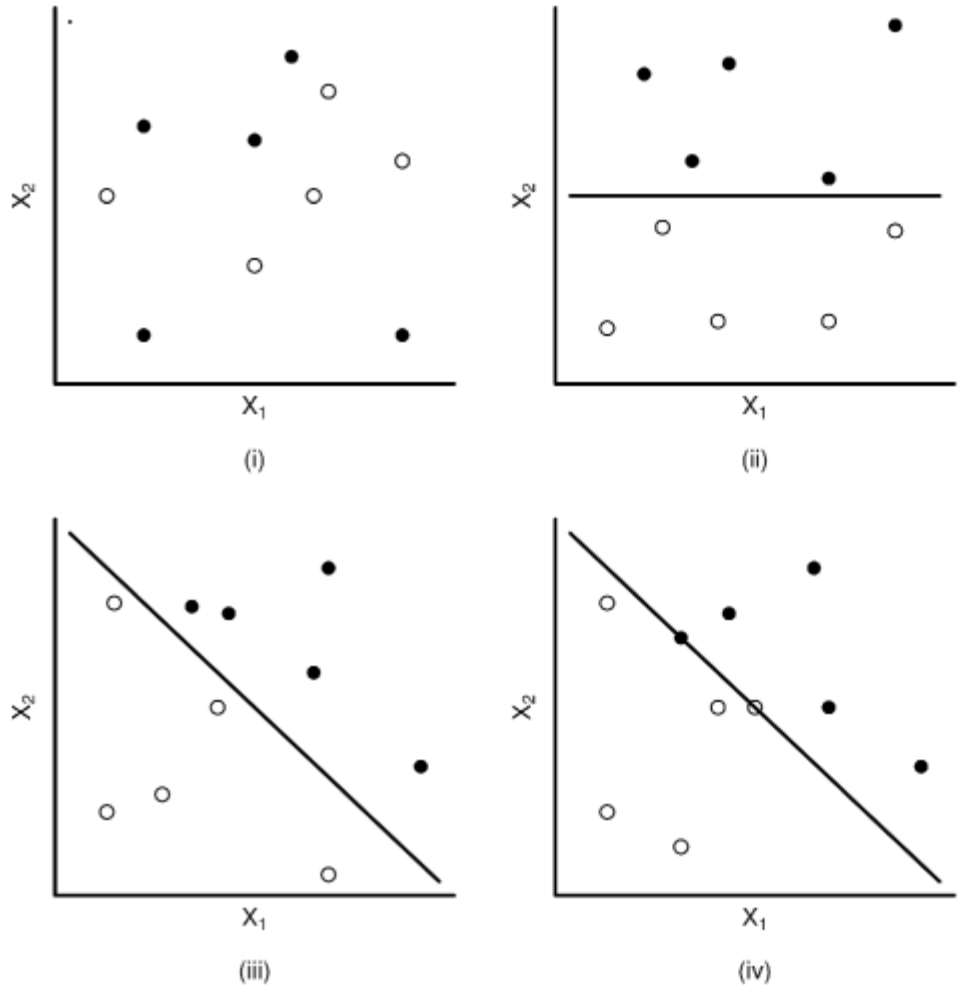


Fig. 1
Graphical representation of separation (complete and quasi-complete) adapted from Albert and Anderson [16]. Sample points for two variables X_1 and X_2 by outcome (Y): open and filled *circles* represent different levels of the outcome (Y=0 or 1). **(i)** No separation; **(ii)** complete separation by variable X_2; **(iii)** complete separation by variables X_1 and X_2; **(iv)** quasi-complete separation by variable X_1 and X_2

Van Smeden et al, BMC Med Res Meth 2016

Mansournia et al,
AmJEpi 2018

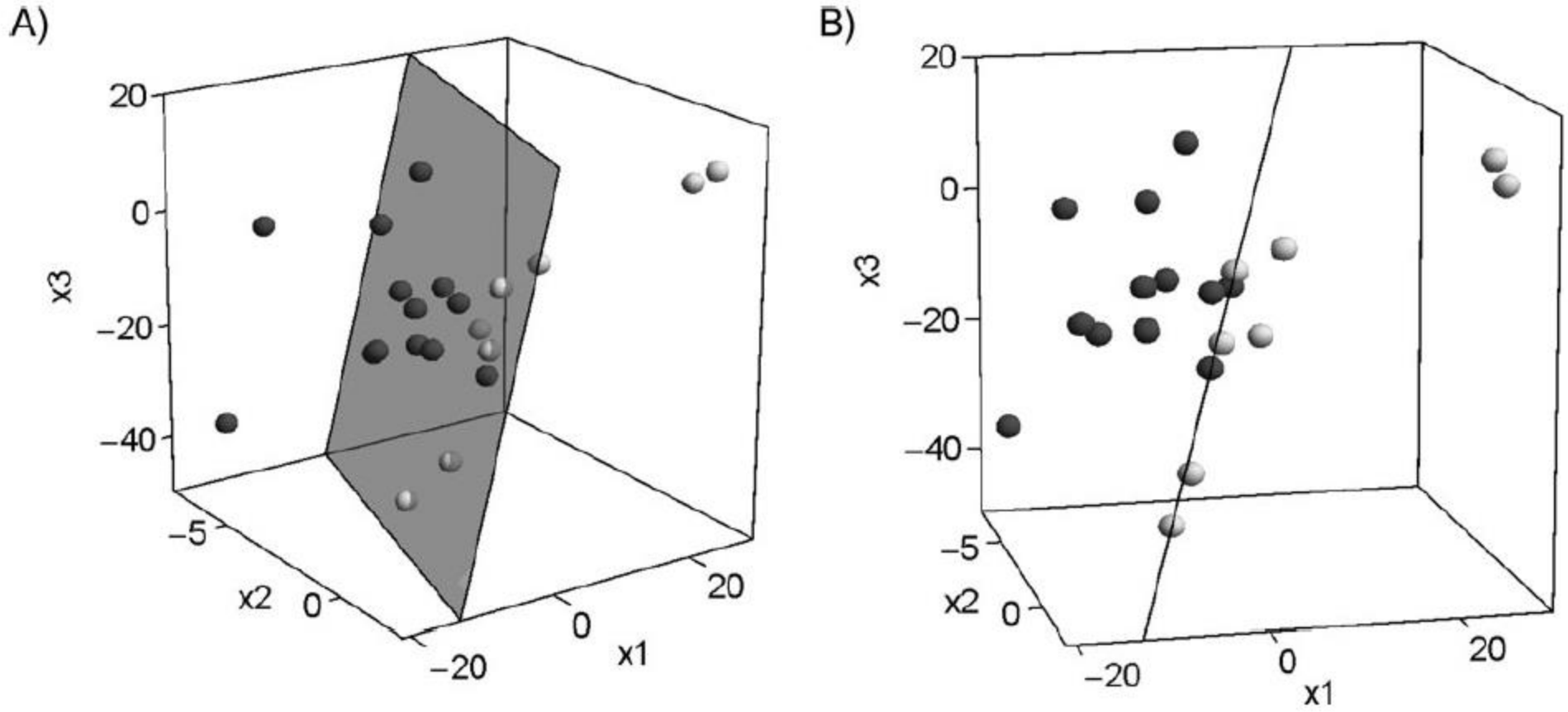


Figure 1. Illustration of data separation for the data from Potter (11), 2005. The axes correspond to the 3 covariates. Treatment success is marked in black and failure in gray. Plots (A) and (B) differ only in the angle of view. The data are an example of quasicomplete separation (i.e., there is a plane (with equation $-112.3x_1 - 165.3x_2 + 21.02x_3 = 5.4$) that separates data points with different outcomes but with observations of both outcomes lying exactly on the plane).

Why a solution is needed

- It is not assumed that event ‚cannot occur' in one of the categories of X
- We just need a bigger sample size
- Hence, the seemingly ∞ odds ratio is a small sample problem
- Normal approximation completely fails: Wald CI for β diverges to $-\infty, +\infty$ (this could be seen as a sign of variance inflation)
- Better are profile likelihood CI, but anticonservative (Heinze, StatMed 2006)

- Methods to correct the higher-level problem of small samples may also work to tackle separation issue

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‘s penalization for logistic regression

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

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

where $I(\beta)$ is the Fisher information matrix and $L(\beta)$ is the likelihood.

Firth-type penalization

- **removes the first-order bias** of the ML-estimates of β,
- is **bias-preventive** rather than corrective,
- is available in **Software** packages such as SAS, R, Stata…

Firth's penalization for logistic regression

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

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

Jeffreys invariant prior

where $I(\beta)$ is the Fisher information matrix and $L(\beta)$ is the likelihood.

Firth-type penalization

- **removes the first-order bias** of the ML-estimates of β,
- is **bias-preventive** rather than corrective,
- is available in **Software** packages such as SAS, R, Stata…

Firth's penalization for logistic regression

In logistic regression, the penalized likelihood is given by

$$L^*(\beta) = L(\beta)\det(X^t W X)^{1/2}, \text{ with}$$

$$W = \text{diag}(\text{expit}(\text{X}_\text{i}\beta)(1 - \text{expit}(\text{X}_\text{i}\beta)))$$
$$= \text{diag}(\pi_i(1-\pi_i))\ .$$

- Firth-type estimates always exist.

  W is maximised at $\pi_i = \frac{1}{2}$, i.e. at $\beta = 0$, thus

- predictions are usually pulled towards $\frac{1}{2}$,
- coefficients towards zero.

Shrinkage!

Firth‘s penalization for logistic regression

Bias reduction also leads to reduction in MSE:

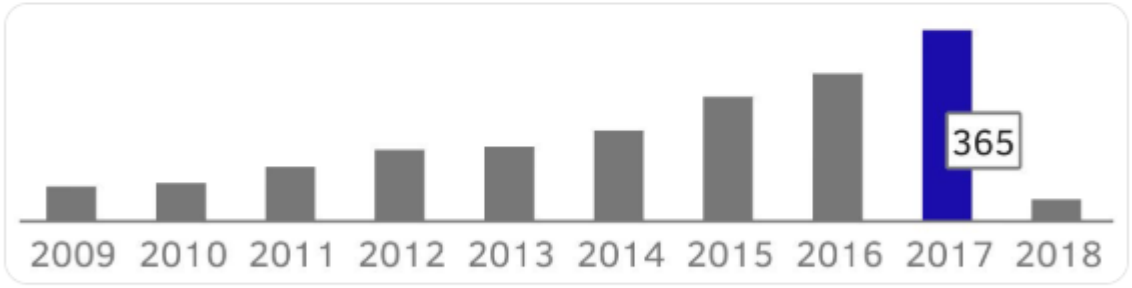
- van Smeden, 2016: ‚By applying Firth’s correction, the problems associated with separation can be avoided.’
  ‘Our simulation study shows that this performance at low values of EPV can be significantly improved using Firth’s correction.‘
  ‚We further show that Firth’s correction can be used to improve the accuracy of regression coefficients and alleviate the problems associated with separation.‘
- Rainey, 2017: Simulation study of LogReg for political science
  ‚Firth‘s methods dominates ML in bias and MSE‘

David Firth
@firthcity

Folge ich

This nice post twitter.com/MaartenvSmeden ... by @MaartenvSmeden prompted me to look at the citation data for that old 1993 paper. Google Scholar currently showing 365 cites in 2017 -- a beautiful number when expressed per day!

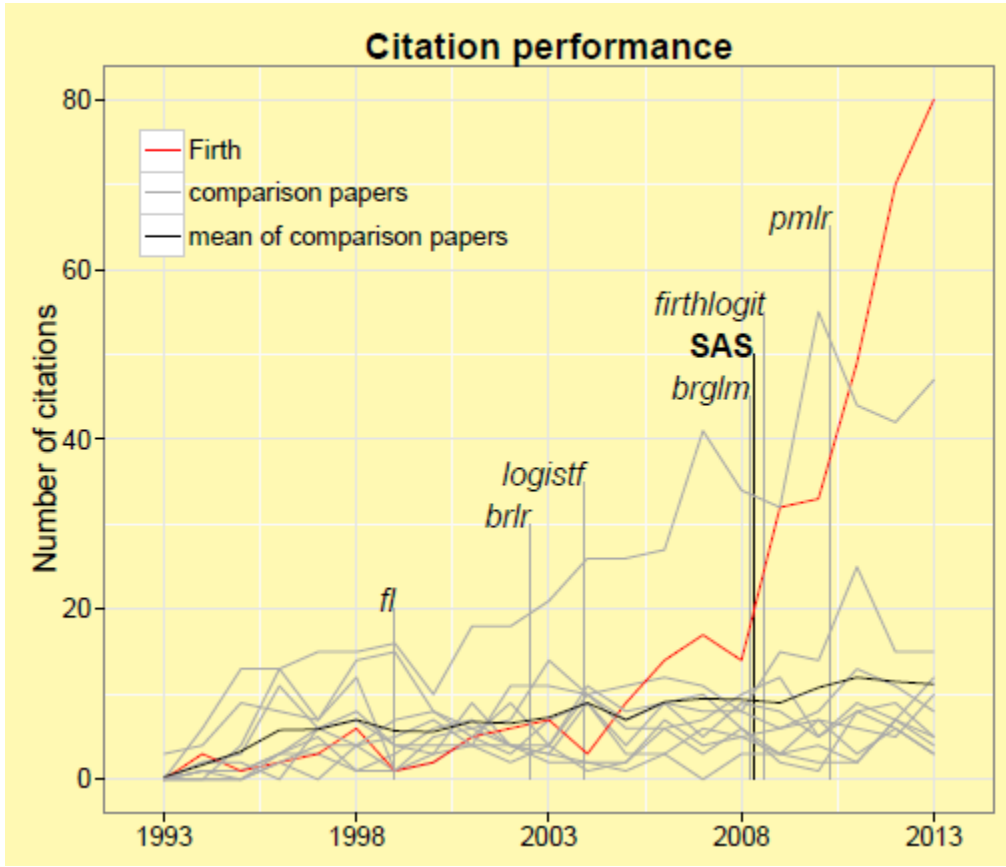
Tweet übersetzen



03:08 - 18. Feb. 2018

3 Retweets **6** „Gefällt mir"-Angaben

3 6



Firth‘s penalization for logistic regression

So, let‘s forget about maximum likelihood for logistic regression and use Firth‘s method throughout?

Firth‘s penalization for logistic regression

So, let‘s forget about maximum likelihood for logistic regression and use Firth‘s method throughout?

Well, but …

the predictions get biased

- Elgmati et al, 2015

… and anti-shrinkage could occasionally arise:

- Greenland and Mansournia, 2015

Firth's Logistic regression

For logistic regression with one binary regressor*,
Firth's bias correction amounts to adding $1/2$ to each cell:

original

	A	B
Y=0	44	4
Y=1	1	1

Firth-type penalization →

augmented

	A	B
0	44.5	4.5
1	1.5	1.5

event rate $= \frac{2}{50} = 0.04$

$OR_{BvsA} = 11$

event rate $= \frac{3}{52} \sim 0.058$

$OR_{BvsA} = 9.89$

av. pred. prob. $= 0.054$

* Generally: for saturated models

Correcting the bias in $\hat{\pi}$: FLIC

Firth's **L**ogistic regression with **I**ntercept **C**orrection:

1. Fit a Firth logistic regression model
2. Modify the estimated intercept $\hat{\beta}_0$ such that $\bar{\hat{\pi}} = \bar{y}$.

unbiased pred. probabilities

effect estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are the same as with original Firth method

Puhr et al, 2017

Example of Greenland 2010

original

	A	B	
Y=0	315	5	*320*
Y=1	31	1	*32*
	346	*6*	*352*

augmented

	A	B	
Y=0	315.5	5.5	*321*
Y=1	31.5	1.5	*33*
	346.5	*6.5*	*354*

$$\text{event rate} = \frac{32}{352} = 0.091$$

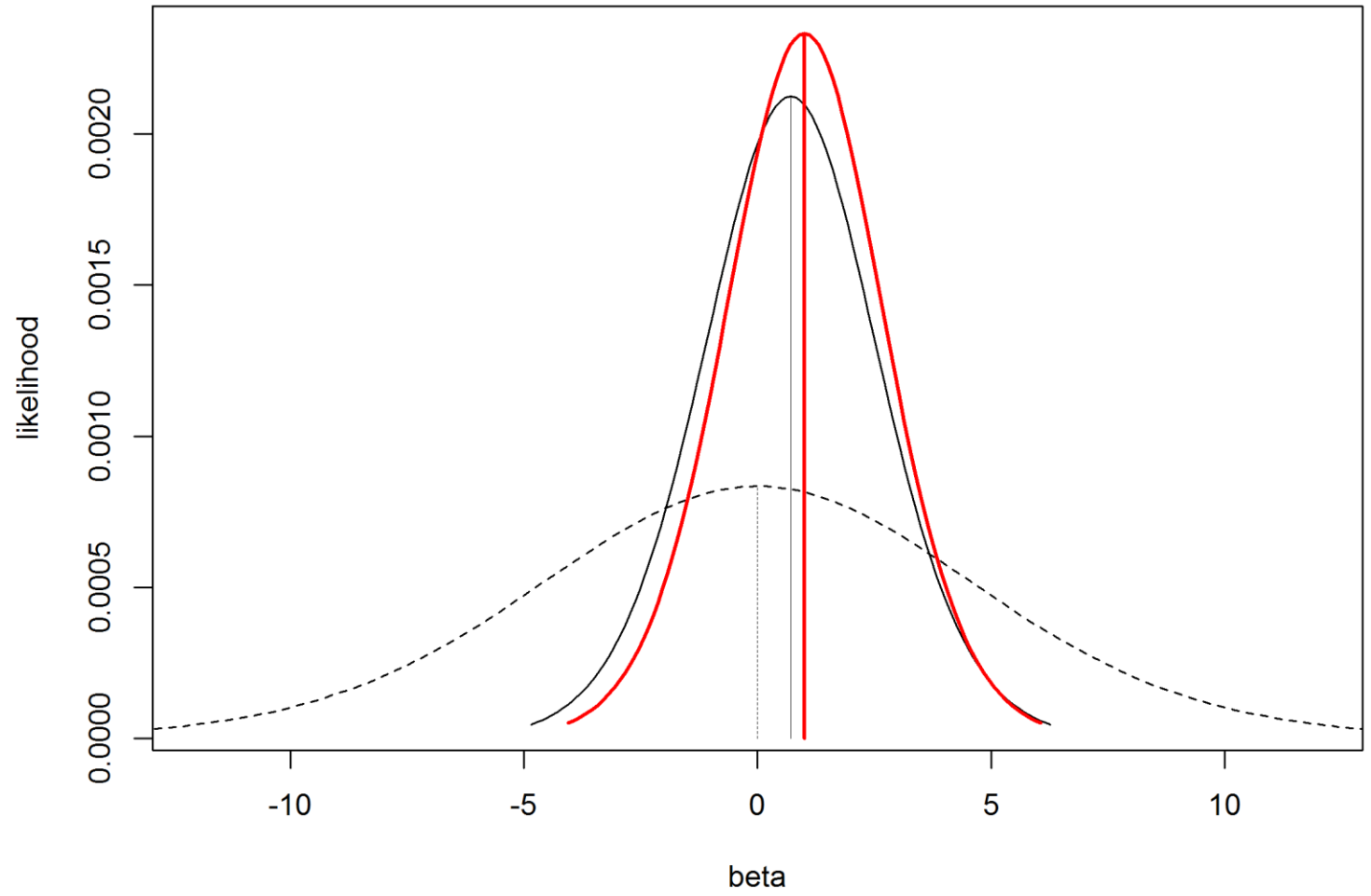
$$OR_{BvsA} = 2.03$$

$$\text{event rate} = \frac{33}{354} = 0.093$$

$$OR_{BvsA} = 2.73$$

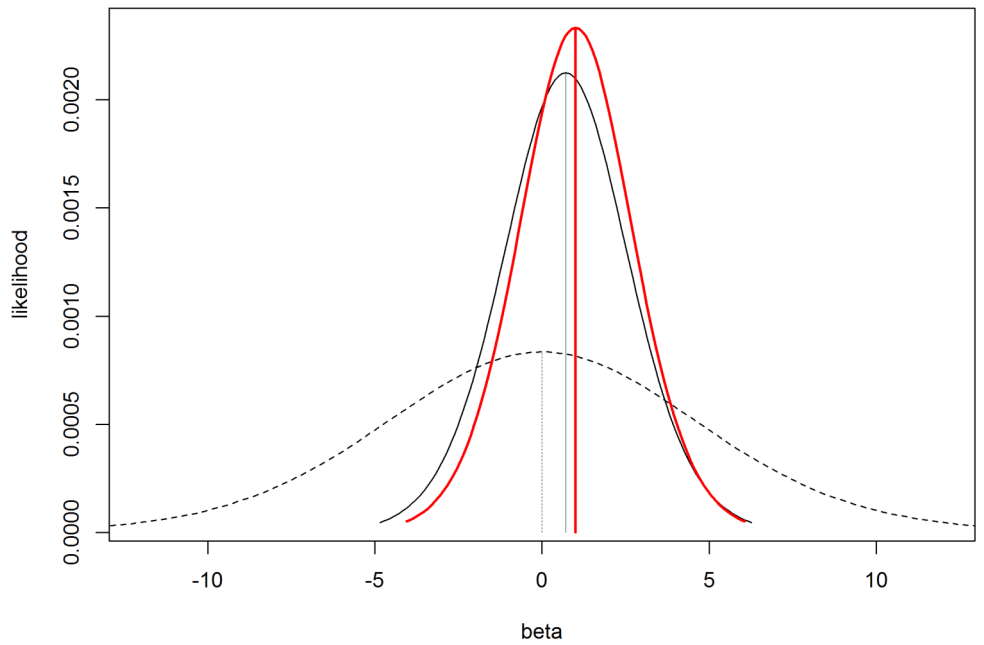
Greenland, AmStat 2010

Greenland example: likelihood, prior, posterior



Bayesian non-collapsibility: anti-shrinkage from penalization

- Prior and likelihood modes do not ‚collapse‘: posterior mode exceeds both
- The ‚shrunken‘ estimate is larger than ML estimate
- How can that happen???



An even more extreme example from Greenland 2010

- 2x2 table

	X=0	X=1	
Y=0	25	5	*30*
Y=1	5	1	*6*
	30	*6*	*36*

- Here we immediately see that the odds ratio = 1 ($\beta_1 = 0$)

- But the estimate from augmented data: odds ratio = 1.26 (try it out!)

Greenland, AmStat 2010

Reason for anti-shrinkage

- We look at the association of X and Y

- We could treat the source of data as a ‚ghost factor' G
  - G=0 for original table
  - G=1 for pseudo data

- We ignore that the conditional association of X and Y is confounded by G

Example of Greenland 2010 revisited

original

	A	B	
Y=0	315	5	*320*
Y=1	31	1	*32*
	346	*6*	*352*

augmented

	A	B	
Y=0	315.5	5.5	*321*
Y=1	31.5	1.5	*33*
	347	*7*	*352*

To overcome both the overestimation and anti-shrinkage problems:

- We propose to adjust for the confounding by including the ‚ghost factor' G in a logistic regression model

FLAC: Firth's Logistic regression with Added Covariate

Split the augmented data into original and pseudo data:

augmented

	A	B
0	315.5	5.5
1	31.5	1.5

original G=0

	A	B
0	315	5
1	31	1

+

pseudo G=1 ← Ghost factor

	A	B
0	0.5	0.5
1	0.5	0.5

OR_{BvsA} =2.03

Define **F**irth type **L**ogistic regression with **A**dditional **C**ovariate as an analysis including the ghost factor as added covariate:

OR_{BvsA} =1.84

FLAC: Firth's Logistic regression with Added Covariate

Beyond 2x2 tables:

Firth-type penalization can be obtained by solving modified score equations:

$$\sum_{i=1}^{N}(y_i - \pi_i)x_{ir} + h_i\left(\frac{1}{2} - \pi_i\right)x_{ir} = 0; \quad r = 0, \ldots, p$$

where the h_i's are the diagonal elements of the hat matrix $H = W^{\frac{1}{2}}X(X'WX)^{-1}XW^{1/2}$

They are equivalent to:

$$\sum_{i=1}^{N}(y_i - \pi_i)x_{ir} + \sum_{i}^{N} h_i\left(\frac{1}{2} - \pi_i\right)x_{ir} =$$

$$= \sum_{i=1}^{N}(y_i - \pi_i)x_{ir} + \sum_{i=1}^{N}\frac{h_i}{2}(y_i - \pi_i) + \sum_{i=1}^{N}\frac{h_i}{2}(1 - y_i - \pi_i) = 0$$

FLAC: **F**irth's **L**ogistic regression with **A**dded **C**ovariate

- A closer inspection yields:

$$\sum_{i=1}^{N}(y_i - \pi_i)x_{ir} + \sum_{i=1}^{N}\frac{h_i}{2}(y_i - \pi_i)x_{ir} + \sum_{i=1}^{N}\frac{h_i}{2}(1 - y_i - \pi_i)x_{ir} = 0$$

The original data

Original data, weighted by $h_i/2$

Data with reversed outcome, weighted by $h_i/2$

Pseudo data

FLAC: Firth's Logistic regression with Added Covariate

- A closer inspection yields:

$$\sum_{i=1}^{N}(y_i-\pi_i)x_{ir}+\sum_{i=1}^{N}\frac{h_i}{2}(y_i-\pi_i)x_{ir}+\sum_{i=1}^{N}\frac{h_i}{2}(1-y_i-\pi_i)x_{ir}=0$$

The original data

Original data, weighted by $h_i/2$

Data with reversed outcome, weighted by $h_i/2$

Pseudo data

Ghost factor: (‚Added covariate')

G=0

G=1

FLAC: **F**irth's **L**ogistic regression with **A**dded **C**ovariate

FLAC estimates can be obtained by the following steps:

1) Define an indicator variable G discriminating between original data ($G = 0$) and pseudo data ($G = 1$).
2) Apply ML on the augmented data including the indicator G in the model.

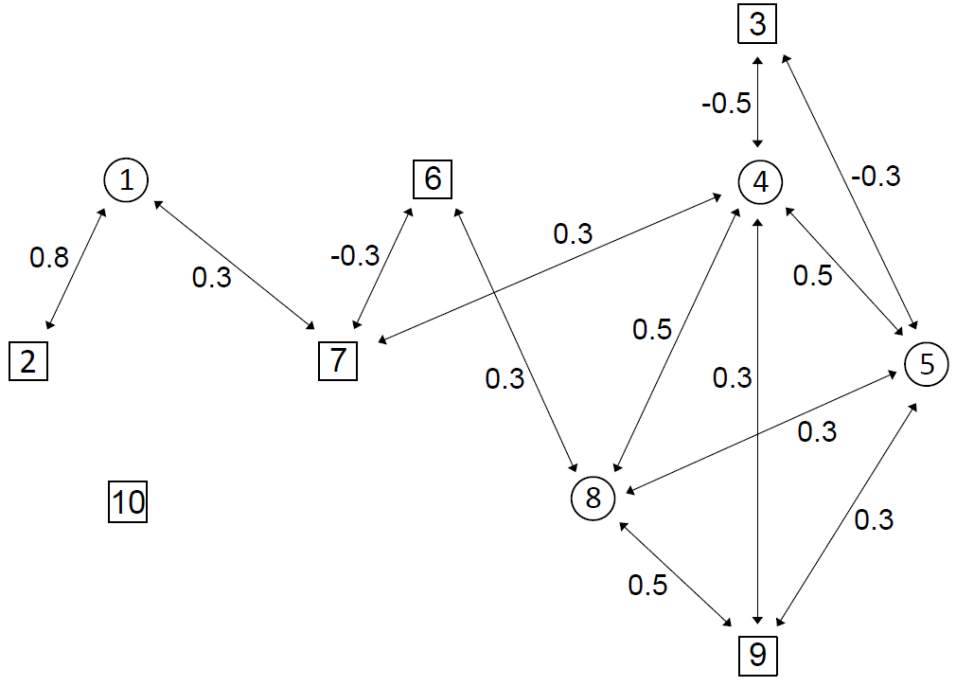
✓ unbiased pred. probabilities

Puhr et al, 2017

Simulation study: the set-up

We investigated the performance of FLIC and FLAC, simulating 1000 data sets for 45 scenarios with:

- 500, 1000 or 1400 observations,
- event rates of 1%, 2%, 5% or 10%
- 10 covariables (6 cat., 4 cont.),
  see Binder et al., 2013
- none, moderate and strong effects
  of positive and mixed signs



Main evaluation criteria:

bias and RMSE of predictions and effect estimates

Other methods for accurate prediction

In our simulation study, we compared FLIC and FLAC to the following methods:

- weakened Firth-type penalization (Elgmati 2015), with $L(\beta)^* = L(\beta)\det(X^t W X)^\tau$, $\tau = 0.1$, — WF
- ridge regression, — RR
- penalization by log-F(1,1) priors, — logF
- penalization by Cauchy priors with scale parameter=2.5. — Cauchy

logF(1,1) prior (Greenland and Mansournia, 2015)

Penalizing by log-F(1,1) prior gives $L(\beta)^* = L(\beta) \cdot \prod \frac{e^{\frac{\beta_j}{2}}}{1+e^{\beta_j}}$.

This amounts to the following modification of the data set:

	x1	x2	y
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*

	x1	x2	y
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
0	1	0	0
0	1	0	1
0	0	1	0
0	0	1	1

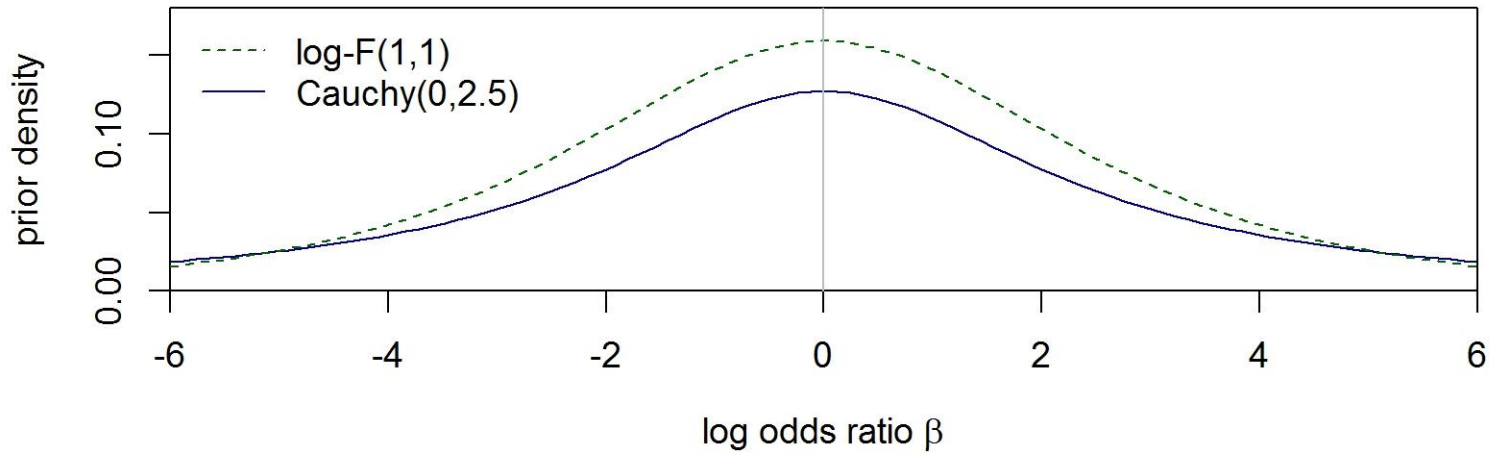
each assigned a weight of 1

each assigned a weight of ½

- No shrinkage for the intercept, no rescaling of the variables

Cauchy priors

Cauchy priors (scale=2.5) have heavier tails than log-F(1,1)-priors:



We follow Gelman 2008:

- all variables are centered,
- binary variables are coded to have a range of 1,
- all other variables are scaled to have standard deviation 0.5,
- the intercept is penalized by Cauchy(0,10).

This is implemented in the function `bayesglm` in the R-package `arm`.

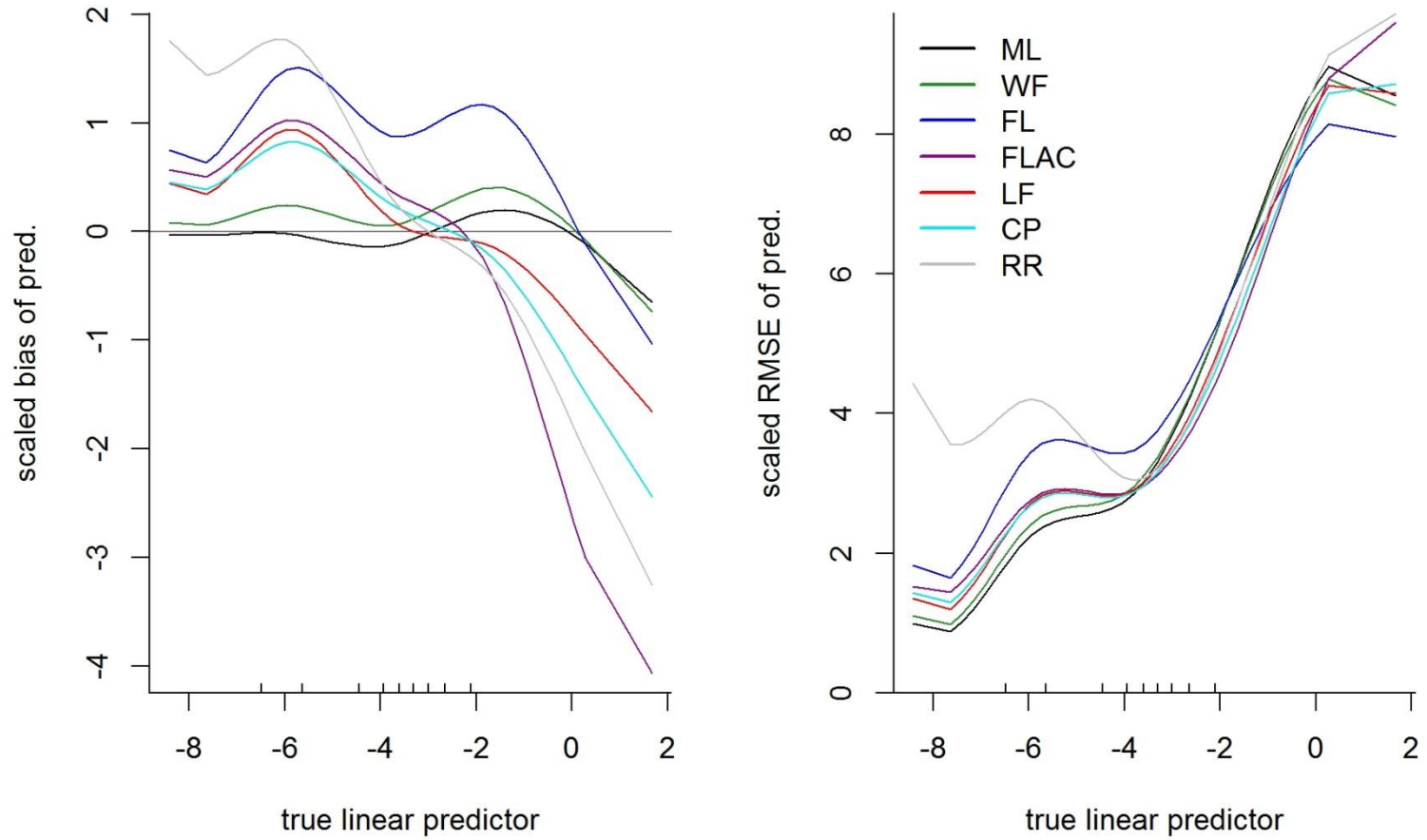
Simulation results

- Bias of $\hat{\beta}$: clear winner is Firth/FLIC method
  FLAC, logF, Cauchy: slight bias towards 0
- RMSE of $\hat{\beta}$:
  equal effect sizes: ridge the winner
  unequal effect sizes: very good performance of FLAC and Cauchy
  closely followed by logF(1,1)

- Calibration of $\hat{\pi}$:
  - often FLAC the winner
  - considerable instability of ridge

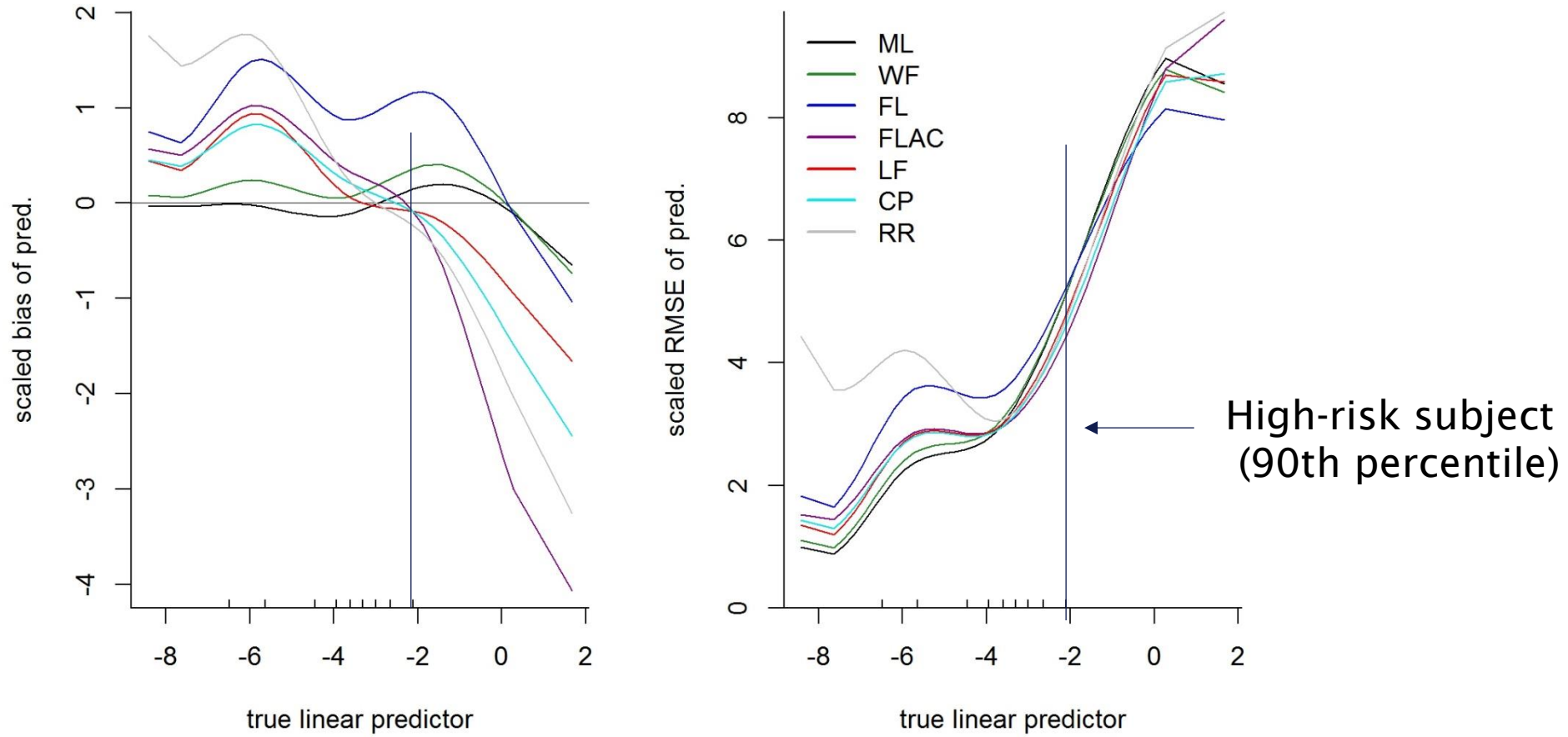
Predictions: bias RMSE

‚scaled' = in multiples of binomial error



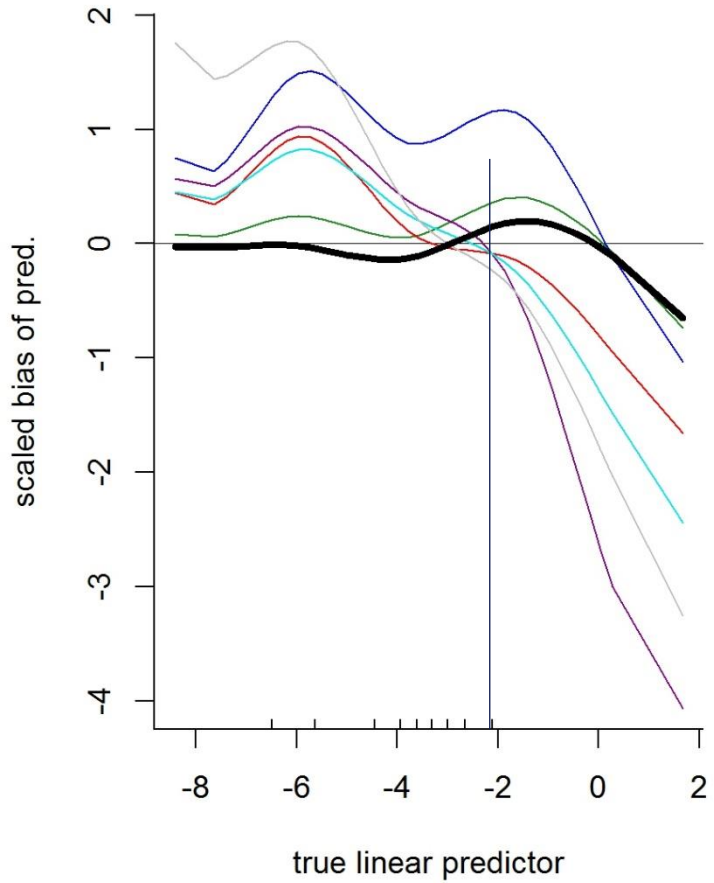
N=500, a=1, ybar=0.05, b.sign=-1

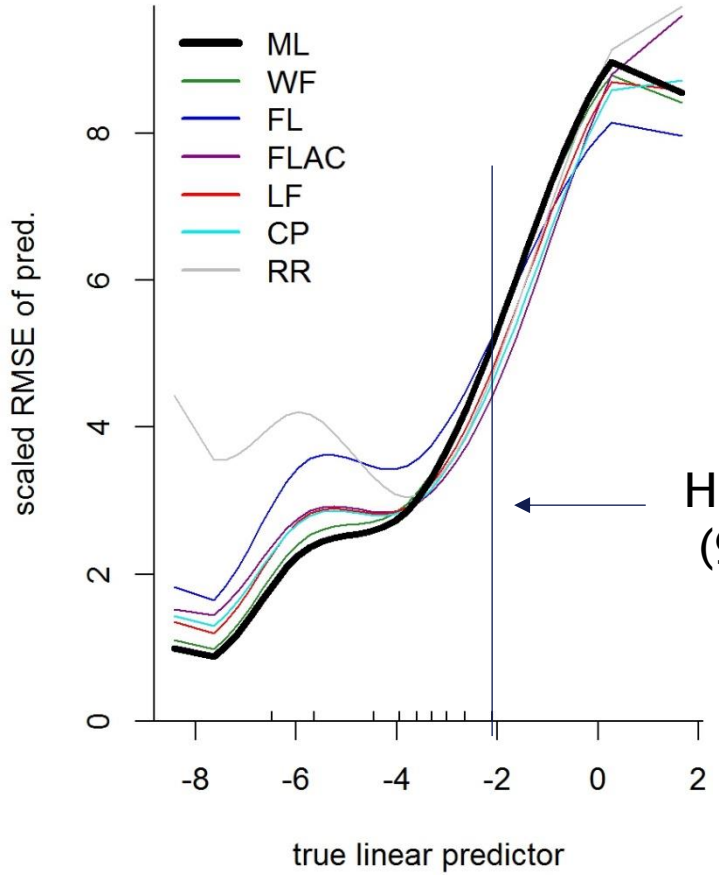
Predictions: bias RMSE



N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE

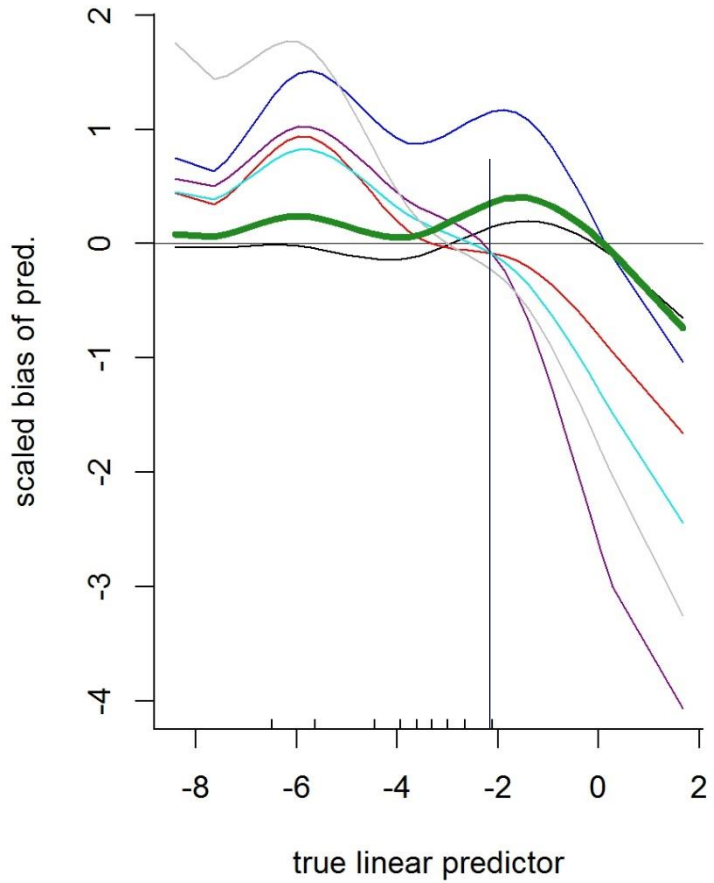


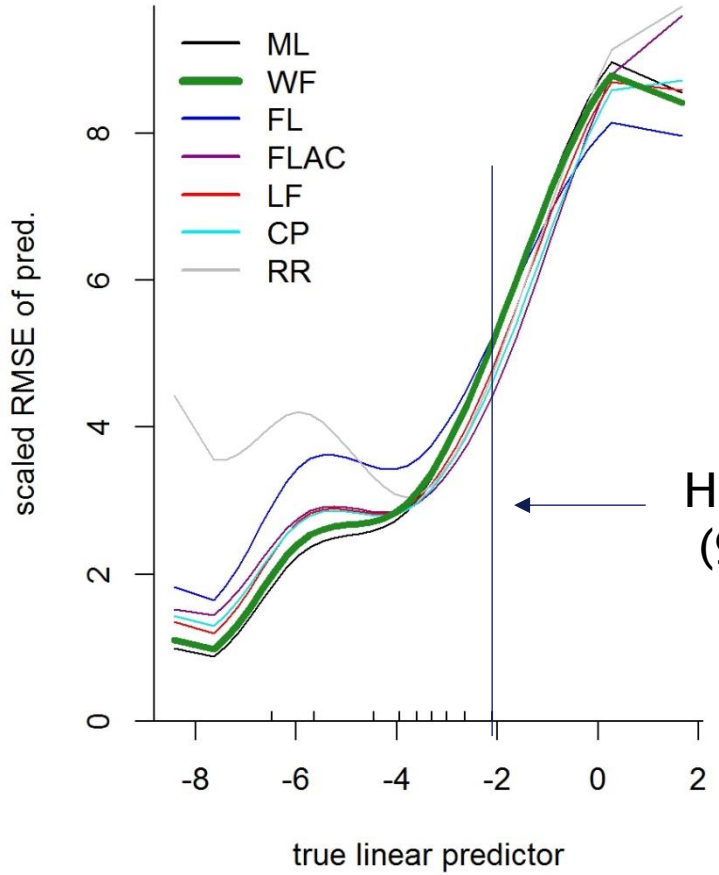


High-risk subject
(90th percentile)

N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE

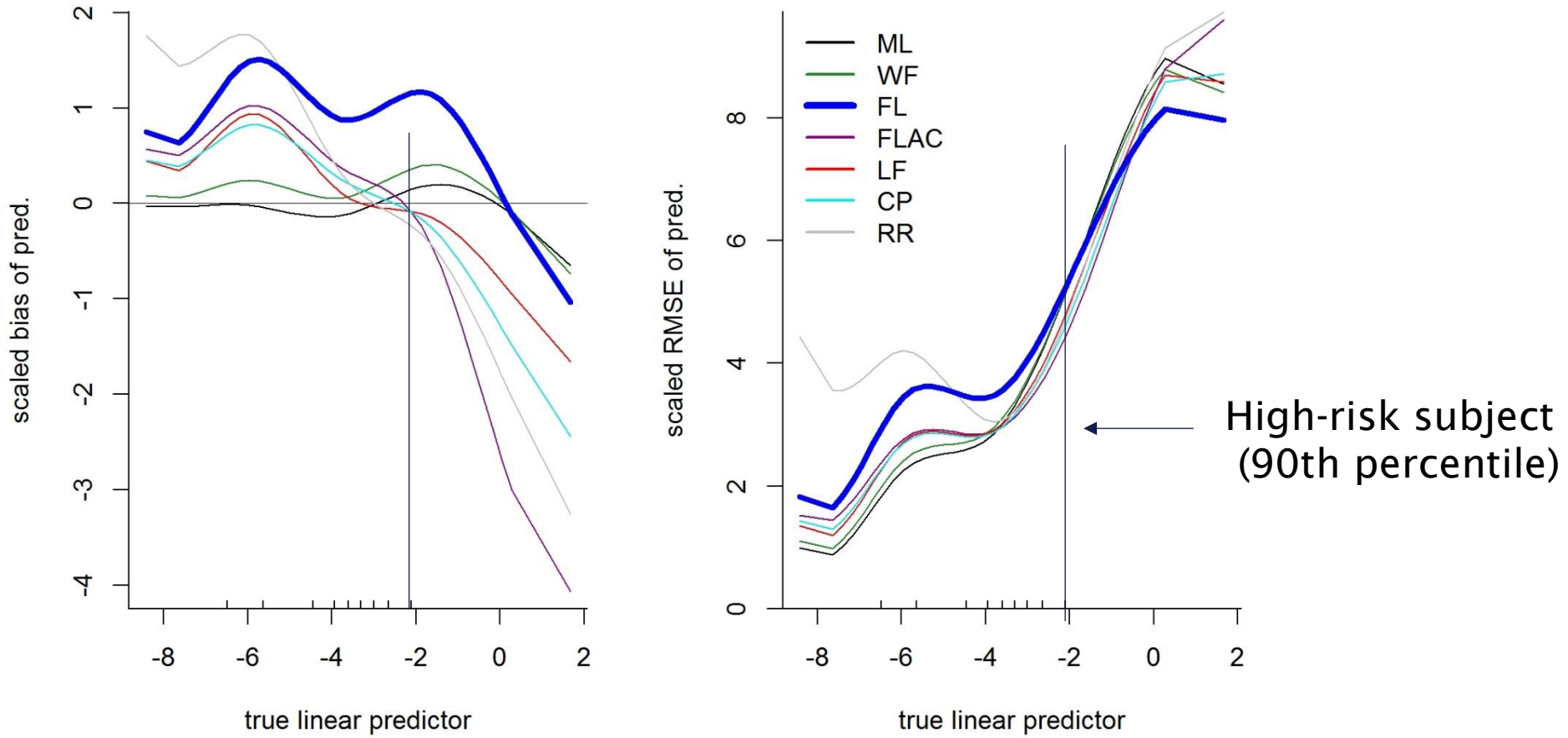




High-risk subject
(90th percentile)

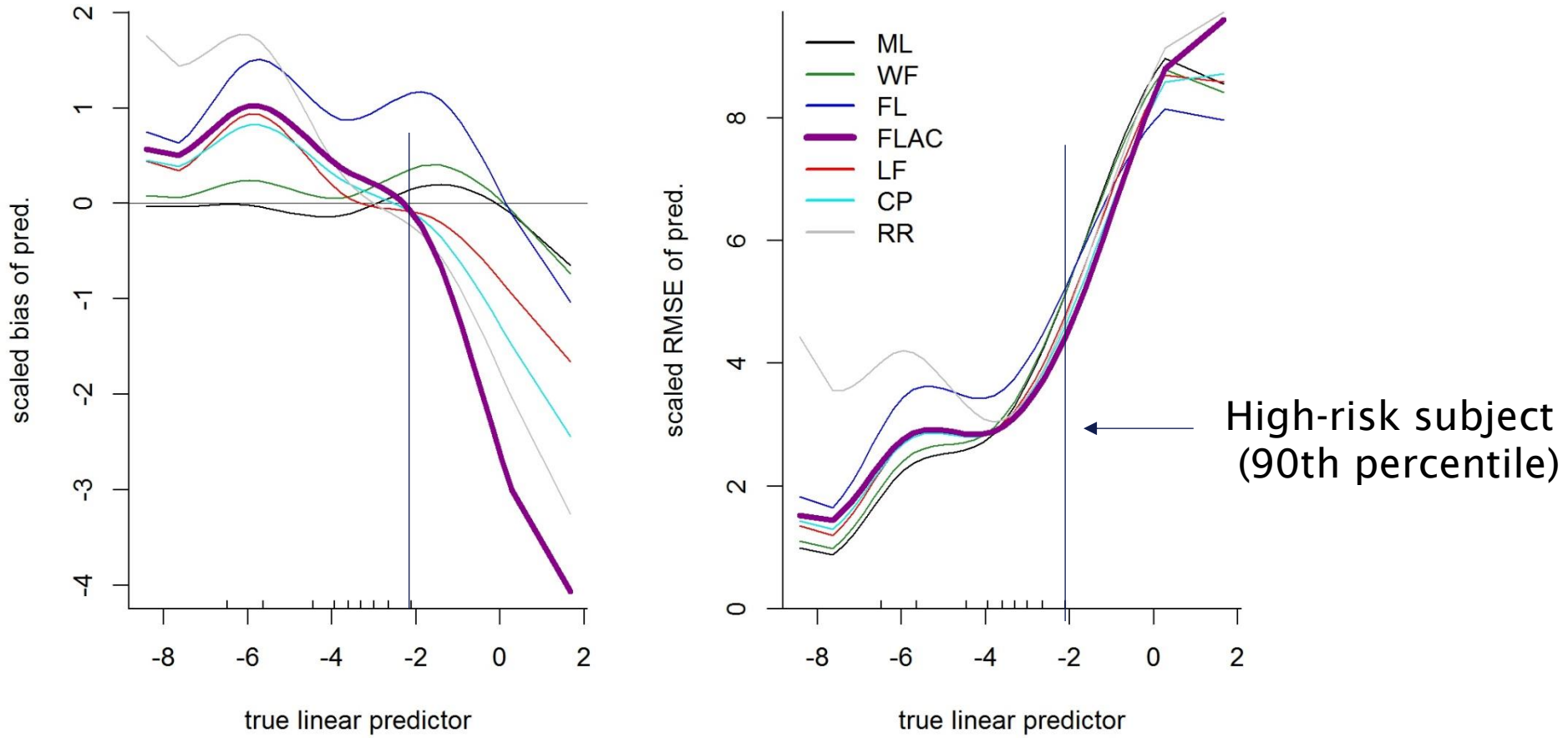
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



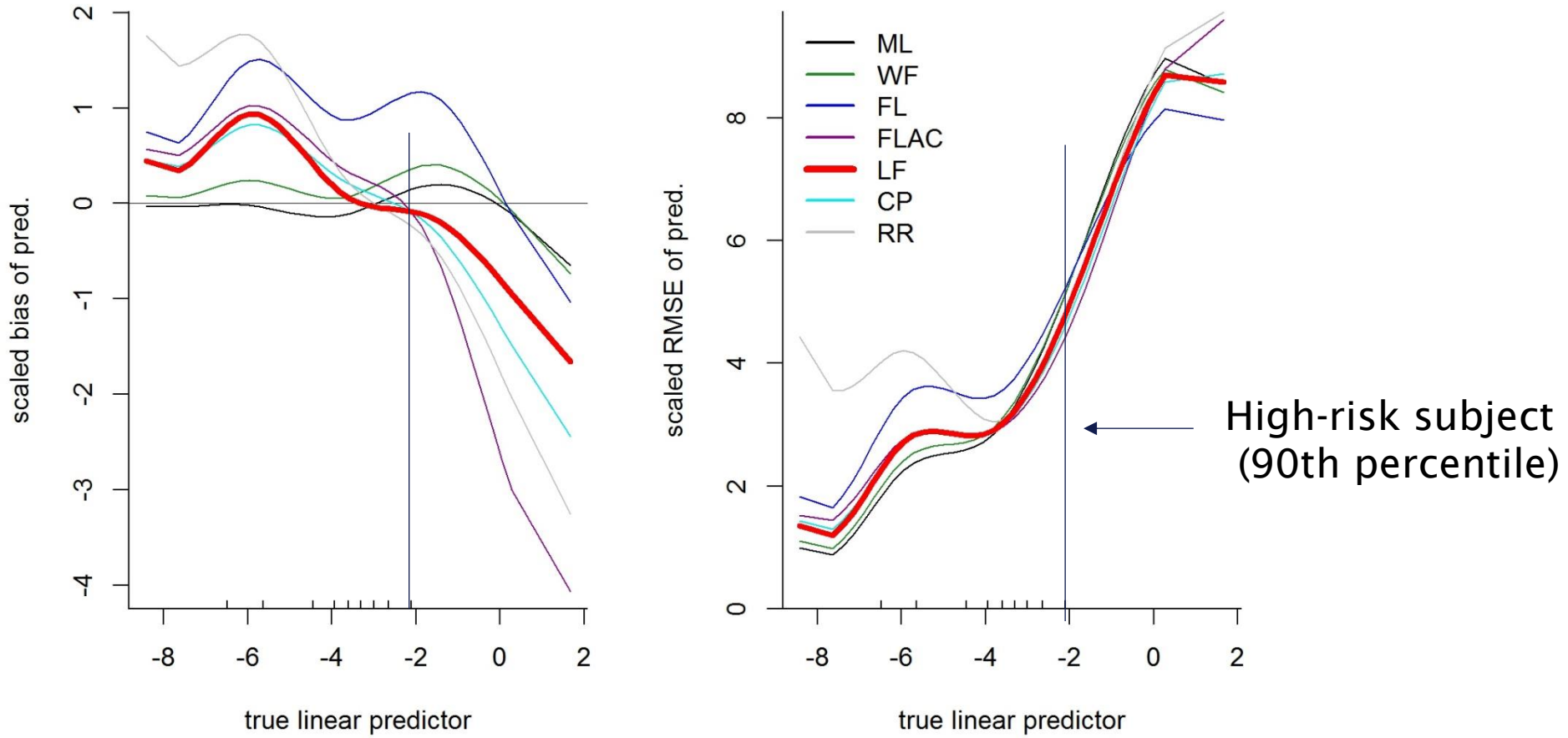
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



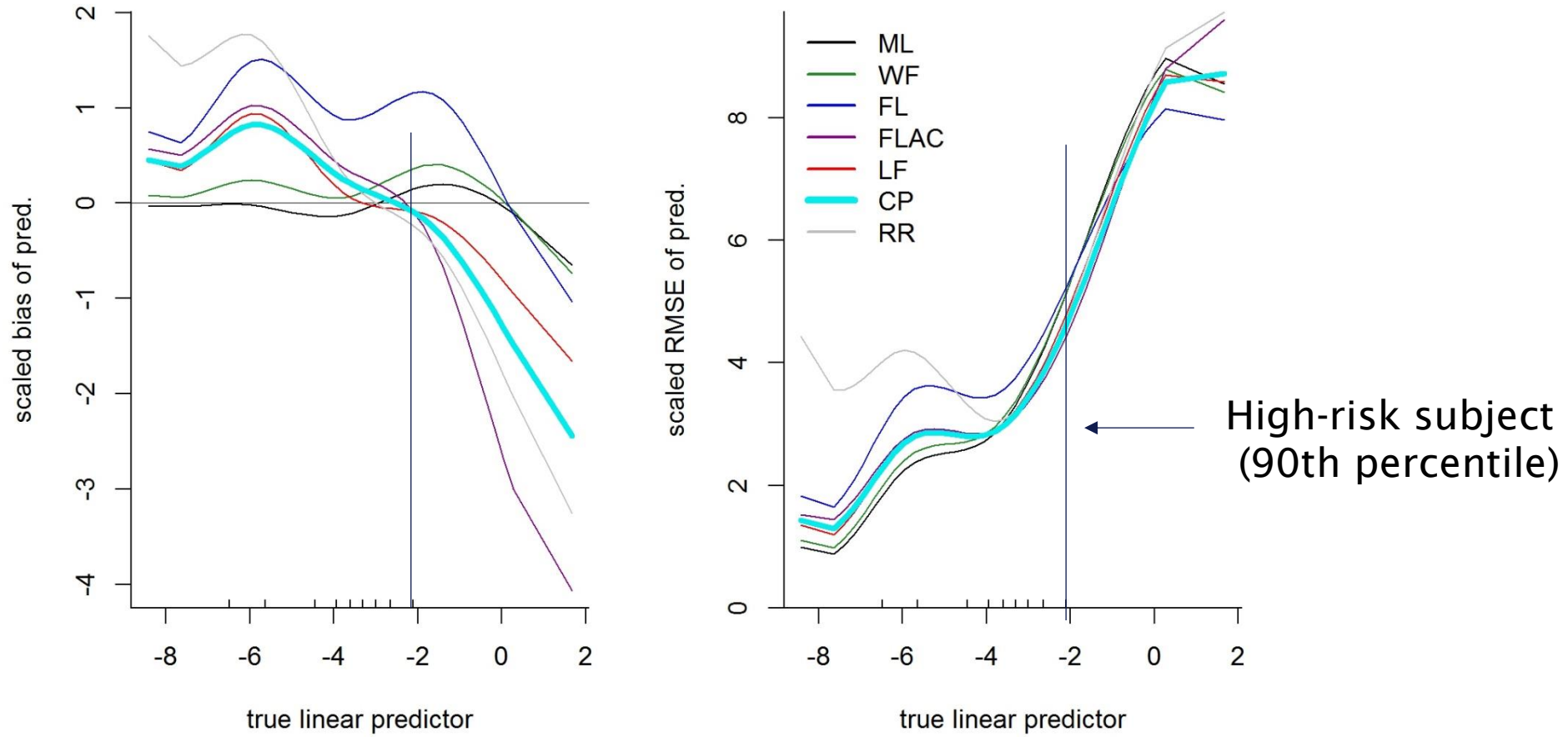
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



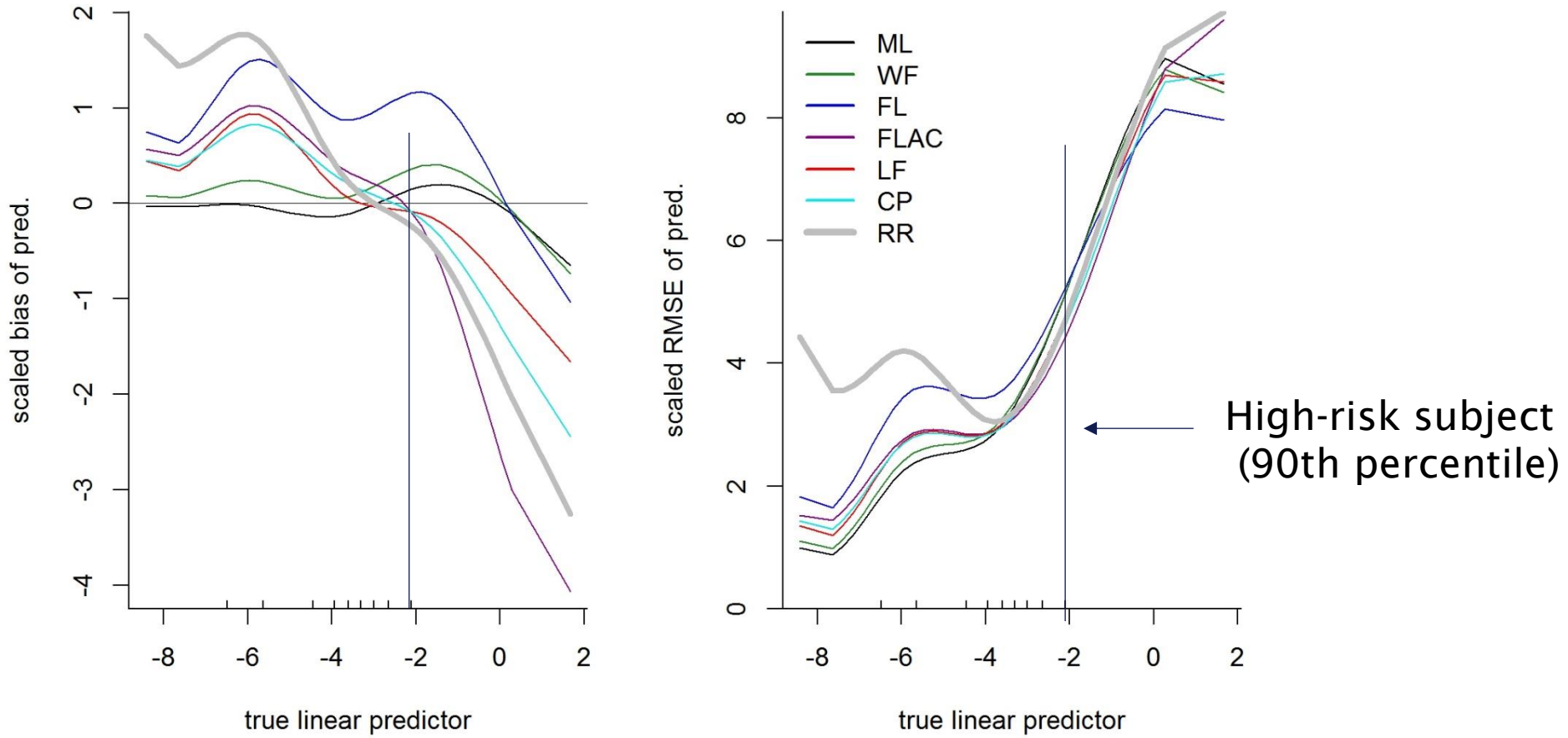
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



N=500, a=1, ybar=0.05, b.sign=-1

Comparison

FLAC

- No tuning parameter
- Transformation-invariant
- Often best MSE, calibration

Ridge

- Standardization is standard
- Tuning parameter
  – no confidence intervals
- Not transformation-invariant
- Performance decreases
  if effects are very different

Bayesian methods (Cauchy, logF)

- Cauchy: in-built standardization (bayesglm),
  no tuning parameter
- logF(*m*,*m*): choose *m* by ’95% prior region’ for parameter of interest
  m=1 for wide prior, *m*=2 less vague
- (in principle, *m* could be tuned as in ridge)
- logF: easily implemented
- Cauchy and logF are not transformation-invariant

Comparison

FLAC

- No tuning parameter
- Transformation-invariant
- Often best MSE, calibration

Recommended!

Ridge

- Standardization is standard
- Tuning parameter
  – no confidence intervals
- Not transformation-invariant
- Performance decreases
  if effects are very different

Bayesian methods (Cauchy, logF)

- Cauchy: in-built standardization (bayesglm),
  no tuning parameter
- logF(*m*,*m*): choose *m* by ’95% prior region’ for parameter of interest
  m=1 for wide prior, *m*=2 less vague
- (in principle, *m* could be tuned as in ridge)
- logF: easily implemented
- Cauchy and logF are not transformation-invariant

References

- Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Statistics in Medicine 2002
- Mansournia M, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression – causes, consequences and control. American Journal of Epidemiology, 2018.
- Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth‘s logistic regression with rare events – accurate effect estimates and predictions? Statistics in Medicine 2017.

Please cf. the reference lists therein for all other citations of this presentation.

Further references:

- Gustafson P, Greenland S. Interval estimation for messy observational data. Statistical Science 2009, 24:328-342.
- Rainey C. Estimating logit models with small samples. Political Science Research and Methods 2018 (cond. accepted).
- van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, Reitsma JB. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Meth 2016.

Confidence intervals

Important:

- With penalized (=shrinkage) methods one cannot achieve nominal coverage over all possible parameter values
- But one can achieve nominal coverage averaging over the implicit prior

- Prior – penalty correspondence can be *a-priori* established if there is no tuning parameter
- Important to use profile penalized likelihood method
- Wald method ($\hat{\beta} \pm 1.96\, SE$) depends on unbiasedness of estimate

Gustafson&Greenland, StatScience 2009

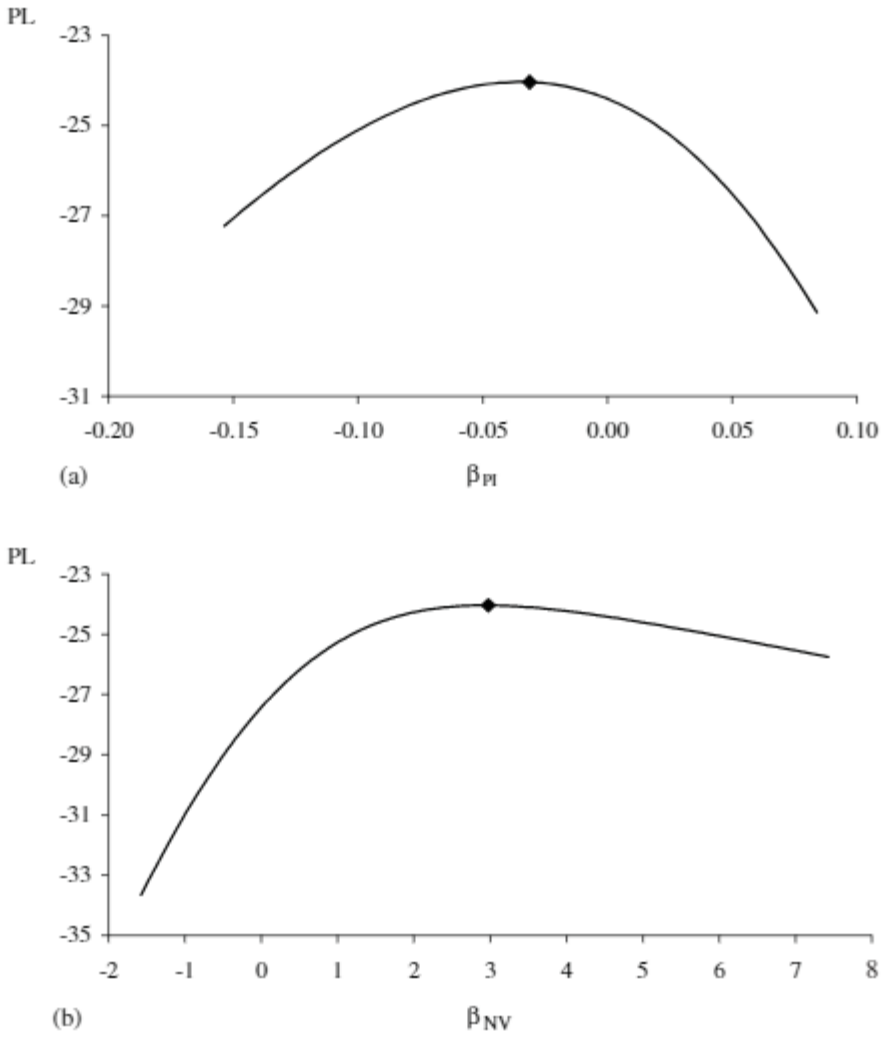


Figure 1. Profile penalized log likelihood function (PL) for factors (a) PI and (b) NV. The functions were obtained by fixing the investigated parameters, β_{PI} and β_{NV}, at 100 predefined values evenly spread within ± 3 standard errors ($\hat{\sigma}(\beta_{PI}) = 0.04$, $\hat{\sigma}(\beta_{NV}) = 1.55$) of the point estimates ($\hat{\beta}_{PI} = -0.03$, $\hat{\beta}_{NV} = 2.93$) denoted by '◇'.

Conclusion

We recommend FLAC for:

- Achieving unbiased predictions
- Good performance
- Invariance to transformations or coding
- Cannot be ‘outsmarted’ by creative coding

Simulating the example of Greenland

- We should distinguish BNC in a single data set from a systematic increase in bias of a method (in simulations)

	X=0	X=1	
Y=0	315	5	*320*
Y=1	31	1	*32*
	346	*6*	*352*

- Simulation of the example:
- Fixed groups x=0 and x=1, P(Y=1|X) as observed in example
- True log OR=0.709

Simulating the example of Greenland

- True value: log OR = 0.709

Parameter	ML	Jeffreys-Firth	
Bias β_1	*	+18%	
RMSE β_1	*	0.86	
Bayesian non-collapsibility β_1		**63.7%**	

* Separation causes β_1 to be undefined ($-\infty$) in 31.7% of the cases

Simulating the example of Greenland

- To overcome Bayesian non-collapsibility,
  Greenland and Mansournia (2015)
  proposed not to impose a prior on the intercept

- They suggest a log-F(1,1) prior for all other regression coefficients

- The method can be used with conventional frequentist software
  because it uses a data-augmentation prior

Greenland and Mansournia, StatMed 2015

logF(1,1) prior (Greenland and Mansournia, 2015)

Penalizing by log-F(1,1) prior gives $L(\beta)^* = L(\beta) \cdot \prod \frac{e^{\frac{\beta_j}{2}}}{1+e^{\beta_j}}$.

This amounts to the following modification of the data set:

	x1	x2	y
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*

	x1	x2	y
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
0	1	0	0
0	1	0	1
0	0	1	0
0	0	1	1

each assigned a weight of 1

each assigned a weight of ½

- No shrinkage for the intercept, no rescaling of the variables

Simulating the example of Greenland

- Re-running the simulation with the log-F(1,1) method yields:

Parameter	ML	Jeffreys-Firth	logF(1,1)
Bias β_1	*	+18%	
RMSE β_1	*	0.86	
Bayesian non-collapsibility β_1		**63.7%**	**0%**

* Separation causes β_1 be undefined ($-\infty$) in 31.7% of the cases

Simulating the example of Greenland

- Re-running the simulation with the log-F(1,1) method yields:

Parameter	ML	Jeffreys-Firth	logF(1,1)
Bias β_1	*	+18%	-52%
RMSE β_1	*	0.86	1.05
Bayesian non-collapsibility β_1		**63.7%**	**0%**

* Separation causes β_1 be undefined ($-\infty$) in 31.7% of the cases

Other, more subtle occurrences of Bayesian non-collapsibility

- Ridge regression: normal prior around 0
- usually implies bias towards zero,
- But:
- With correlated predictors with different effect sizes, for some predictors the bias can be away from zero

Simulation of bivariable log reg models

- $X_1, X_2 \sim \text{Bin}(0.5)$ with correlation $r = 0.8, n = 50$
- $\beta_1 = 1.5$, $\beta_2 = 0.1$, ridge parameter λ optimized by cross-validation

Parameter	ML	Ridge (CV λ)	Log-F(1,1)	Jeffreys-Firth
Bias β_1	+40% (+9%*)	-26%	-2.5%	+1.2%
RMSE β_1	3.04 (1.02*)	1.01	0.73	0.79
Bias β_2	-451% (+16%*)	+48%	+77%	+16%
RMSE β_2	2.95 (0.81*)	0.73	0.68	0.76
Bayesian non-collapsibility β_2		**25%**	**28%**	**23%**

*excluding 2.7% separated samples

Anti-shrinkage from penalization?

Bayesian non-collapsibility/anti-shrinkage

- can be avoided in univariable models,
  but no general rule to avoid it in multivariable models
- Likelihood penalization can often decrease RMSE
  (even *with* occasional anti-shrinkage)
- **Likelihood penalization ≠ guaranteed shrinkage**