MEDICAL UNIVERSITY OF VIENNA

Accurate Prediction of Rare Events with Firth's Penalized Likelihood Approach

Angelika Geroldinger, Daniela Dunkler, Rainer Puhr,
Rok Blagus, Lara Lusa, Georg Heinze

12th Applied Statistics
September 2015

Overview

1. Introduction: Bias, bias reduction
2. Firth's method: Definition and properties
3. PREMA: Accurate prediction and Firth's method

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:

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

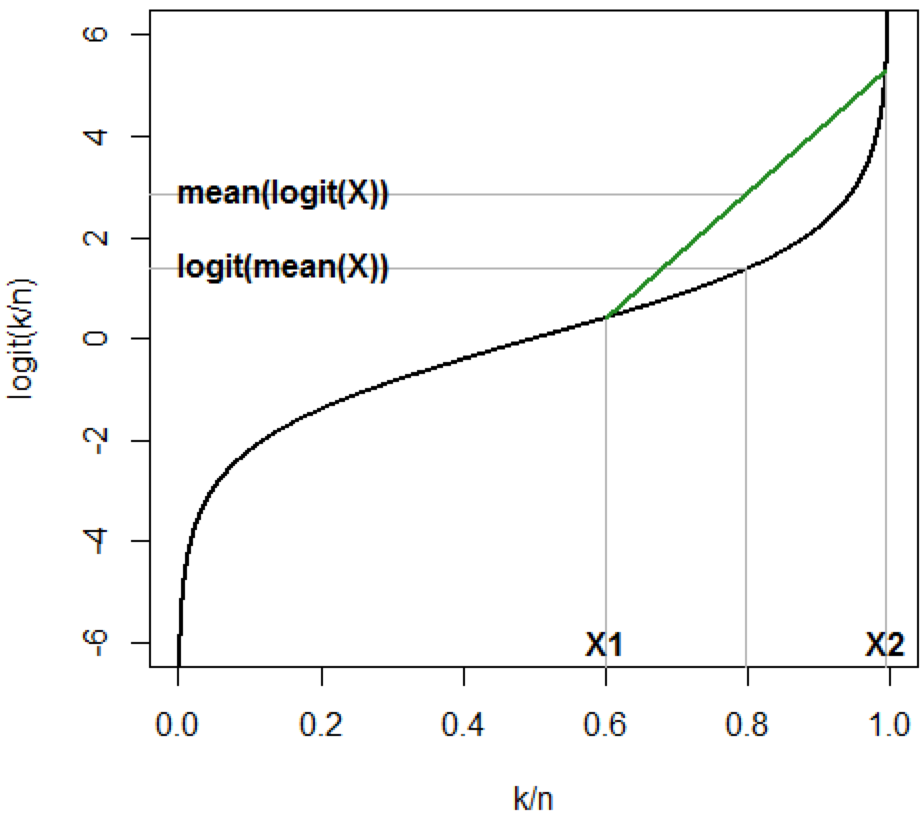
Since k/n is unbiased,

$\hat{\alpha}$ is biased!

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

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

Jensen's inequality



Example: Bias in logistic regression

Consider a model containing only intercept, no regressors:

$$\operatorname{logit}\left(P(Y=1)\right)=\alpha.$$

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

$$\hat{\alpha}=\operatorname{logit}\left(\mathrm{k/n}\right).$$

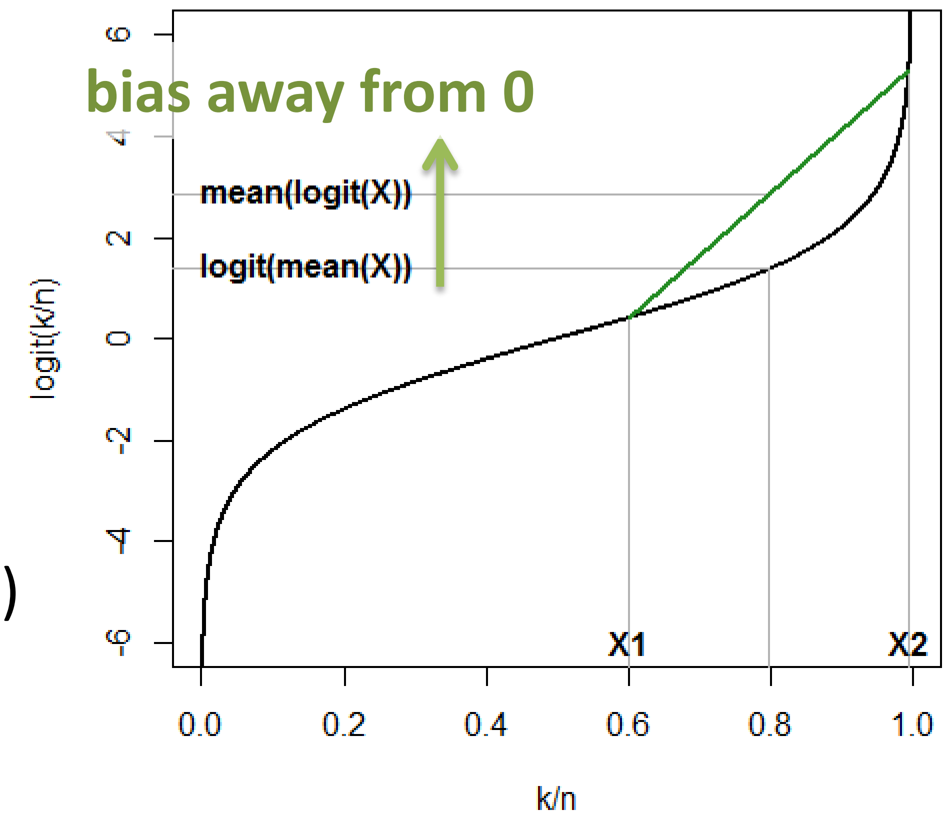
Since $\mathrm{k/n}$ is unbiased,

$\hat{\alpha}$ is biased!

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

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

Jensen's inequality



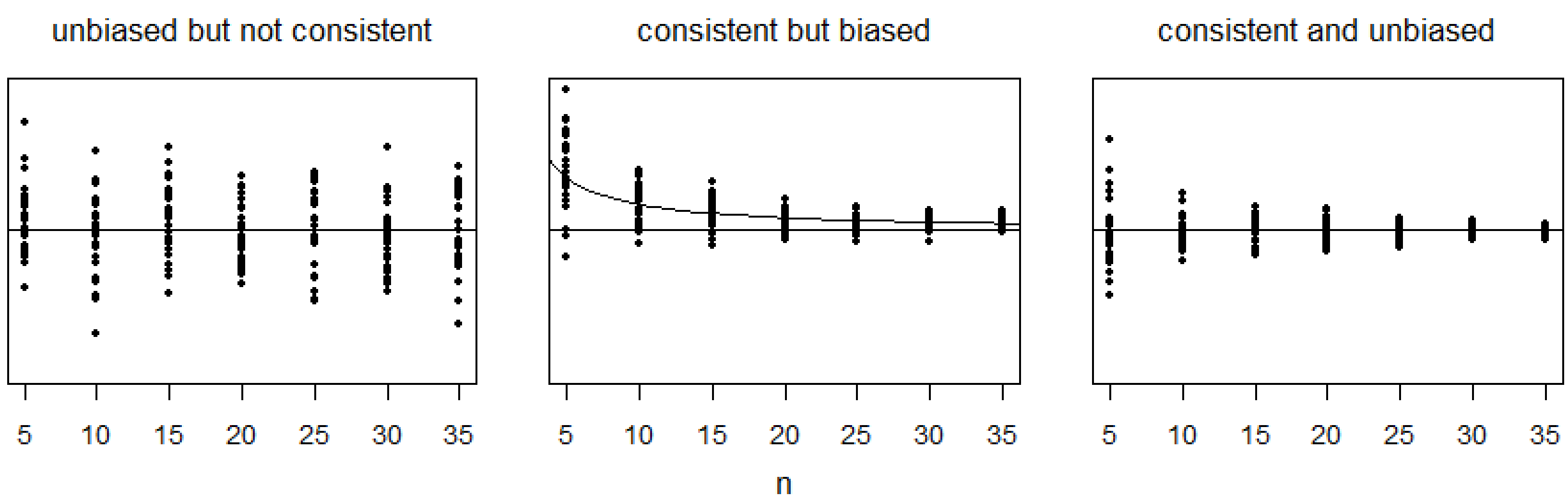
Bias and consistency

The **bias of an estimator** $\hat{\theta}$ for a true value θ is defined as

$$\mathrm{bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta.$$

Example: Let $Y_1, \ldots Y_n$ be i.i. normally distributed. Then, Y_n is an unbiased estimator for the mean.

An estimator $\hat{\theta}$ is called **consistent** if it converges in probability to θ.



Bias reduction

For ML-estimates in regular models one can show that

$$\text{bias}(\hat{\theta}) = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \ \dots .$$

Some approaches to a bias-reduced estimate $\hat{\theta}_{bc}$:

bias-corrective

- jackknife,
- bootstrap,
- explicitly determine the function b_1 and set $\hat{\theta}_{bc} = \theta - b_1(\hat{\theta})$,

bias-preventive

- Firth type penalization.

Firth type penalization

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

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

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 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 type penalization

We are interested in logistic regression:

Here the penalized likelihood is given by $L(\theta)\,\det(X^t W X)^{1/2}$ with

$$W = \mathrm{diag}(\mathrm{expit}(\mathrm{X_i}\theta)(1 - \mathrm{expit}(\mathrm{X_i}\theta))).$$

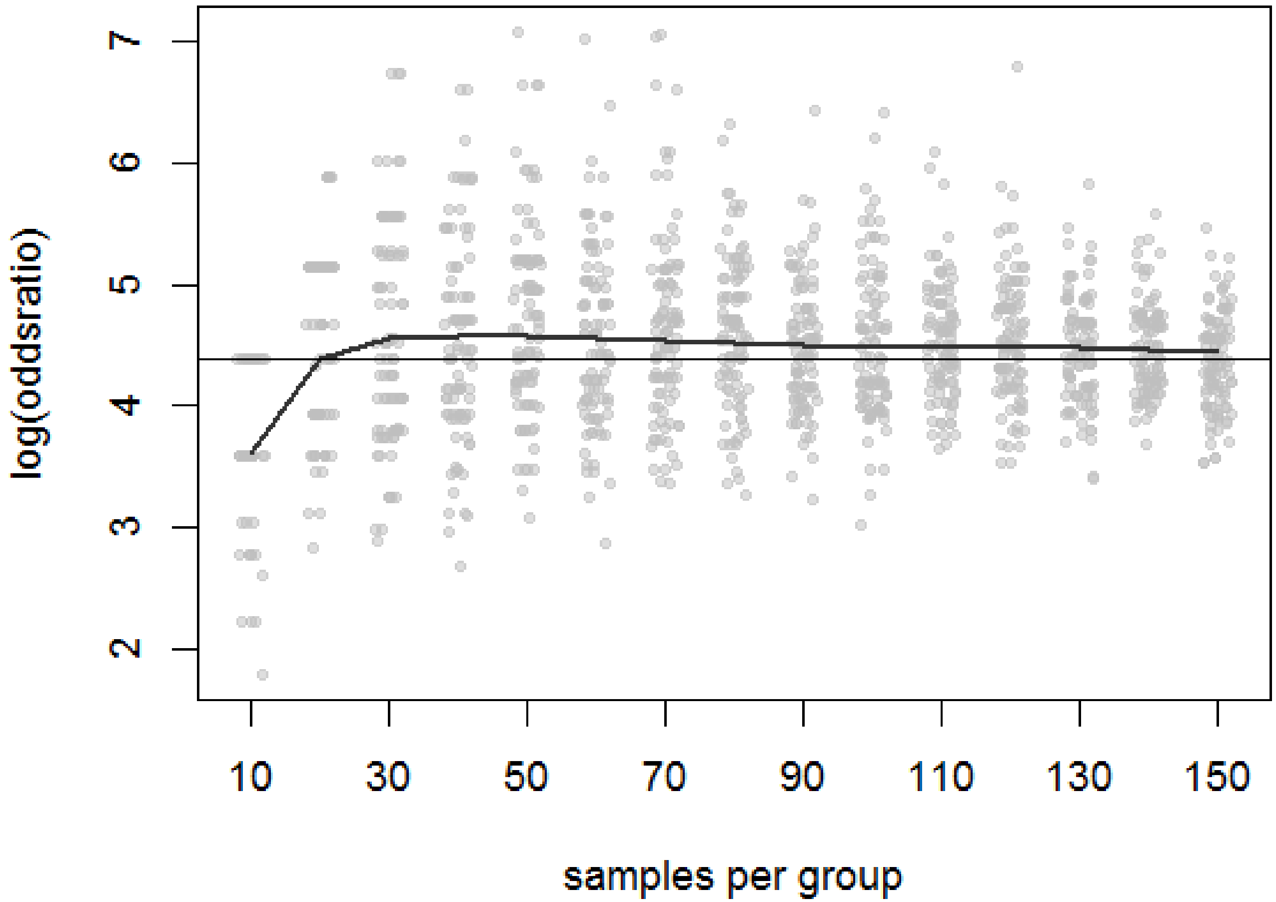
- W is maximised at $\theta = 0$, i.e. the ML estimates are shrunken towards zero,
- for a 2×2 table (logistic regression with one binary regressor), the Firth's bias correction amounts to adding $1/2$ to each cell.

Example: 2 × 2 table

Two groups with event probabilities 0.9 and 0.1.
ML-Estimates:

X

Y		A	B
	0	0.1	0.9
	1	0.9	0.1

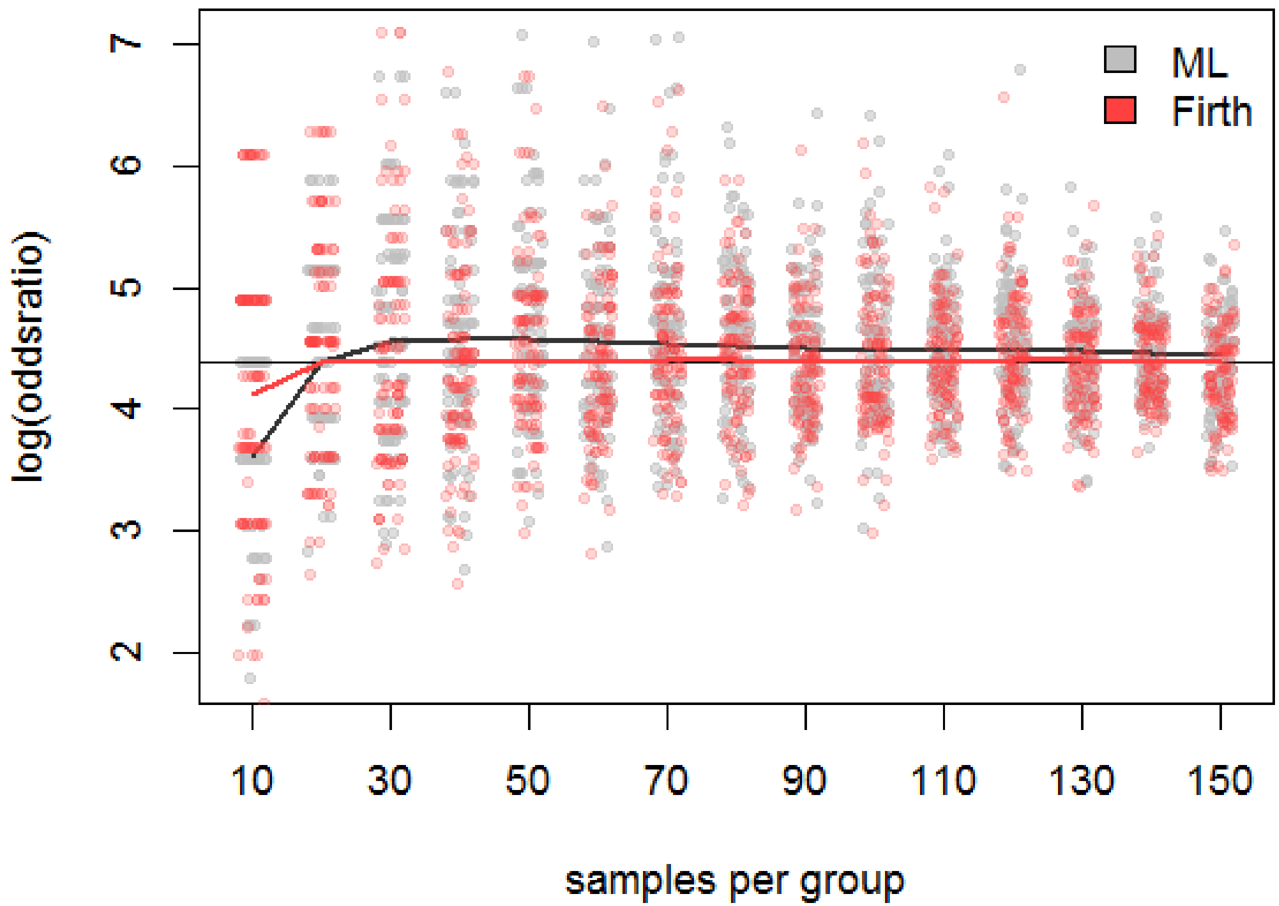


Example: 2 × 2 table

Two groups with event probabilities 0.9 and 0.1.
ML-Estimates and Firth-Estimates:

X

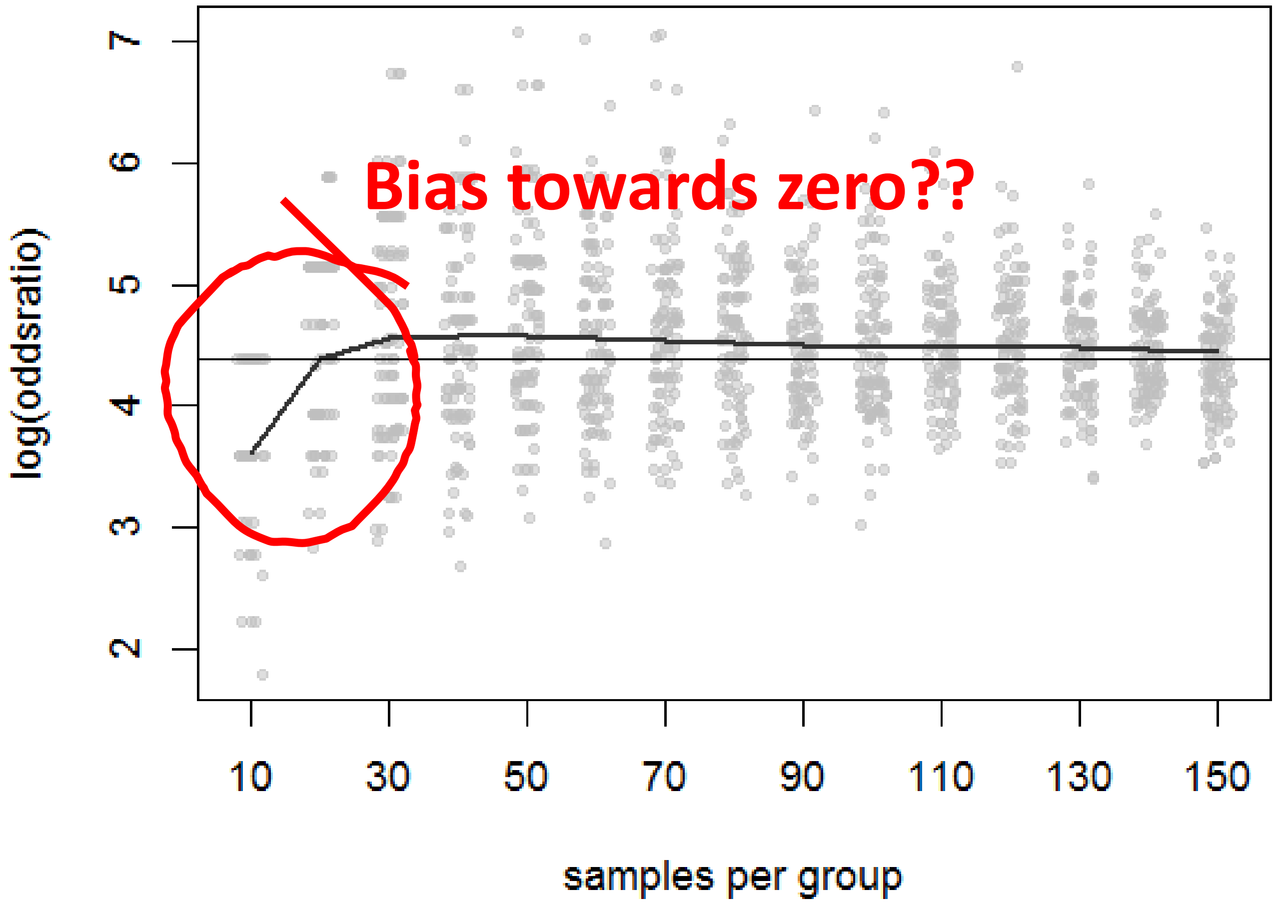
Y	A	B
0	0.1	0.9
1	0.9	0.1



Example: 2 × 2 table

Two groups with event probabilities 0.9 and 0.1.
ML-Estimates:

		X	
		A	B
Y	0	0.1	0.9
	1	0.9	0.1



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:

complete separation

	A	B
0	0	10
1	10	

Firth

	A	B
0	0.5	10.5
1	10.5	0.5

quasi-complete separation

	A	B
0	0	7
1	10	

Firth

	A	B
0	0.5	7.5
1	10.5	.5

PREMA

We are interested in

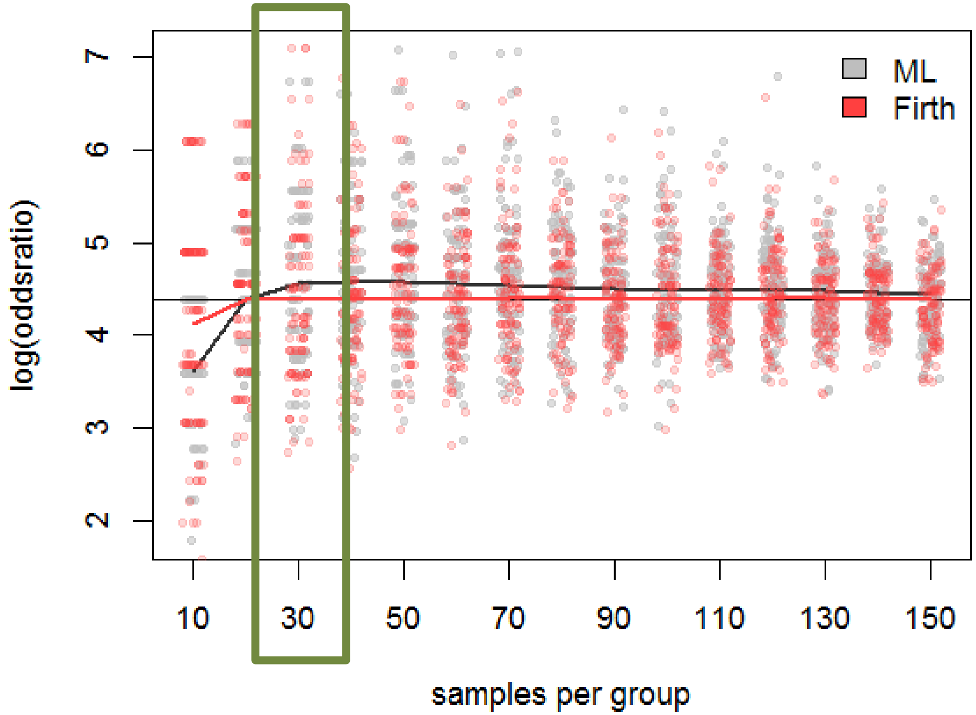
accurate prediction of rare events

in particular in the presence of **high-dimensional data**.

What can we expect from Firth's method?

Accurate prediction

Recall the example:

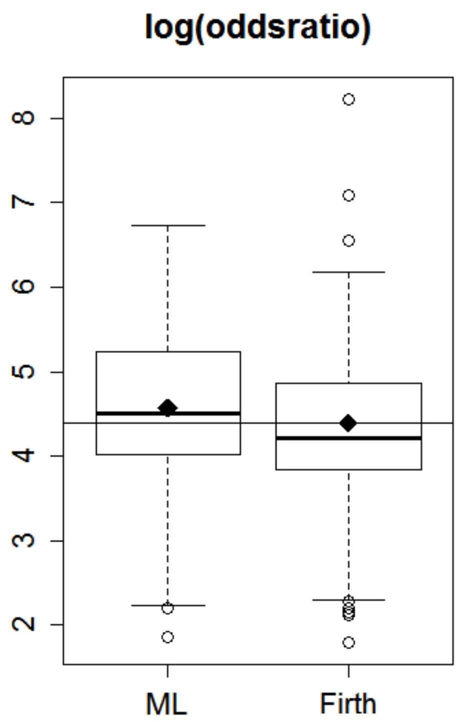


	X	
Y	A	B
0	0.1	0.9
1	0.9	0.1

Now we take a closer look at n=30…

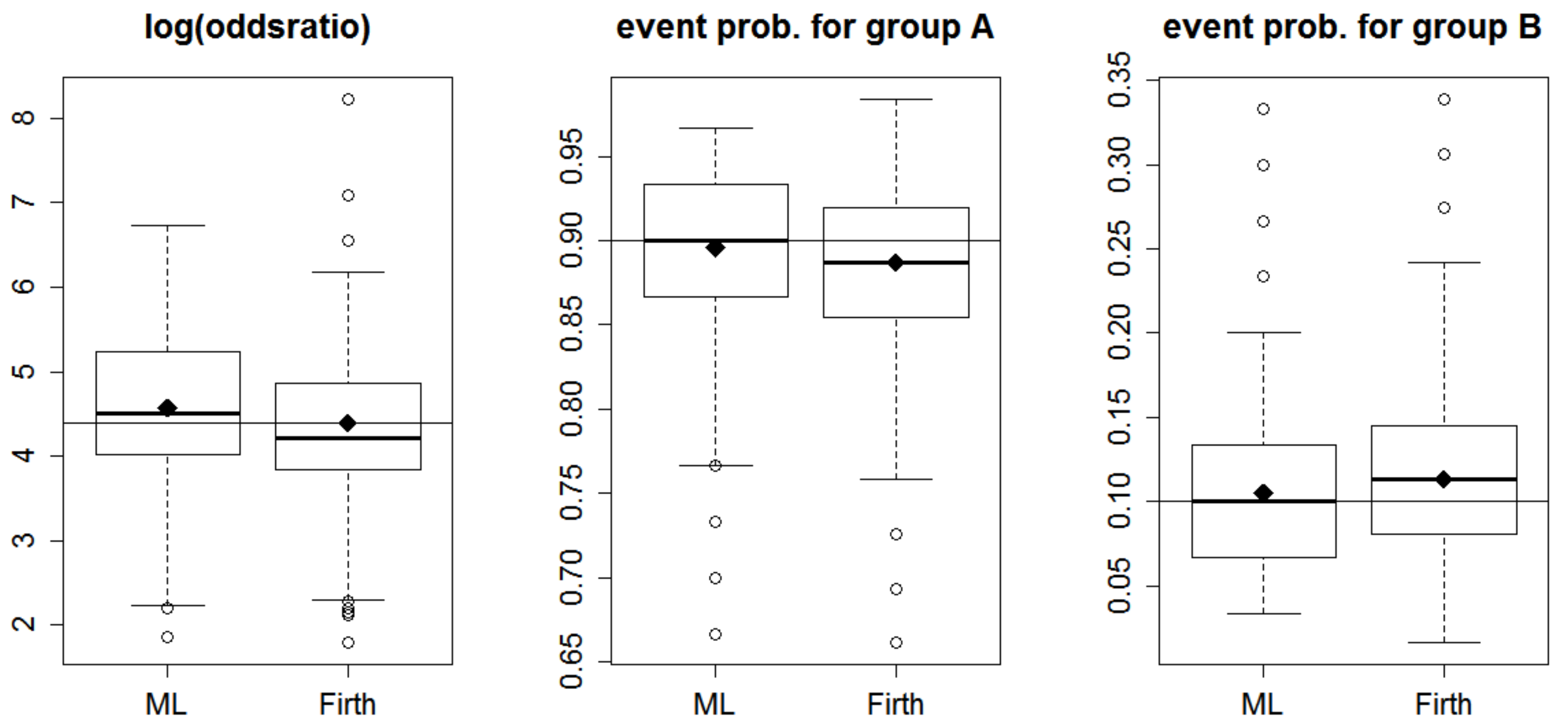
Accurate prediction

Firth's method aims at removing the bias of the coefficients.



Accurate prediction

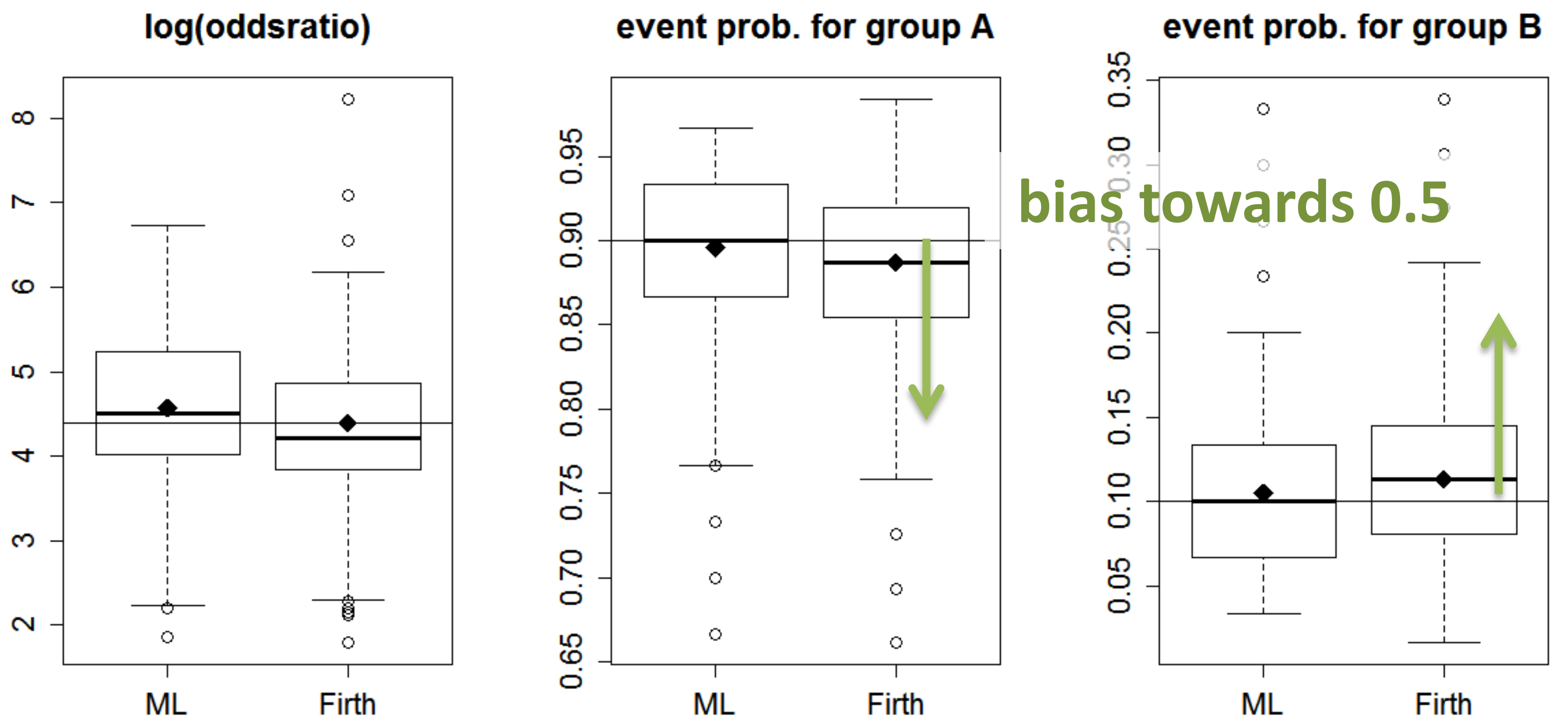
Firth's method aims at removing the bias of the coefficients.
Though, this results in biased event probabilities.



for ML estimates 791 out of 10000 scenarios were excluded due to separation

Accurate prediction

Firth's method aims at removing the bias of the coefficients. Though, this results in biased event probabilities.

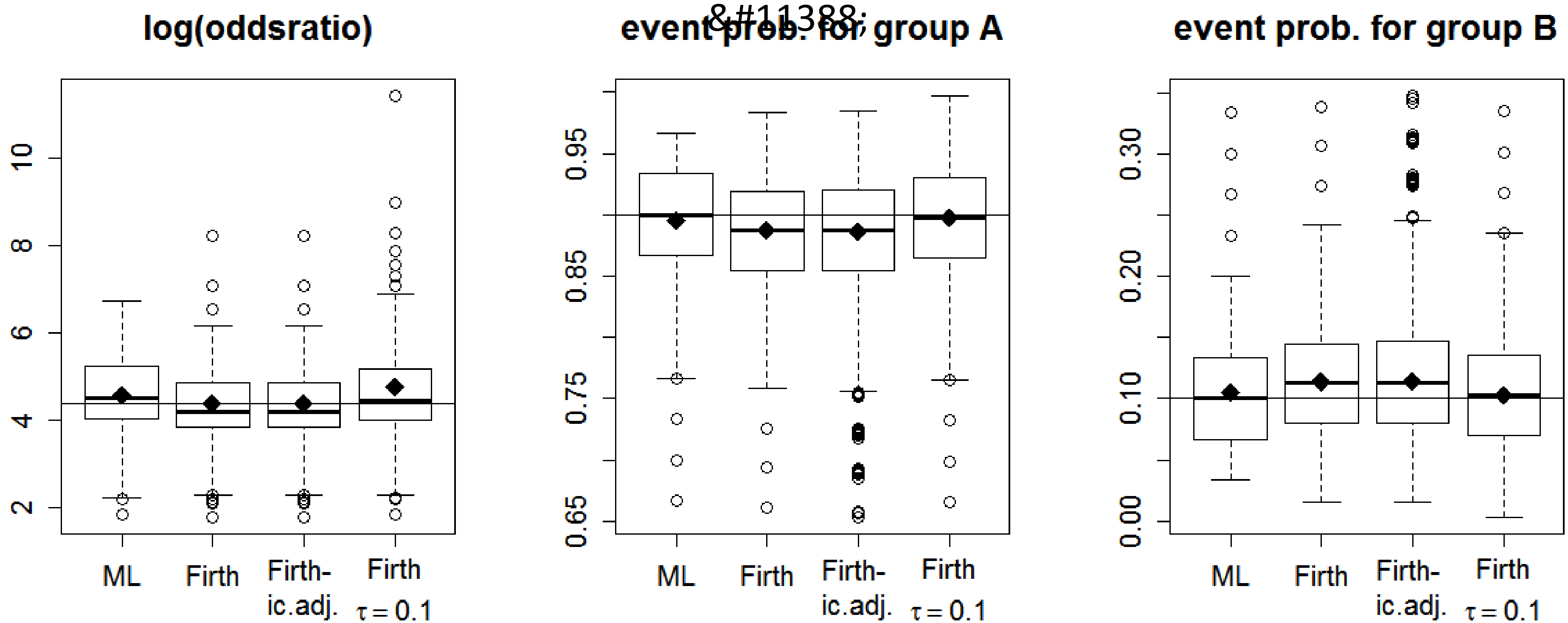


for ML estimates 791 out of 10000 scenarios were excluded due to separation

Accurate prediction

Approaches for unbiased event-probabilities:

- Puhr R and Heinze G: **adjust the intercept**, such that the mean predicted probability is equal to the proportion of events
- Elgmati E et al.: **weaken the Firth type penalty** (replace 0.5 by factor < 0.5), for instance $\tau = 0.1$



Accurate prediction

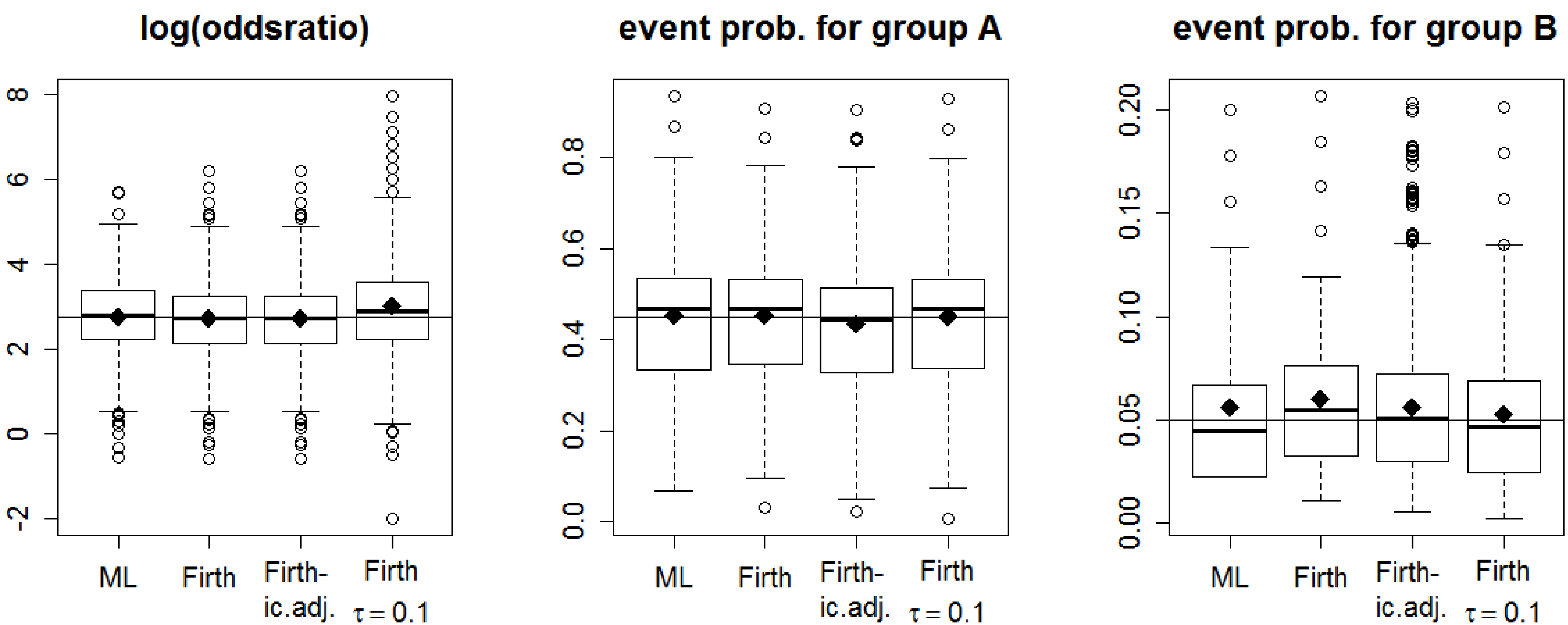
With rare events:

- group A: 45% events, N=15
- group B: 5% events, N=45

  → 15% events in total,
  separation in ~10% of scenarios

X

		A	B
Y	0	0.55	0.95
	1	0.45	0.05
		N=15	N=45



High dimensional data

If $n \ll p$ then, in general, the sample outcome can be perfectly predicted.

Just another case of complete separation?

Unfortunately, Firth estimates are **not unique for $\boldsymbol{n < p}$**.

(For the same reason why ML estimates are not unique.)

However, ridge and LASSO estimates give “reasonable” results for $n < p$.

⇨ Combine ridge or LASSO with Firth?

Combination of Firth's method and ridge

Shen and Gao (2008), for $n > p$ **:**

$$l^*(\theta) = l(\theta) + \underbrace{\frac{1}{2}\log(\det(I(\theta)))}_{\text{Firth penalty}} - \underbrace{\lambda \, \|\theta\|^2}_{\text{ridge penalty}}$$

Motivation: to deal with multicollinearity AND separation

Conclusion: reduces MSE but introduces bias of coefficients

~~Conclusions~~ QUESTIONS

- Other modifications of Firth’s penalty favouring accurate prediction of event probabilities?
- Performance of these modifications in combination with weighting, tuning? In the situation of rare events?
- Combination of Firth’s and ridge for high-dimensional data?
- Combination of Firth’s method and LASSO? In high-dimensional data?
- Tuning Firth’s penalty?

Literature

- Elgmati E, Fiaccone RL, Henderson R and Matthews JNS. Penalised logistic regression and dynamic prediction for discrete-time recurrent event data. Lifetime Data Analysis 2015; doi: 10.1007/s10985-015-9321-4.
- Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993; 80(1): 27-38.
- Heinze G and Schemper M. A solution to the problem of separation in logistic regression. Statistics in Medicine 2002; 21(16): 2409-2419.
- Puhr R and Heinze G. Predicting rare events with penalized logistic regression. Work in progress.
- Shen J and Gao S. A solution to separation and multicollinearity in multiple logistic regression. Journal of Data Science 2008; 6(4): 515-531.