

Logistic regression with rare events: problems and solutions

Georg Heinze

Medical University of Vienna

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

Barcelona, 27 September 2017

Georg.heinze@meduniwien.ac.at

[@Georg__Heinze](https://twitter.com/Georg__Heinze)

<http://prema.mf.uni-lj.si>

<http://cemsiiis.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

- Models
 - for prediction of binary outcomes
 - should be interpretable,
i.e., betas should have a meaning
→ explanatory models

Logistic regression

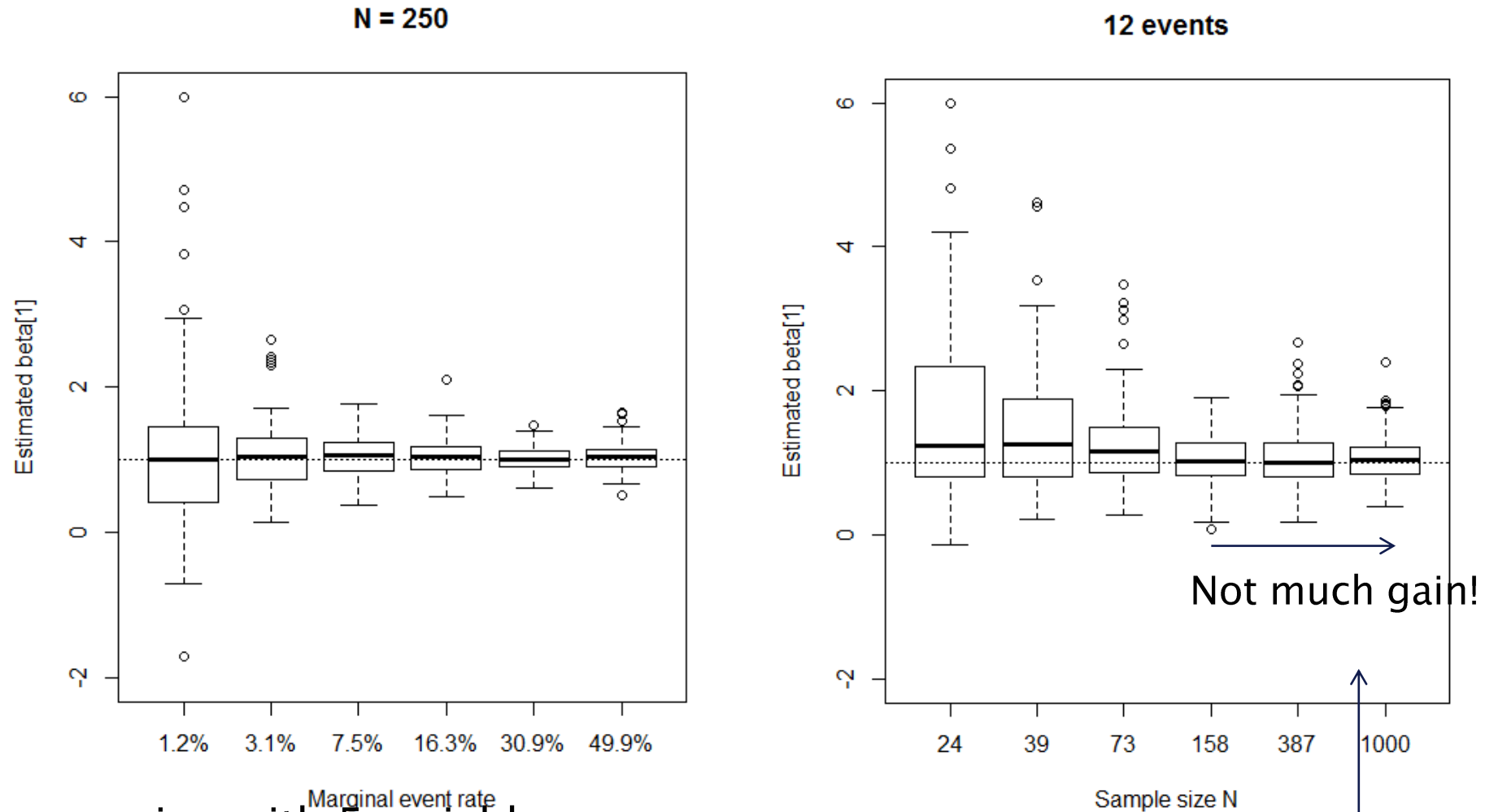
$$\Pr(Y = 1) = \pi = [1 + \exp(-X\beta)]^{-1}$$

- 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: $L(\beta|X) = \prod_{i=1}^n \hat{\pi}_i^{y_i} (1 - \hat{\pi}_i)^{1-y_i}$
- 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'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

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

$$\begin{aligned} W &= \text{diag}(\text{expit}(X_i \beta)(1 - \text{expit}(X_i \beta))) \\ &= \text{diag}(\pi_i(1 - \pi_i)) . \end{aligned}$$

- Firth-type estimates always exist.

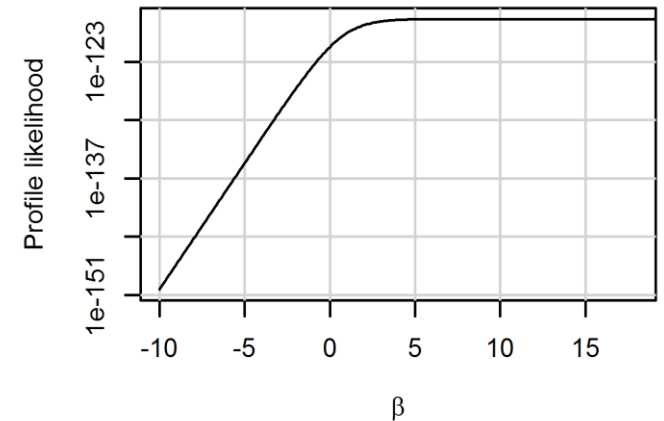
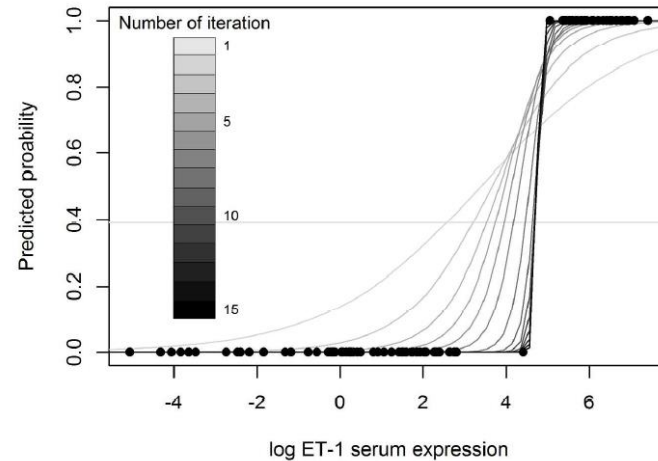
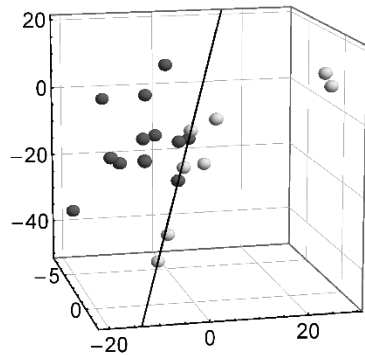
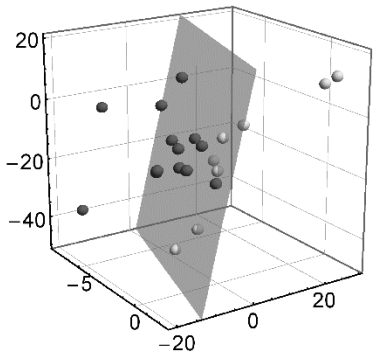
W is maximised at $\pi_i = \frac{1}{2}$, i.e. at $\beta = 0$, thus

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

← ← Shrinkage!

Firth's penalization for logistic regression

- Separation of outcome classes by covariate values (Figs. from Mansournia et al 2017)



- Firth's bias reduction method was proposed as solution to the problem of separation in logistic regression (Heinze and Schemper, 2002)
- Penalized likelihood has a unique mode
- It prevents infinite coefficients to occur

Firth's penalization for logistic regression

Bias reduction also leads to reduction in MSE:

- Rainey, 2017: Simulation study of LogReg for political science
,Firth's methods dominates ML in bias and MSE'

However, 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

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

$$\text{OR}_{B \text{ vs } A} = 11$$

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

$$\text{OR}_{B \text{ vs } A} = 9.89$$

$$\text{av. pred. prob.} = 0.054$$

* Generally: for saturated models

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

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

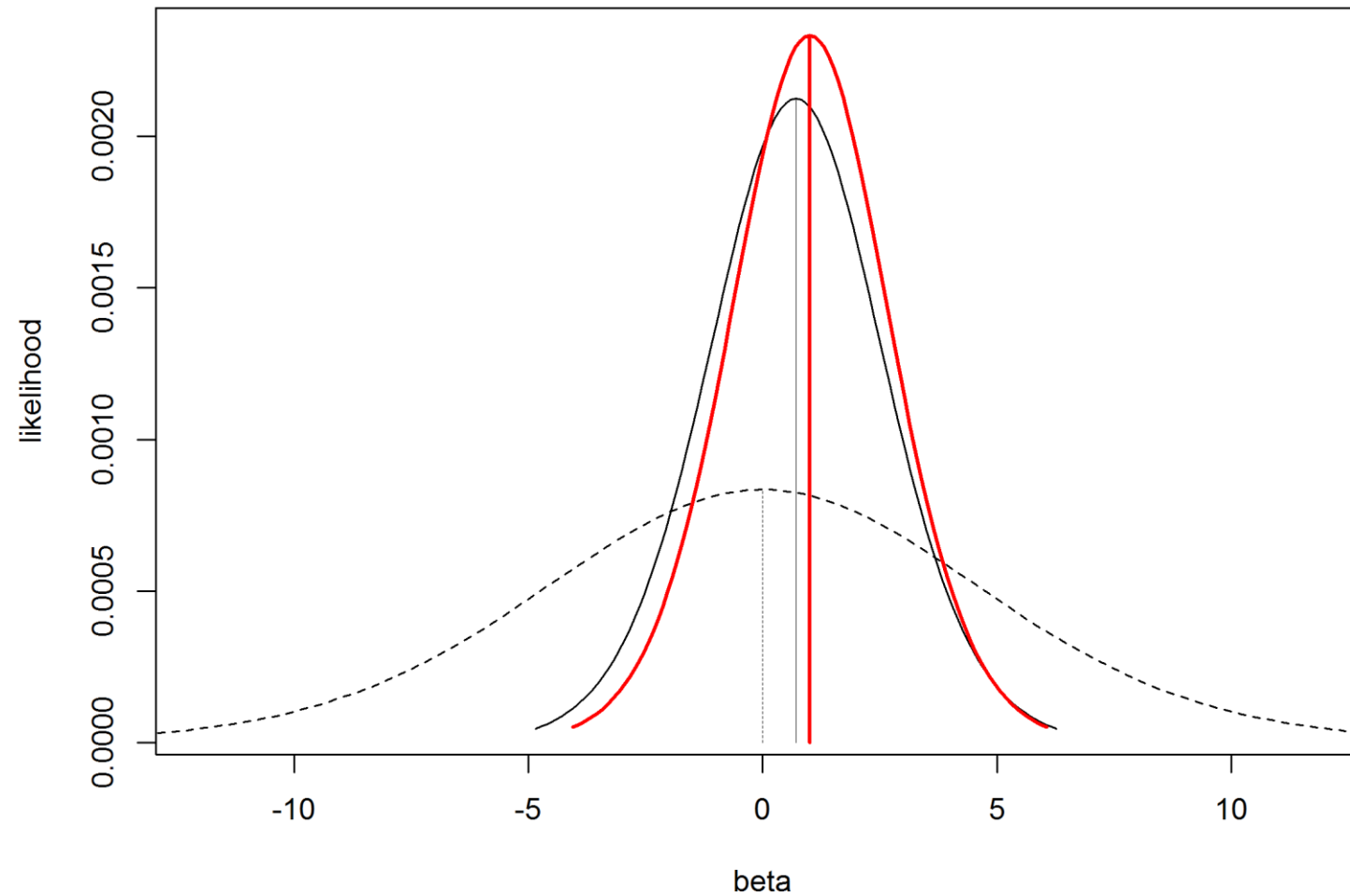
$$\text{OR}_{B\text{vs}A} = 2.03$$

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

$$\text{OR}_{B\text{vs}A} = 2.73$$

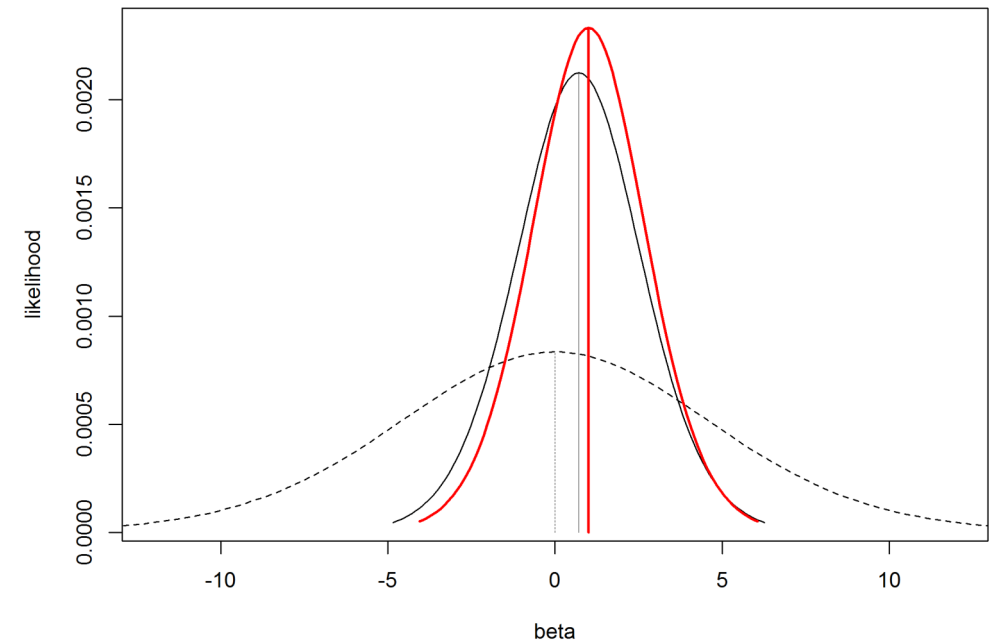
Greenland, AmStat 2010

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

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: $\log \text{OR} = 0.709$

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

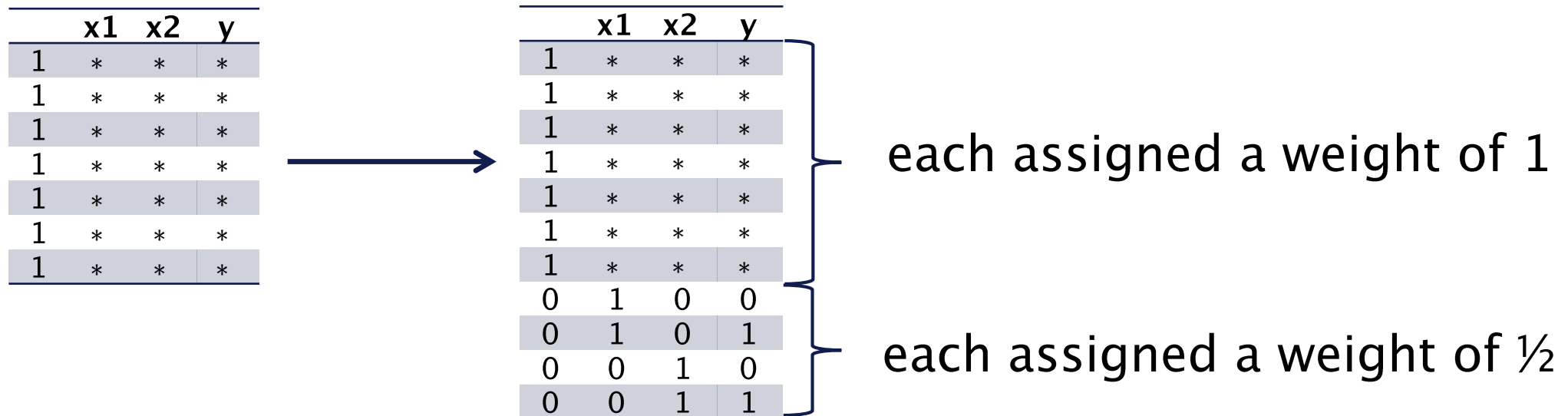
- To overcome Bayesian non-collapsibility, Greenland and Mansournia (2015) proposed not to impose a prior on the intercept
- They suggest a log-F(1,1) prior for all other regression coefficients
- The method can be used with conventional frequentist software because it uses a data-augmentation prior

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:



- 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 \neq guaranteed shrinkage**

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

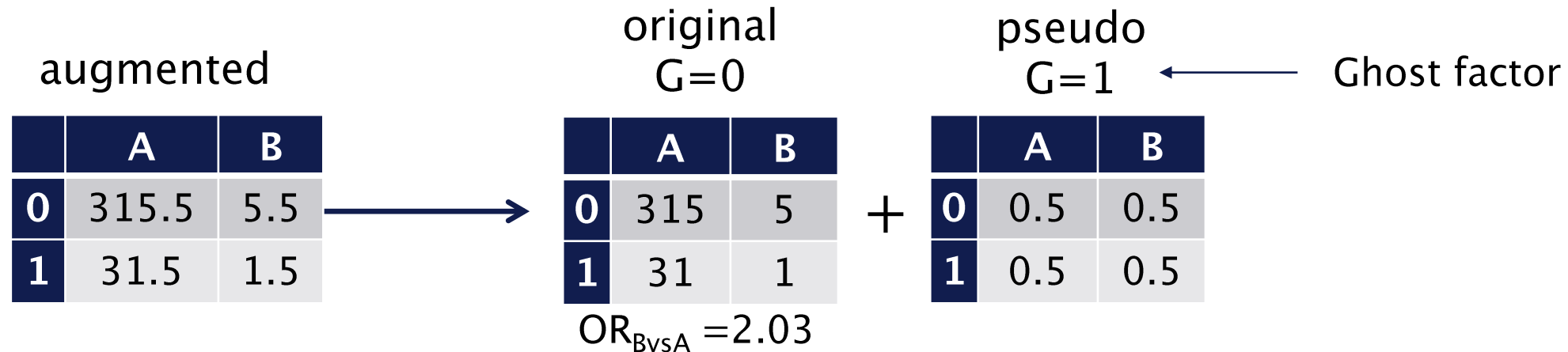
	A	B	
Y=0	315.5	5.5	321
Y=1	31.5	1.5	33
	347	7	352

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 the original and pseudo data:



Define Firth type Logistic regression with Additional Covariate 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:

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

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

They are equivalent to:

$$\begin{aligned} & \sum_{i=1}^N (y_i - \pi_i)x_{ir} + \sum_i^N h_i \left(\frac{1}{2} - \pi_i \right) x_{ir} = \\ & = \sum_{i=1}^N (y_i - \pi_i)x_{ir} + \sum_{i=1}^N \frac{h_i}{2} (y_i - \pi_i) + \sum_{i=1}^N \frac{h_i}{2} (1 - y_i - \pi_i) = 0 \end{aligned}$$

FLAC: Firth's Logistic regression with Added Covariate

- A closer inspection yields:

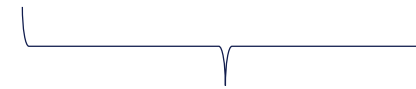
$$\sum_{i=1}^N (y_i - \pi_i) x_{ir} + \sum_{i=1}^N \frac{h_i}{2} (y_i - \pi_i) x_{ir} + \sum_{i=1}^N \frac{h_i}{2} (1 - y_i - \pi_i) x_{ir} = 0$$



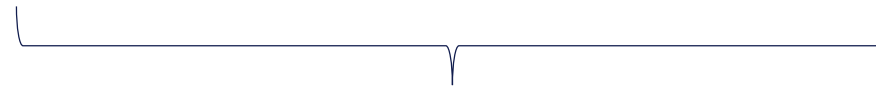
The original data



Original data,
weighted by $h_i/2$



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

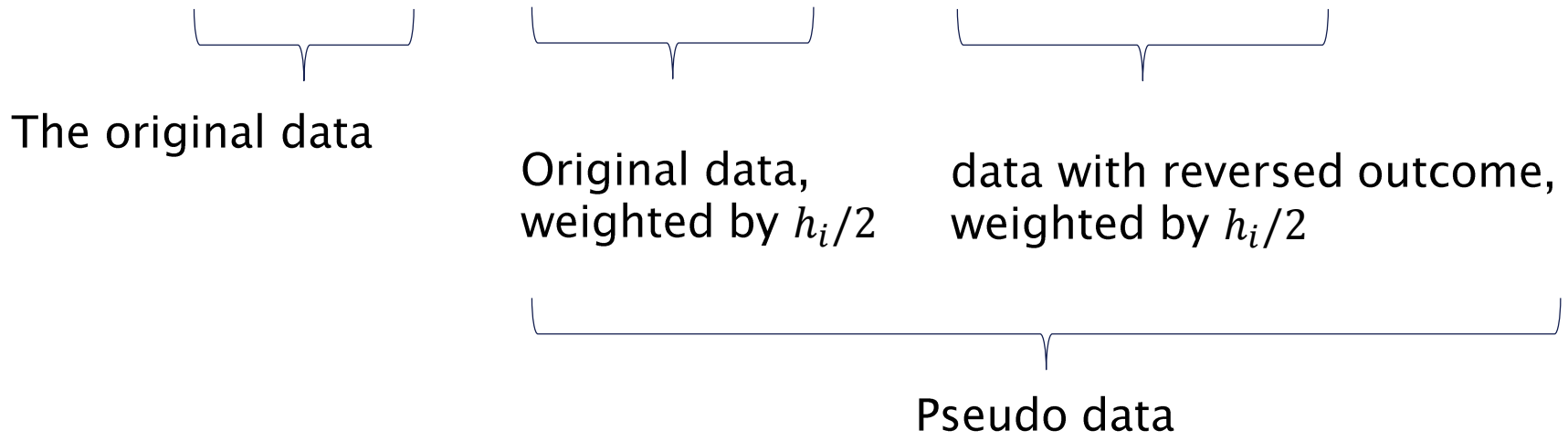


Pseudo data

FLAC: Firth's Logistic regression with Added Covariate

- A closer inspection yields:

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



Ghost factor:
(,Added covariate')

G=0

G=1

FLAC: Firth's Logistic regression with Added Covariate

FLAC estimates can be obtained by the following steps:

- 1) Define an indicator variable discriminating between original and pseudo data.
- 2) Apply ML on the augmented data including the indicator.

 unbiased pred. probabilities

FLIC

Firth type Logistic regression with Intercept Correction:

1. Fit a Firth logistic regression model
2. Modify the intercept in Firth-type estimates such that the average pred. prob. becomes equal to the observed proportion of events.



unbiased pred. probabilities

effect estimates are the same as in Firth type logistic regression

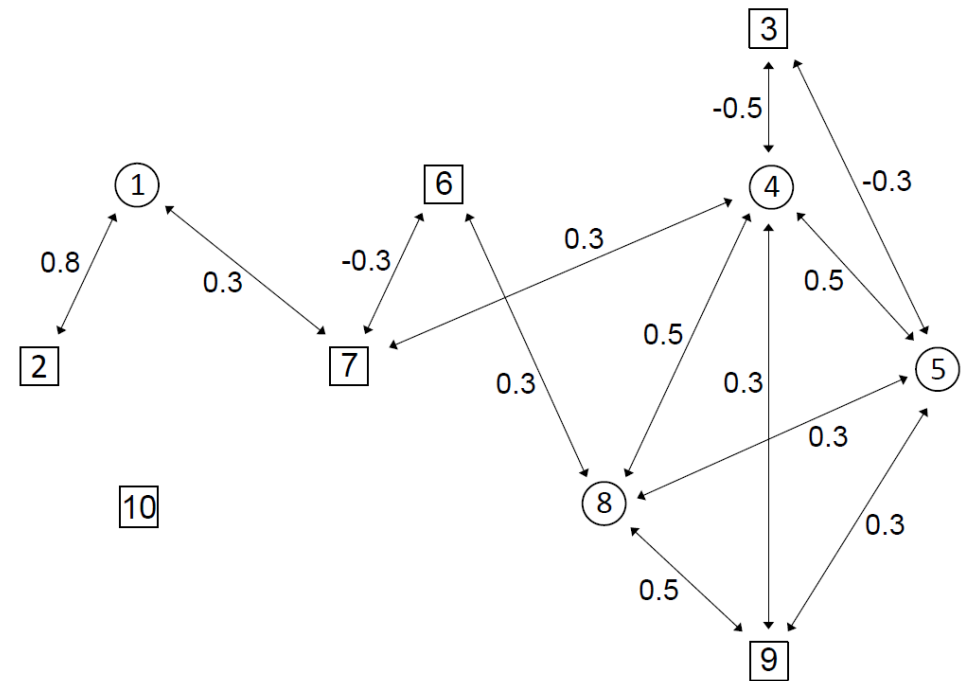
Simulation study: the set-up

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., 2011
- none, moderate and strong effects
of positive and mixed signs

Main evaluation criteria:

bias and RMSE of predictions and effect estimates



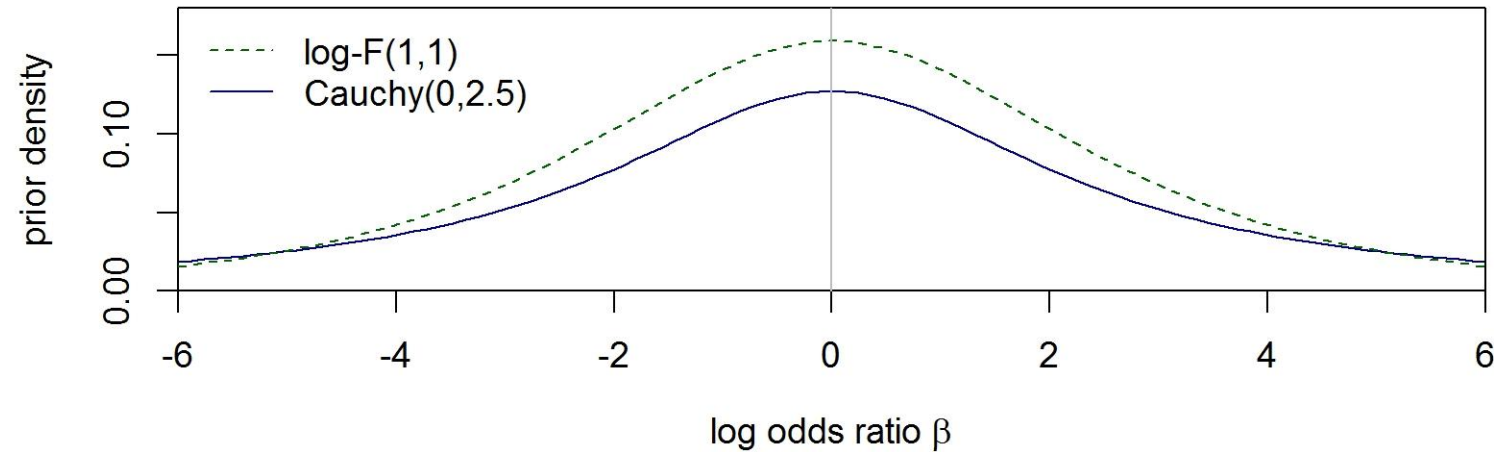
Other methods for accurate prediction

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, LF
- penalization by Cauchy priors with scale parameter=2.5. CP

Cauchy priors (CP)

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



We follow Gelman 2008:

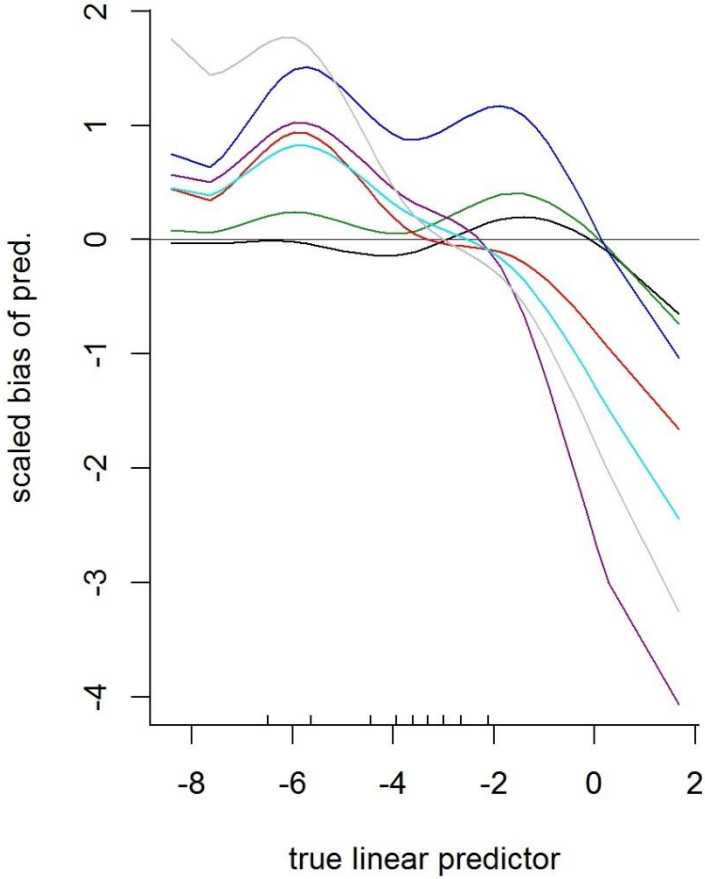
- all variables are centered,
- binary variables are coded to have a range of 1,
- all other variables are scaled to have standard deviation 0.5,
- the intercept is penalized by Cauchy(0,10).

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

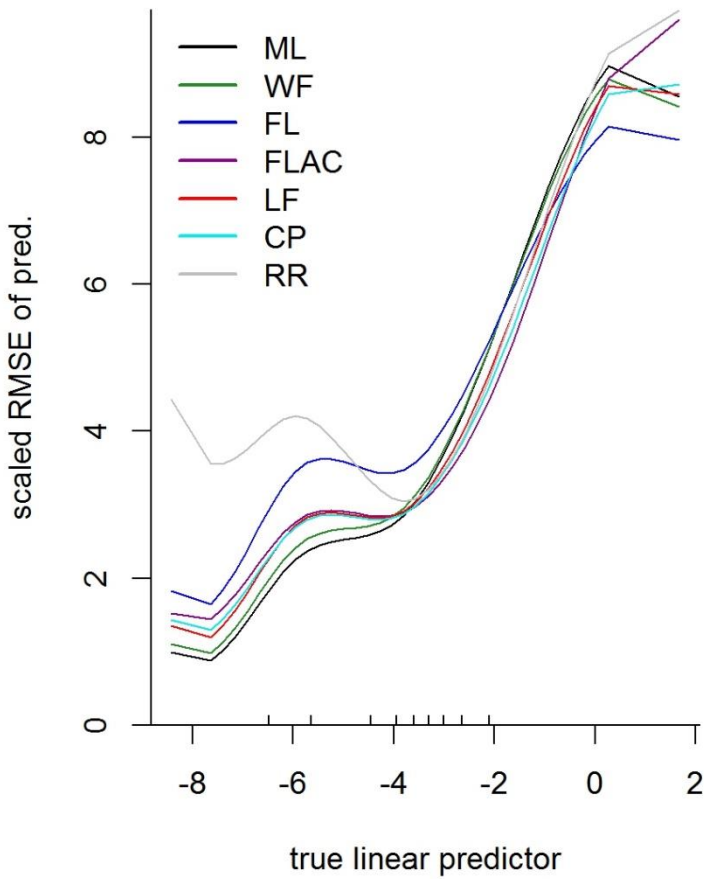
Simulation results

- Bias of $\hat{\beta}$: clear winner is Firth method
FLAC, logF, CP: slight bias towards 0
- RMSE of $\hat{\beta}$:
equal effect sizes: ridge the winner
unequal effect sizes: very good performance of FLAC and CP
closely followed by logF(1,1)
- Calibration: often FLAC the winner; considerable instability of ridge
- Bias and RMSE of $\hat{\pi}$: see following slides

Predictions: bias

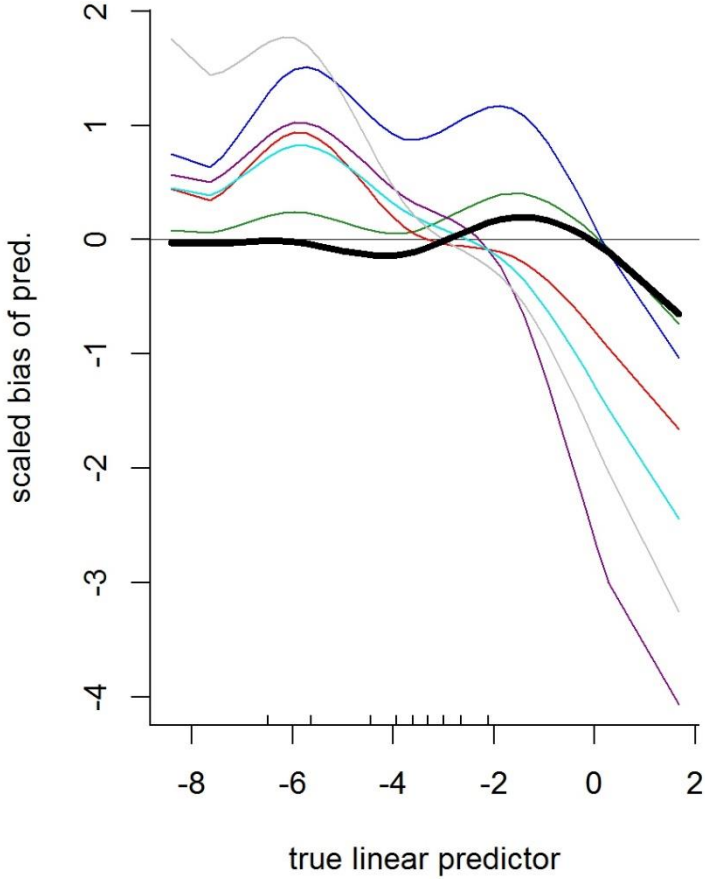


RMSE

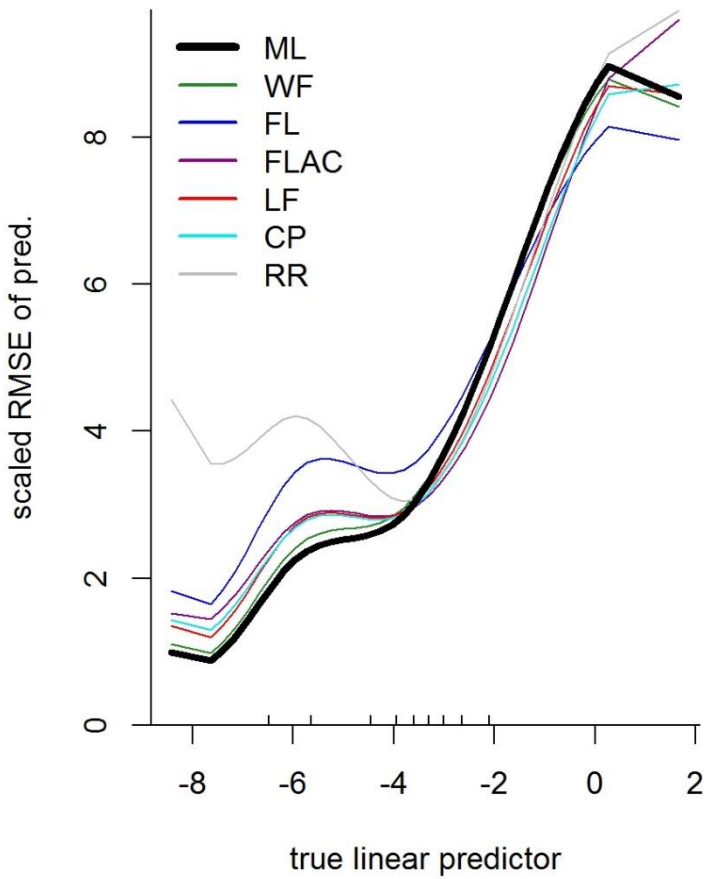


N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias

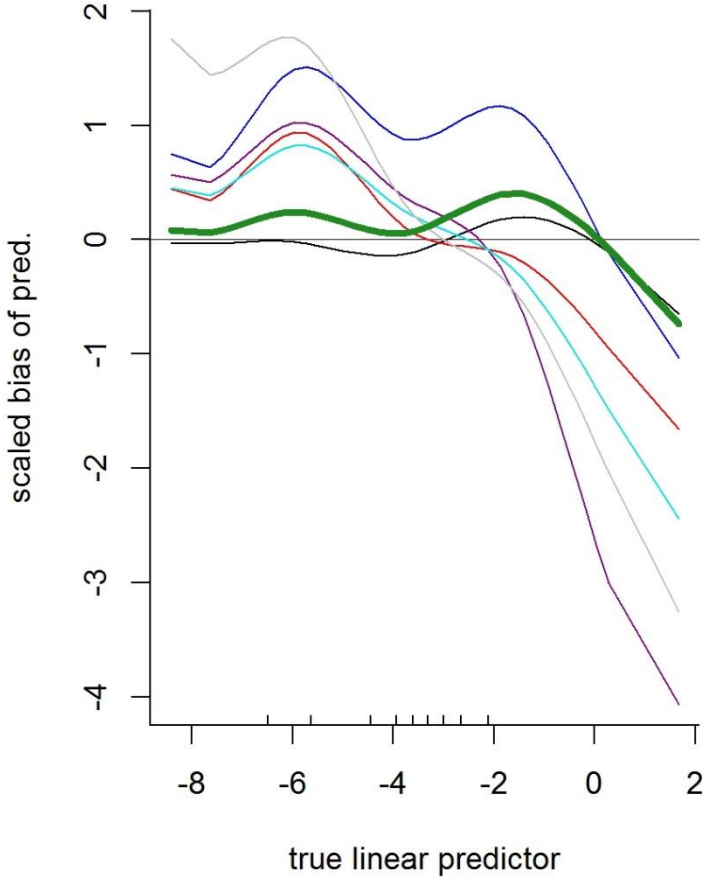


RMSE

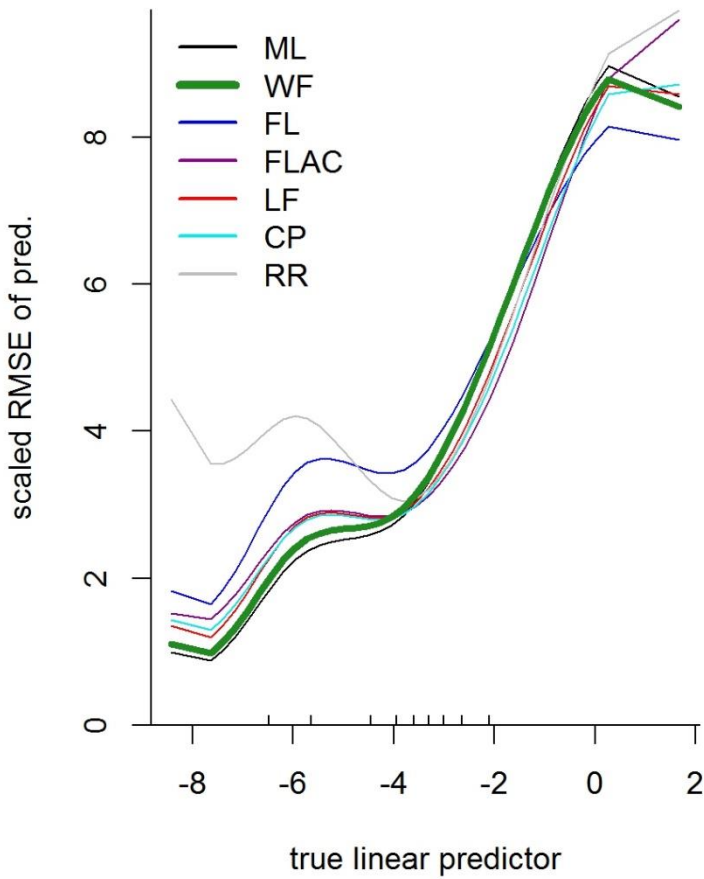


N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias

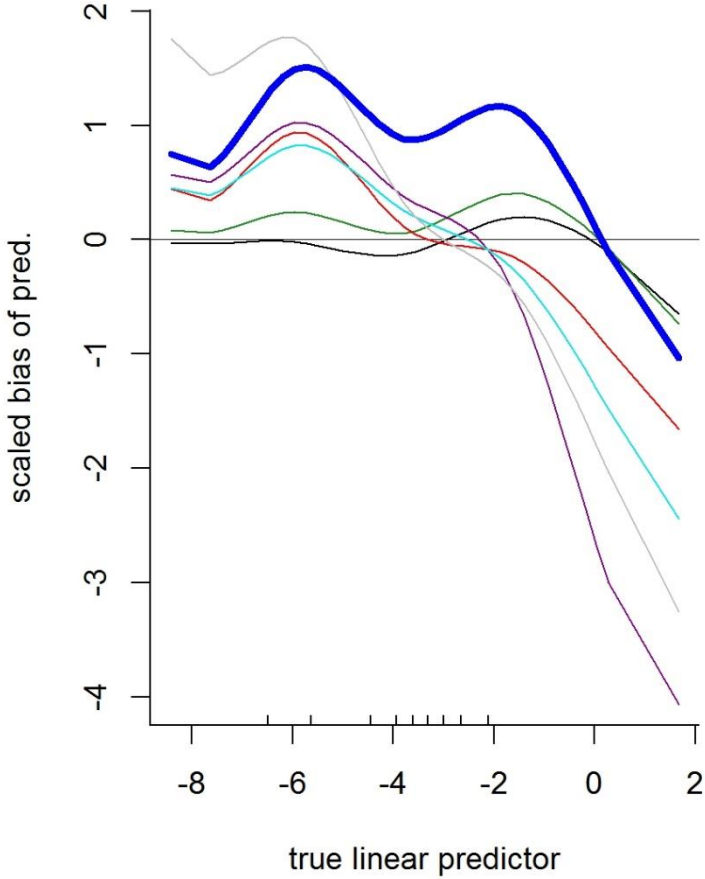


RMSE

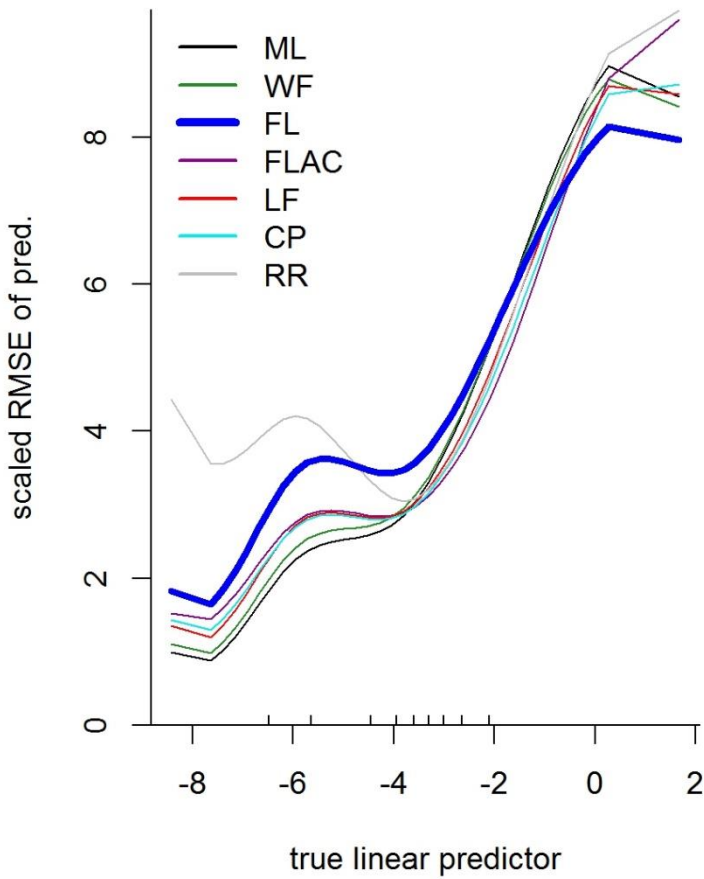


N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias

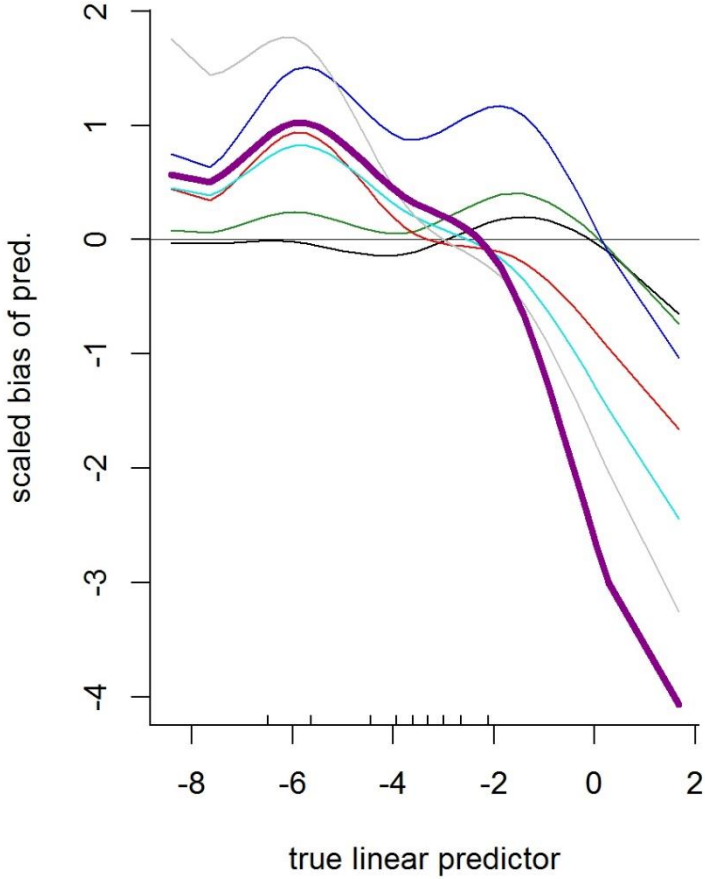


RMSE

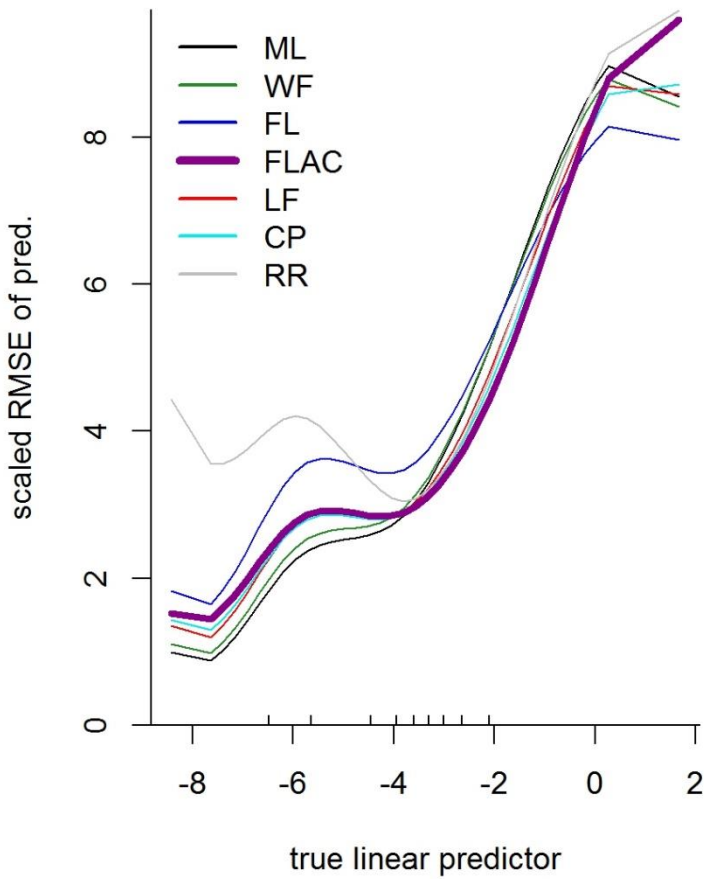


N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias

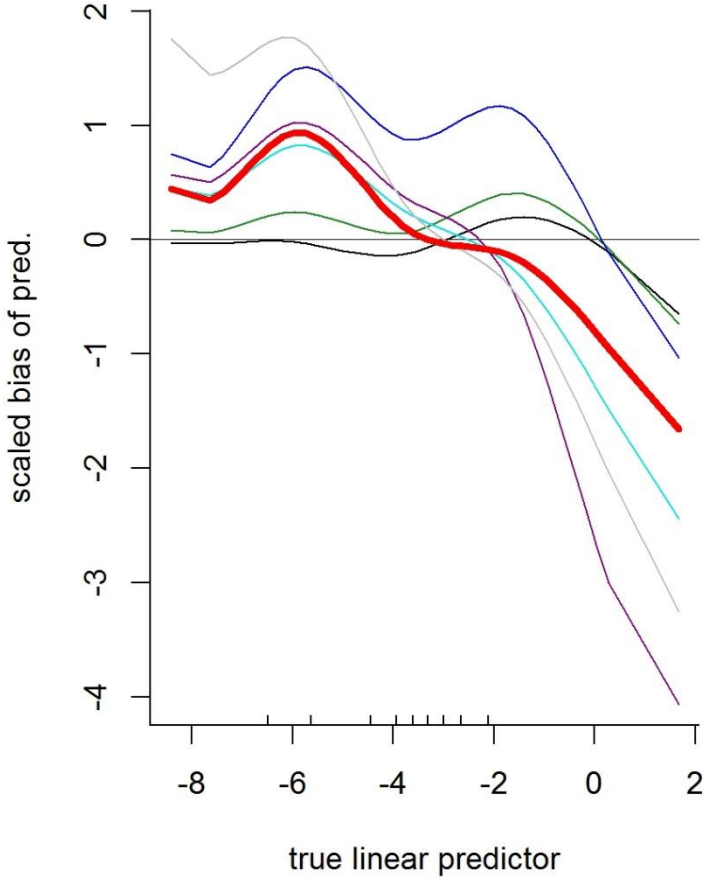


RMSE

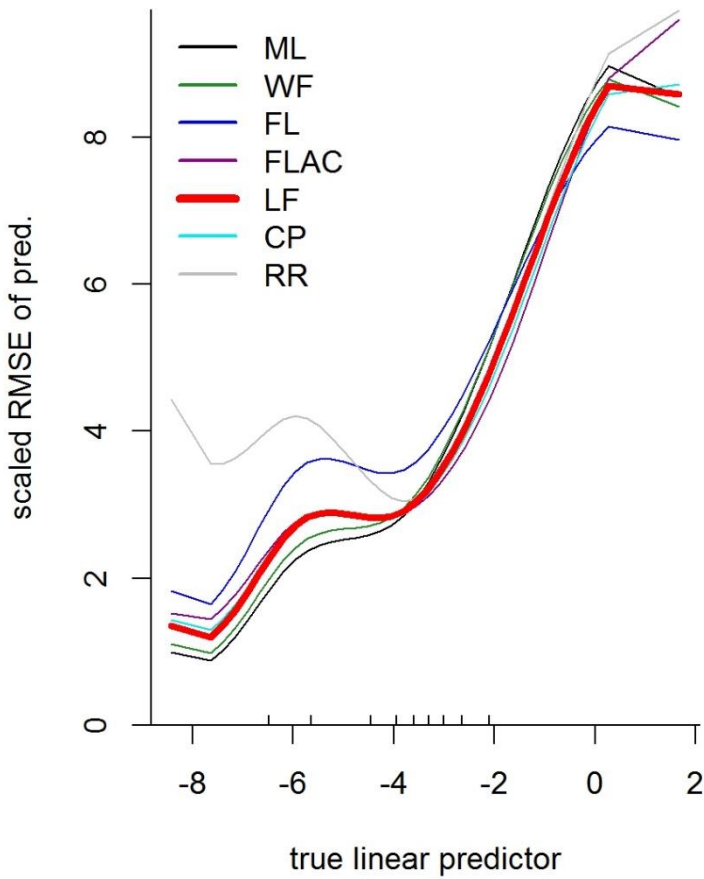


N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias

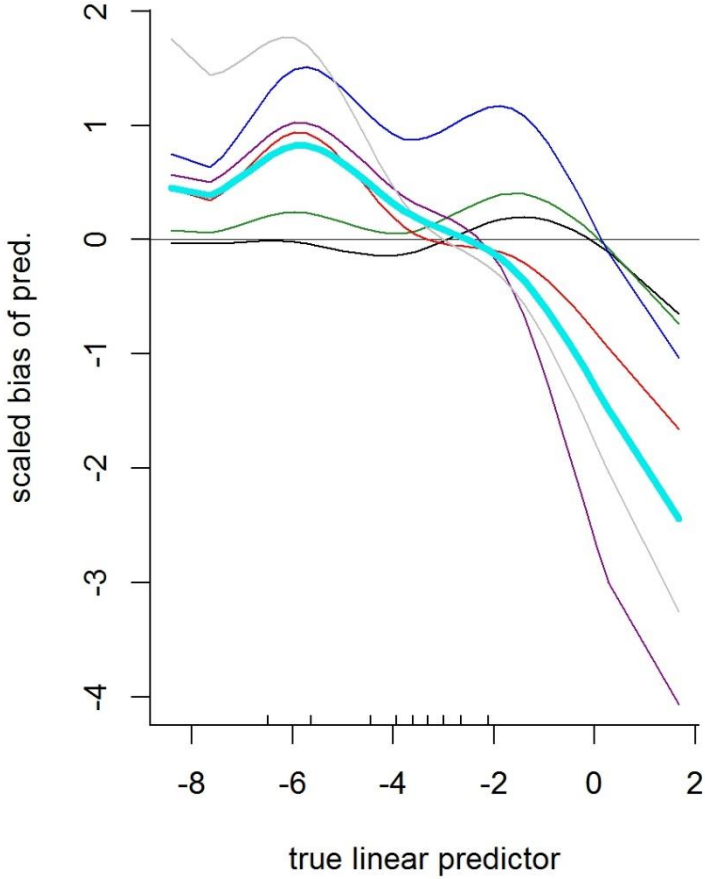


RMSE

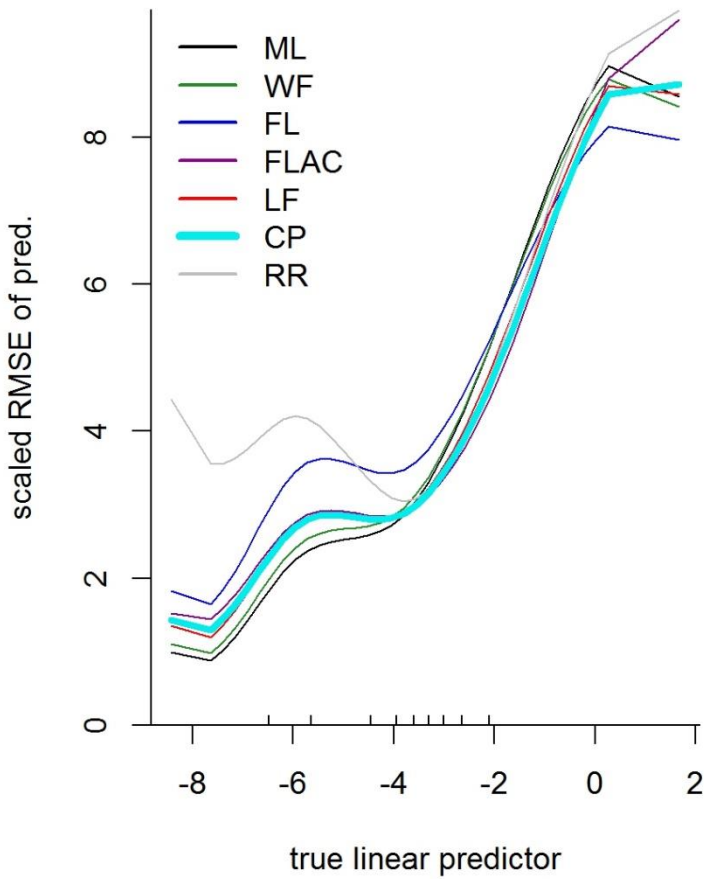


N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias

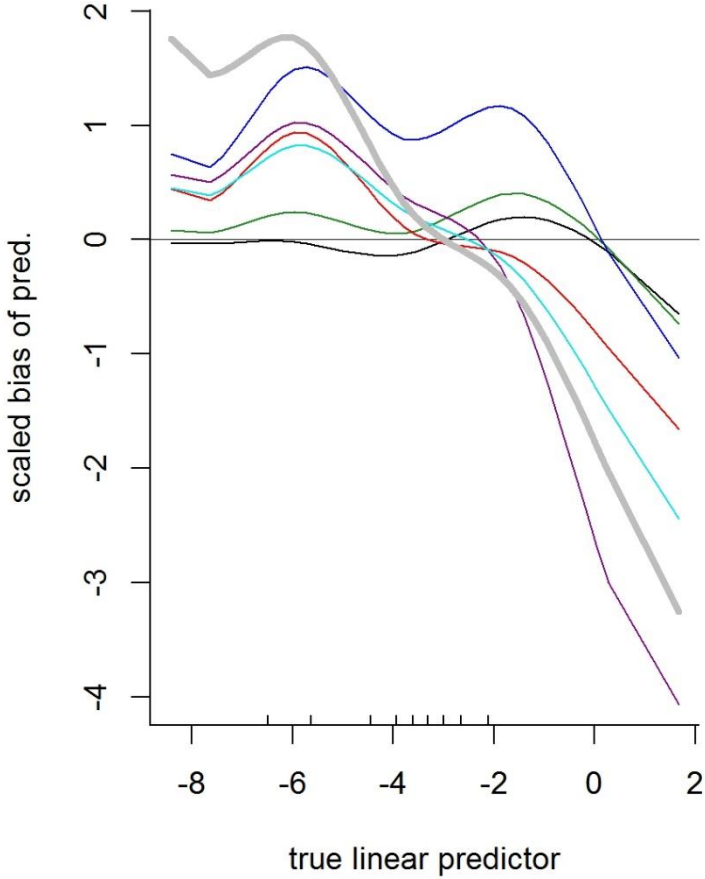


RMSE

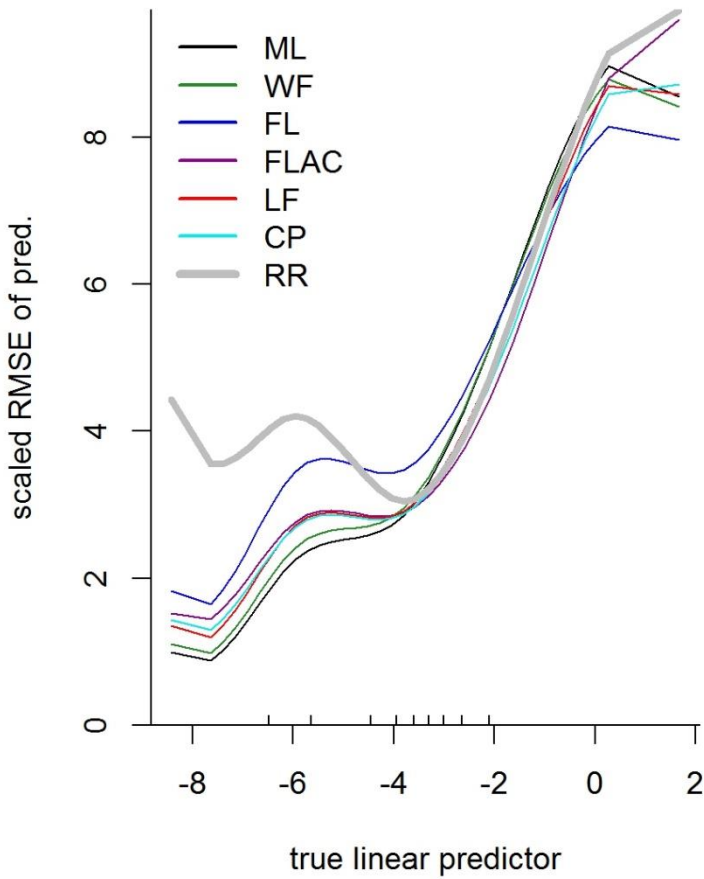


N=500, a=1, ybar=0.05, b.sign=-1

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 (CP, logF)

- CP: in-built standardization, no tuning parameter
- $\log F(m, m)$: 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
- CP and logF are not transformation-invariant

Confidence intervals

It is important to note that:

- 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

Conclusion

- We recommend FLAC for:
- Good performance
- Invariance to transformations or coding
- Cannot be 'outsmarted' by creative coding

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*, 2017.
- 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. www.carlislerainey.com/papers/small.pdf (27 March 2017)