Ridge regression – a solution to separation in logistic regression?

Hana Šinkovec[1], Angelika Geroldinger[2], Rok Blagus[3], Georg Heinze[1]

[1] Medical University of Vienna, Austria
[2] INSERM, UMRS 1138, CRC, Paris, France
[3] University of Ljubljana, Slovenia

Supported by the Austrian Science Fund FWF, Project I2276-N33

Our interests and methods

- the relationship between a binary outcome variable and covariates *X*

  Y=1 (event)

  Y=0 (non-event)

- prediction of binary outcome → logistic regression

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

- estimation of the parameters → maximum likelihood (ML)

$$\ell(\beta) = \log L(\beta) = \sum_{i}^{n} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]$$

- $\exp(\beta)$ = *odds ratio* should be interpretable

Separation

- under certain conditions:
- small/sparse data set
- rare outcomes/exposures
- covariates with strong correlations/effects
- Example:

complete separation

	1	**0**
A	15	0
B	0	15

quasi-complete separation

	1	**0**
A	12	3
B	15	0

→ events and non-events are perfectly separated by the values of a covariate or a linear combination of covariates

ML parameter estimates:

$$\hat{\beta} = \log\left(\frac{f_{11}f_{22}}{f_{12}f_{21}}\right)$$

does not exist!

A Solution

STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2409–2419 (DOI: 10.1002/sim.1047)

A solution to the problem of separation in logistic regression

Georg Heinze*,† and Michael Schemper

Section of Clinical Biometrics, Department of Medical Computer Sciences, University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

SUMMARY

The phenomenon of separation or monotone likelihood is observed in the fitting process of a logistic model if the likelihood converges while at least one parameter estimate diverges to $\pm$ infinity. Separation primarily occurs in small samples with several unbalanced and highly predictive risk factors. A procedure by Firth originally developed to reduce the bias of maximum likelihood estimates is shown to provide an ideal solution to separation. It produces finite parameter estimates by means of penalized maximum likelihood estimation. Corresponding Wald tests and confidence intervals are available but it is shown that

A Solution

STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2409–2419 (DOI: 10.1002/sim.1047)

A solution to the problem of separation in logistic regression

Georg Heinze*,† and Michael Schemper

Section of Clinical Biometrics, Department of Medical Computer Sciences, University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

SUMMARY

The phenomenon of separation or monotone likelihood is observed in the fitting process of a logistic model if the likelihood converges while at least one parameter estimate diverges to ± infinity. Separation primarily occurs in small samples with several unbalanced and highly predictive risk factors. A procedure by Firth originally developed to reduce the bias of maximum likelihood estimates is shown to provide an ideal solution to separation. It produces finite parameter estimates by means of penalized maximum likelihood estimation. Corresponding Wald tests and confidence intervals are available but it is shown that

A Solution

ACCEPTED MANUSCRIPT

Separation in Logistic Regression – Causes, Consequences, and Control

Mohammad Ali Mansournia, Angelika Geroldinger, Sander Greenland, Georg Heinze

American Journal of Epidemiology, kwx299, https://doi.org/10.1093/aje/kwx299
Published: 17 August 2017

Abstract

Separation is encountered in regression models with a discrete outcome (such as logistic regression) where the covariates perfectly predict the outcome. It is most frequent under the same conditions that lead to small-sample and sparse-data bias, such as presence of a rare outcome, rare exposures, highly correlated covariates, or covariates with strong effects. In theory separation will produce infinite estimates for some coefficients. In practice however separation may be unnoticed or mishandled because of software limits in recognizing and handling the problem, and notifying the user. We discuss causes of separation in logistic regression and describe how common software packages deal with it. We then describe methods that remove separation, focusing on the same penalized-likelihood techniques used to address more general sparse-data problems. These methods improve accuracy, avoid software problems, and allow interpretation as Bayesian analyses with weakly informative priors. We discuss likelihood penalties and their relative advantages and disadvantages, including some that can be implemented easily with any software package. We illustrate ideas and methods using a case-control study of contraceptive practices and urinary tract infection.

Penalized likelihood logistic regression

- intended to provide shrinkage of the parameter estimates → parameter estimates do not diverge

$$\ell^P(\beta) = \log L(\beta) + P(\beta)$$

- Firth: $$P(\beta) = \frac{1}{2} \log \det(I(\beta))$$

Penalized likelihood logistic regression

- intended to provide shrinkage of the parameter estimates → parameter estimates do not diverge

$$\ell^P(\beta) = \log L(\beta) + P(\beta)$$

- Firth: $P(\beta) = \frac{1}{2}\log \det(I(\beta))$

- Ridge: $P(\beta) = -\lambda \sum \beta^2$
- LASSO: $P(\beta) = -\lambda \sum |\beta|$

- λ is usually optimized by cross-validating the deviance

Real data example

The histology of endometrium (HG):

- n=30 grading 0–II -> HG=0
- n=49 grading III–IV -> HG=1

can be explained by:

- neovasculization (NV):
  - present for n=13 and absent for n=66;
- pulsatility index of arteria uterina (PI):
  - median=16 (range: 0–49)
- endometrium height (EH):
  - median=1.64 (range: 0.27–3.61)

Estimating the model by ML

```
Call:
glm(formula = HG ~ NV + PI + EH, family = "binomial", data = asser)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50137  -0.64108  -0.29432   0.00016   2.72777

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.30452    1.63730   2.629 0.008563 **
NV           18.18556 1715.75089   0.011 0.991543
PI           -0.04218    0.04433  -0.952 0.341333
EH           -2.90261    0.84555  -3.433 0.000597 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Estimating the model by ML

```
Call:
glm(formula = HG ~ NV + PI + EH, family = "binomial", data = asser)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50137  -0.64108  -0.29432   0.00016   2.72777

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.30452    1.63730   2.629 0.008563 **
NV          18.18556 1715.75089   0.011 0.991543
PI          -0.04218    0.04433  -0.952 0.341333
EH          -2.90261    0.84555  -3.433 0.000597 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

??

Estimating the model by ML

```
Call:
glm(formula = HG ~ NV + PI + EH, family = "binomial", data = asser)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50137  -0.64108  -0.29432   0.00016   2.72777

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.30452    1.63730   2.629 0.008563 **
NV          18.18556 1715.75089   0.011 0.991543
PI          -0.04218    0.04433  -0.952 0.341333
EH          -2.90261    0.84555  -3.433 0.000597 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of Fisher Scoring iterations: 17
```

	HG=1	HG=0
NV=0	17	49
NV=1	13	0

```
> model.ml$converged
[1] TRUE
```

-> only likelihood ‚converged‘, not the parameters!

Estimating the model by Firth

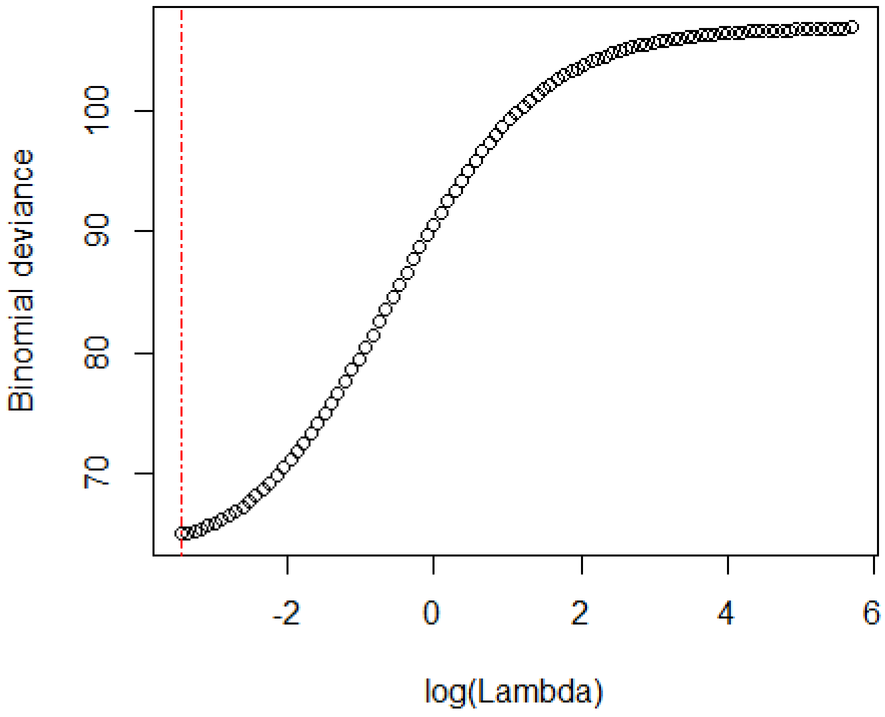
```
logistf(formula = HG ~ NV + PI + EH, data = asser, family = "binomial")

Model fitted by Penalized ML
Confidence intervals and p-values by Profile Likelihood

                   coef   se(coef) lower 0.95  upper 0.95      Chisq            p
(Intercept)  3.77455968 1.48869166  1.0825417  7.20928050  8.1980136 4.193628e-03
NV           2.92927334 1.55076372  0.6097274  7.85463171  6.7984572 9.123668e-03
PI          -0.03475176 0.03957815 -0.1244587  0.04045547  0.7468285 3.874822e-01
EH          -2.60416391 0.77601764 -4.3651832 -1.23272106 17.7593175 2.506867e-05

Likelihood ratio test=43.65582 on 3 df, p=1.78586e-09, n=79
Wald test = 17.47967 on 3 df, p = 0.0005630434
```

Estimating the model using (tuned) ridge regression

```
> model.cv<-cv.glmnet(y=asser$HG, x=x, family="binomial", nfolds=nrow(asser), alpha=0)
Warning message:
Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per fold
> coef(model.cv, s="lambda.min")
4 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept)  2.47125685
NV           2.47998054
PI          -0.01756067
EH          -1.91025366
```

Estimating the model using (tuned) ridge regression

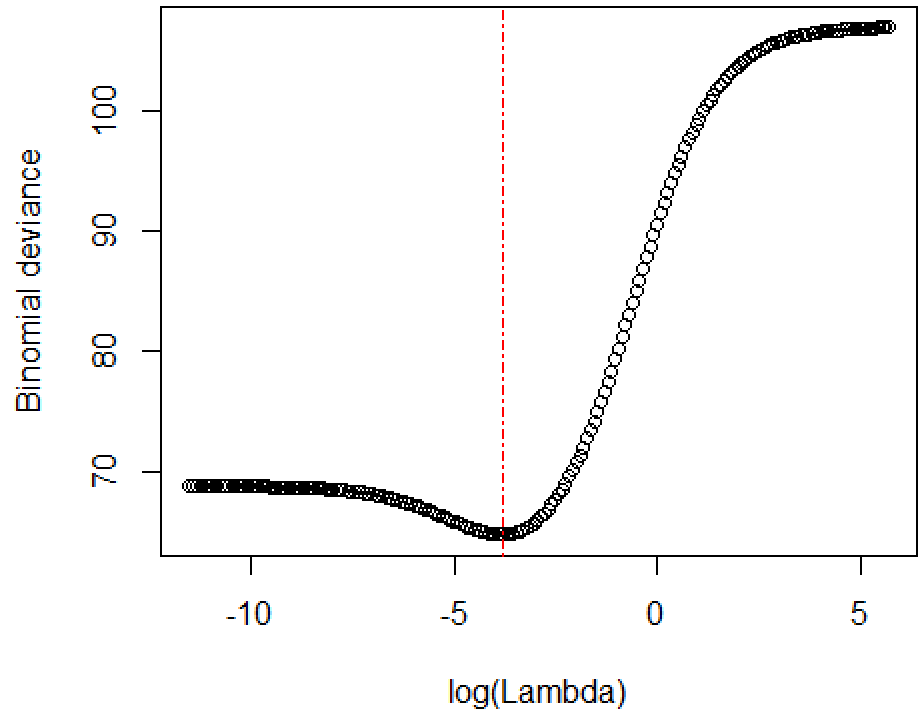
```
> plot(model.cv)
```



-> converged, but at lowest lambda

Extending the range of λ

```
> plot(model.cv2)
> coef(model.cv2, s="lambda.min")
4 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept)  2.80892893
NV           2.73293270
PI          -0.02126937
EH          -2.10097872
```

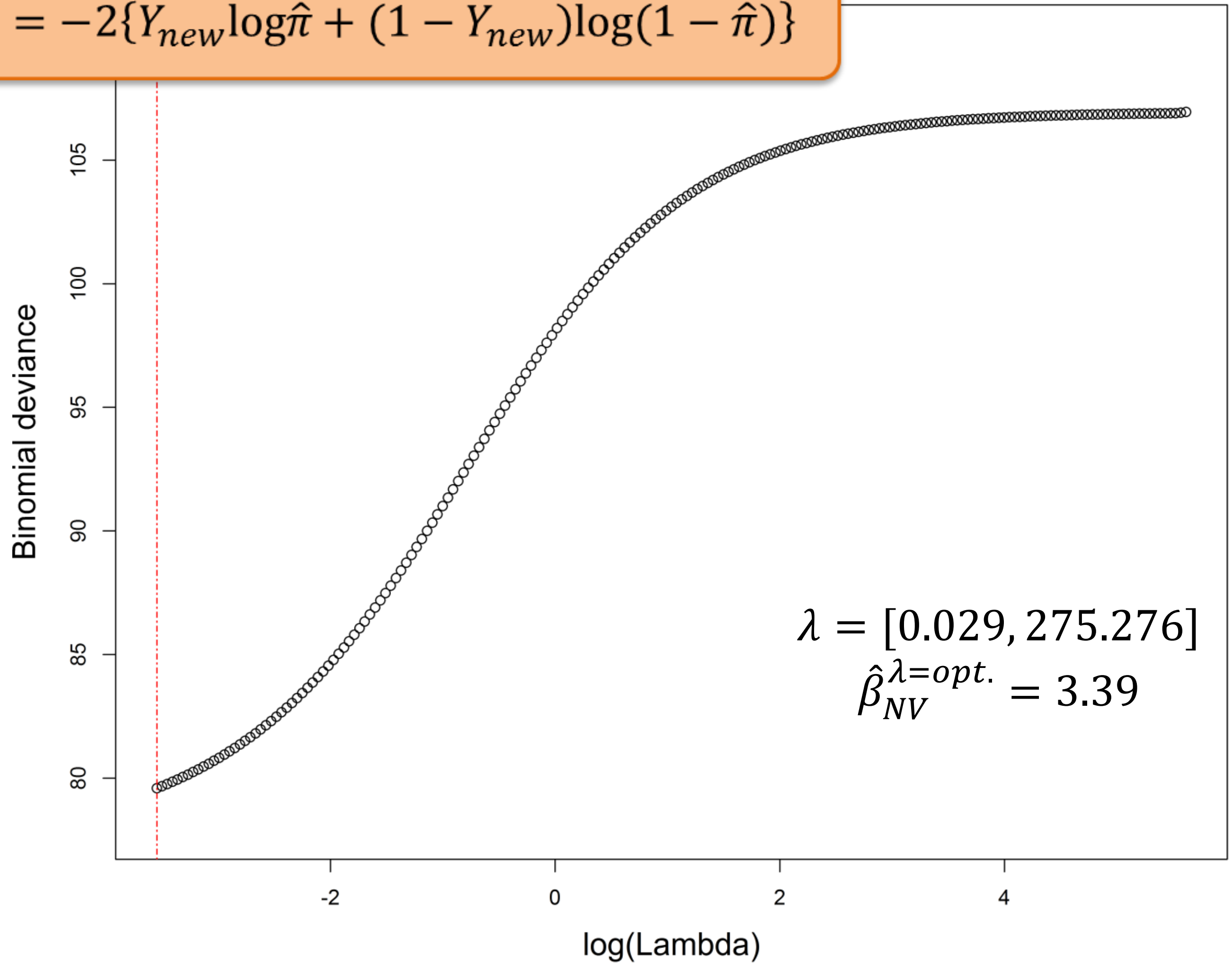


‚real convergence‘!

Is it a solution to separation?

Univariable model for $\hat{\beta}_{NV}$

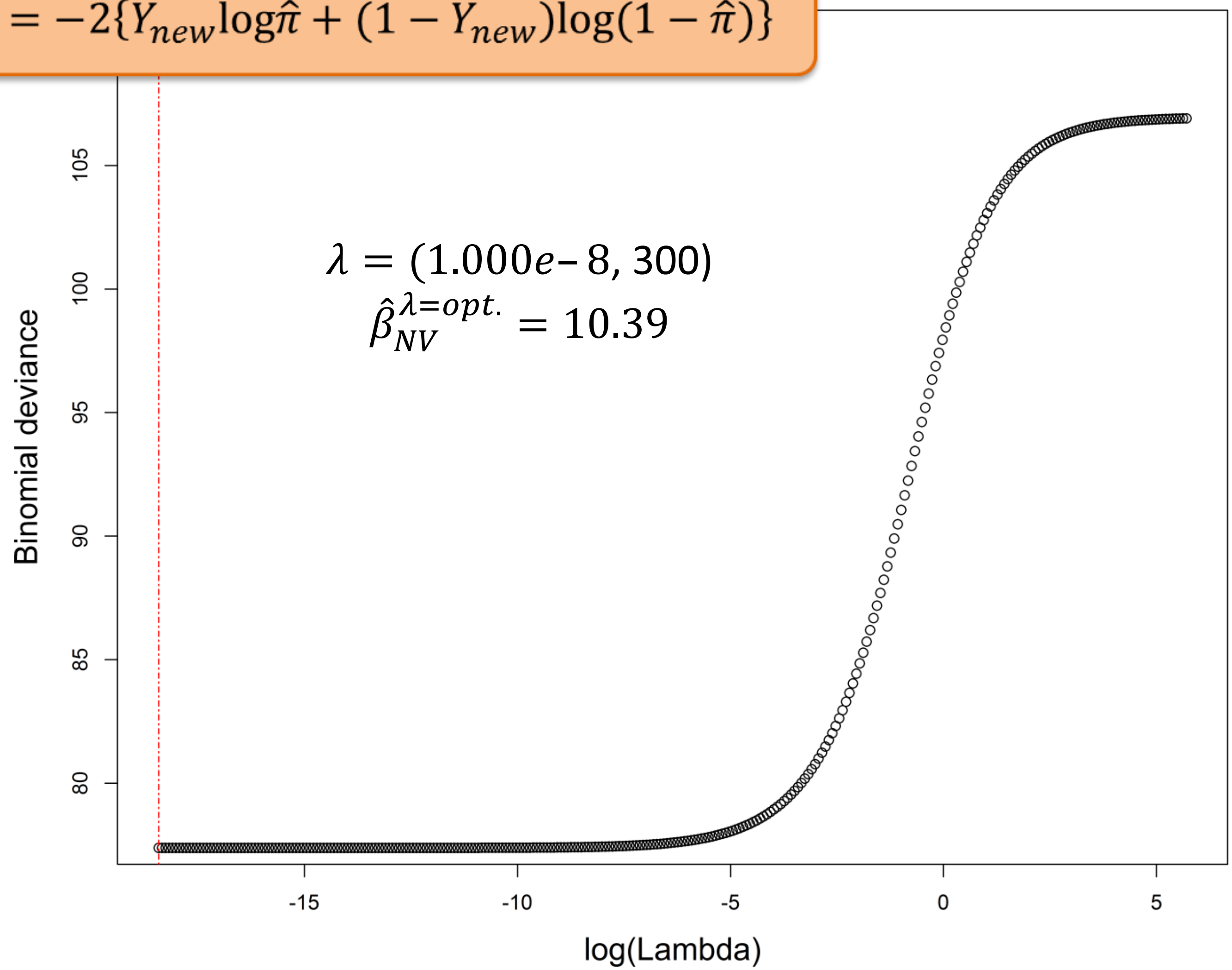
$$D_i = -2\{Y_{new}\log\hat{\pi} + (1 - Y_{new})\log(1 - \hat{\pi})\}$$



$$\lambda = [0.029, 275.276]$$
$$\hat{\beta}_{NV}^{\lambda=opt.} = 3.39$$

Univariable model for $\hat{\beta}_{NV}$

$$D_i = -2\{Y_{new}\log\hat{\pi} + (1 - Y_{new})\log(1 - \hat{\pi})\}$$



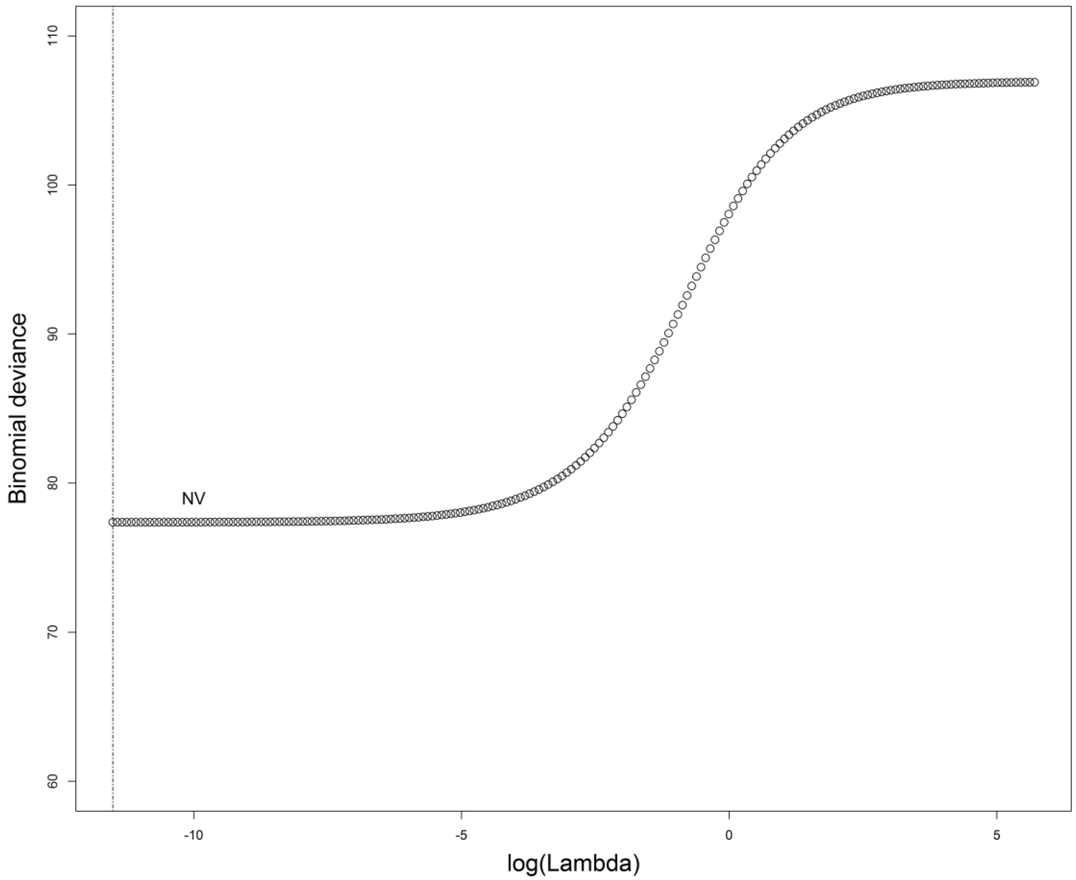
Intermediate conclusion

- For the multivariable model,
  ridge regression converged

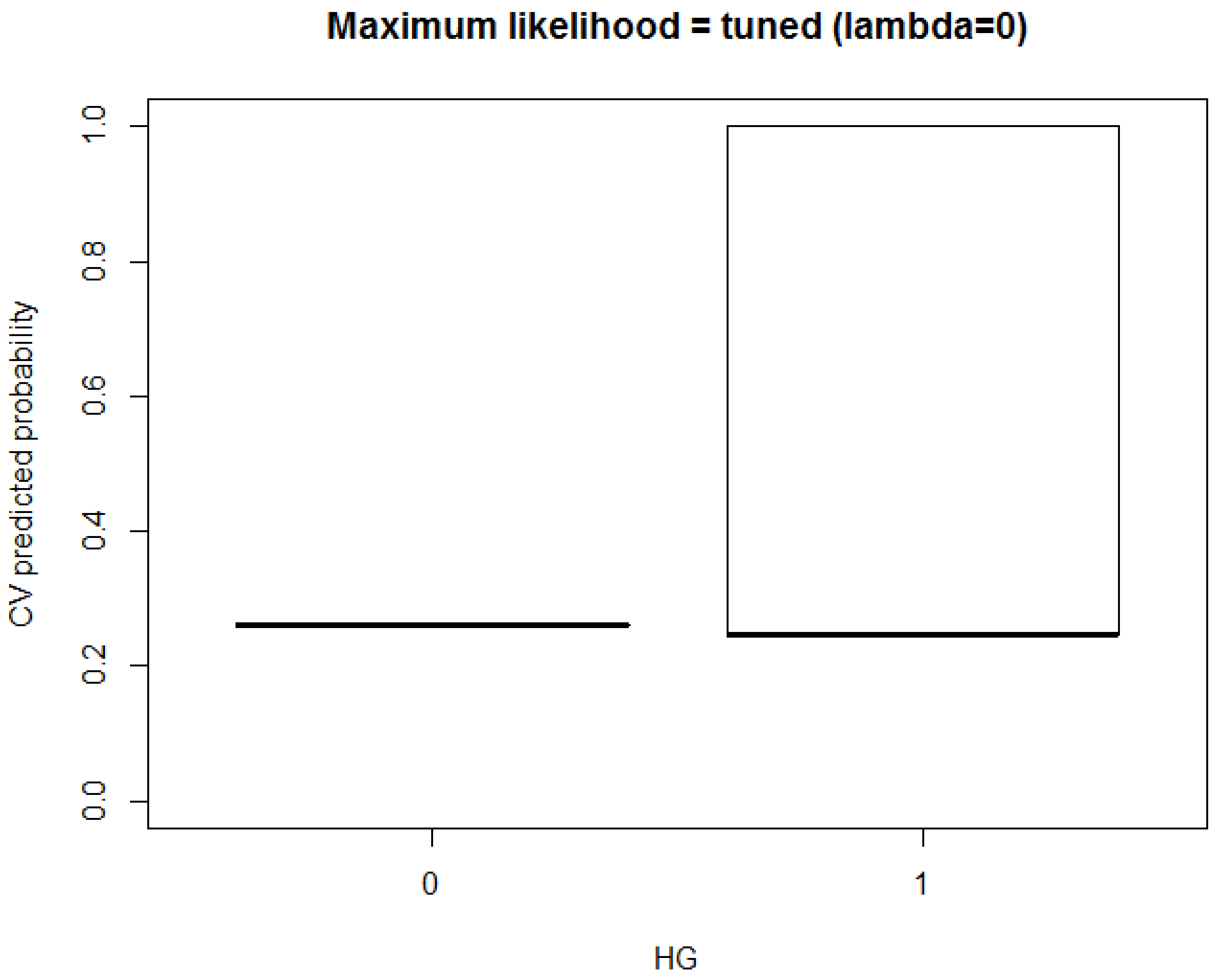
- For the univariable model,
  ridge regression did not converge

- Why does this happen?

Univariable model with NV only

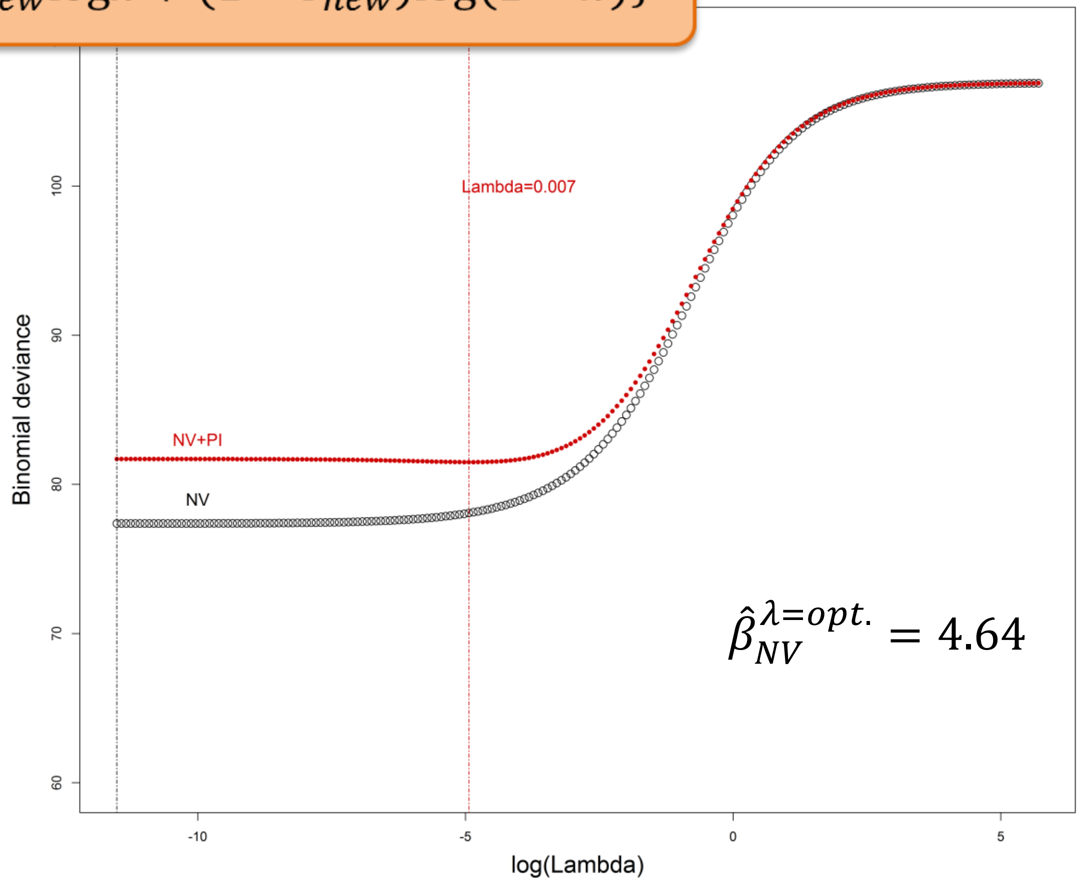


Cross-validating predicted probabilities



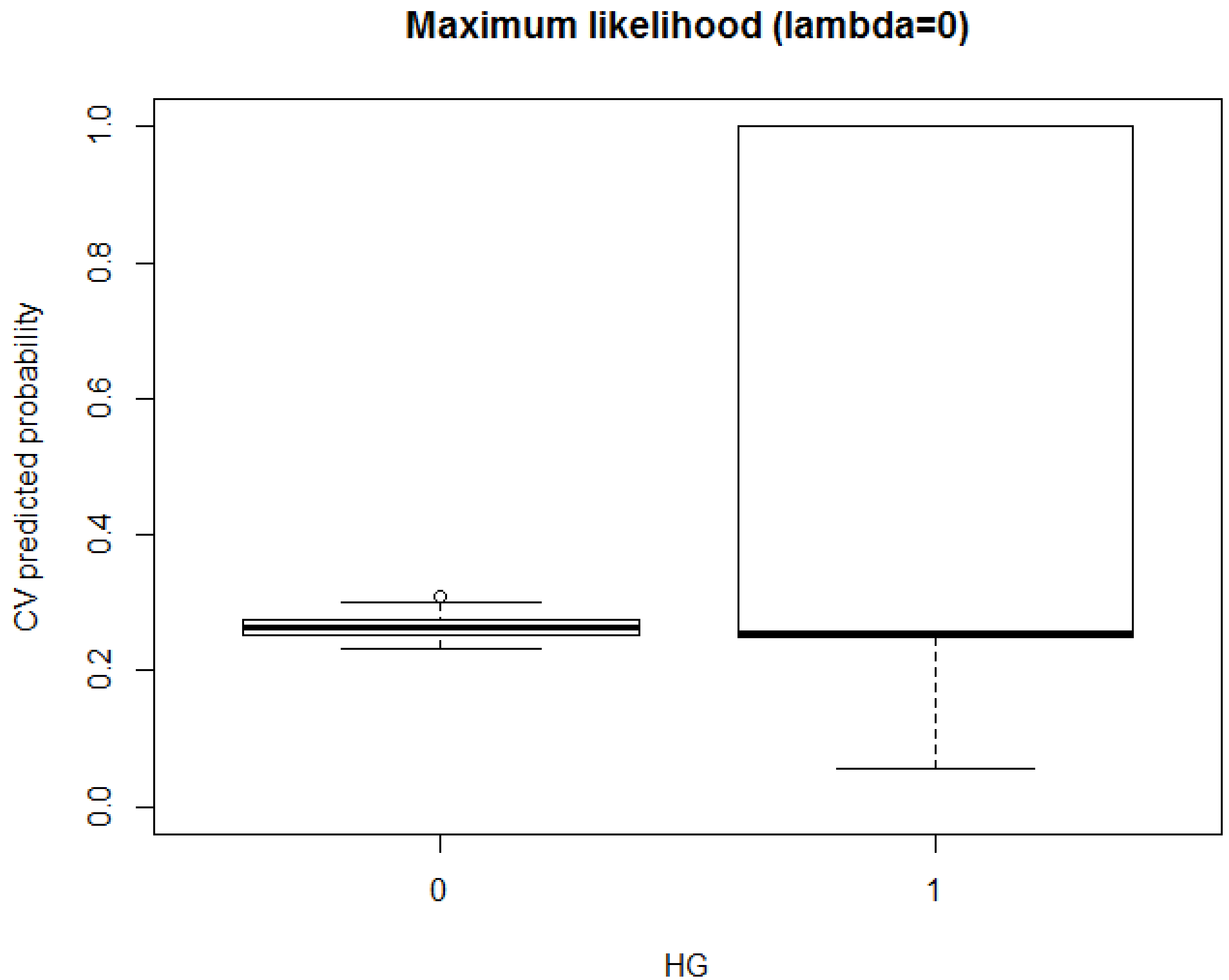
Bivariable model: NV+PI

$$D_i = -2\{Y_{new}\log\hat{\pi} + (1 - Y_{new})\log(1 - \hat{\pi})\}$$

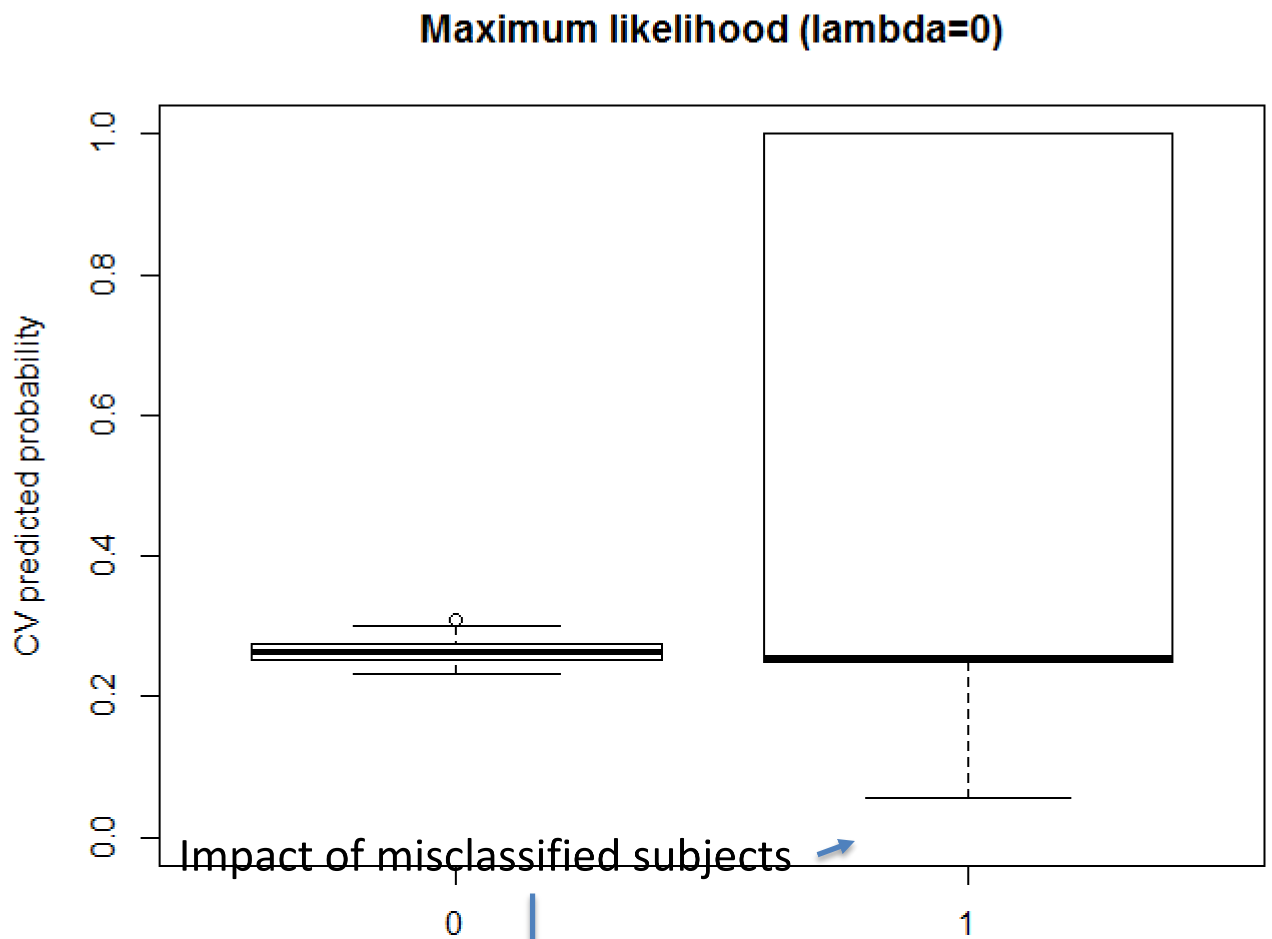


$$\hat{\beta}_{NV}^{\lambda=opt.} = 4.64$$

Cross-validating predicted probabilities



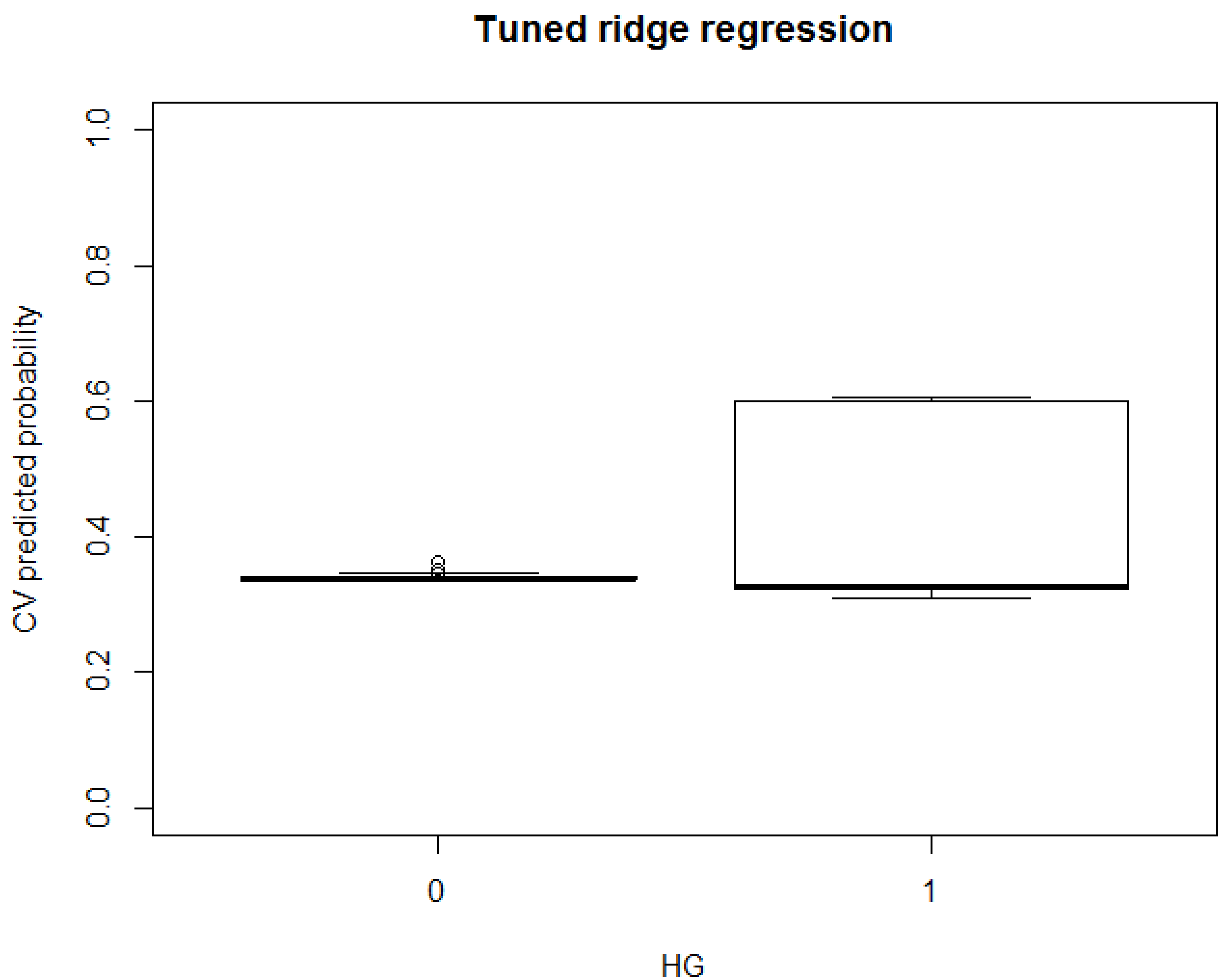
Cross-validating predicted probabilities



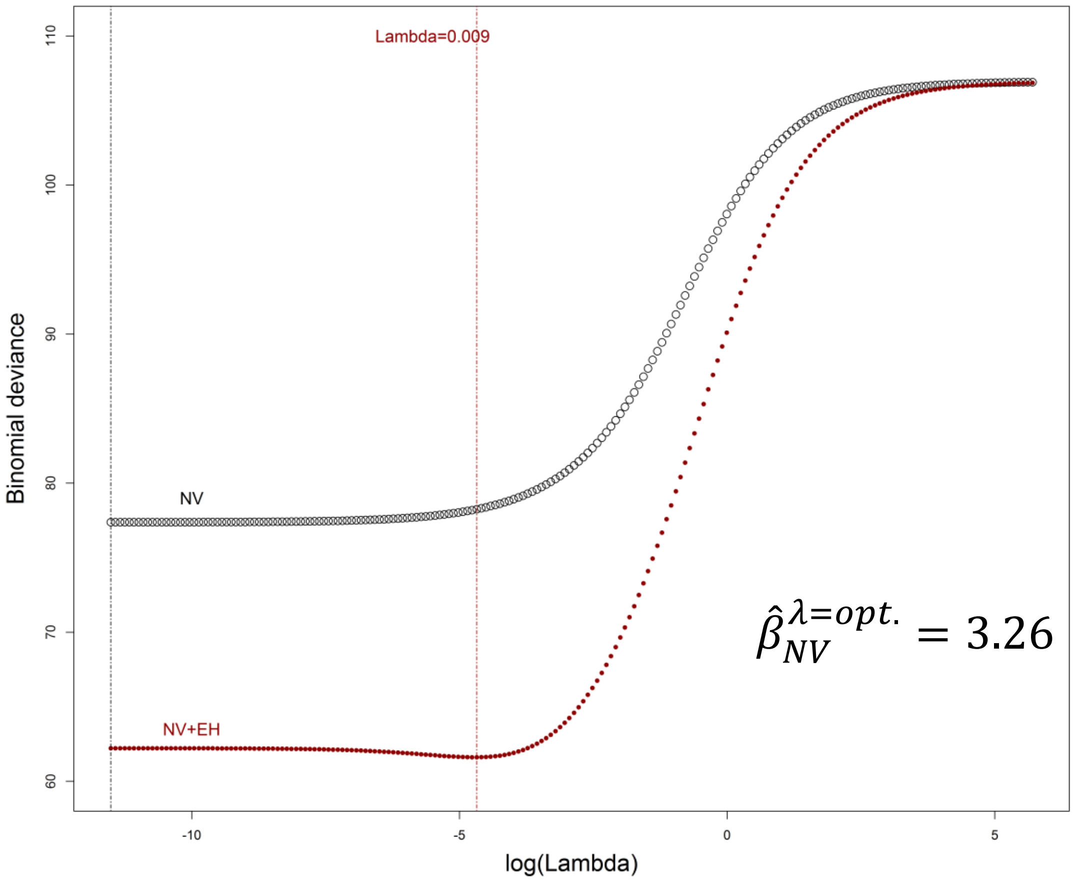
Impact of misclassified subjects

$$D_i = -2\{Y_{new}\log\hat{\pi} + (1 - Y_{new})\log(1 - \hat{\pi})\}$$

Cross-validating predicted probabilities

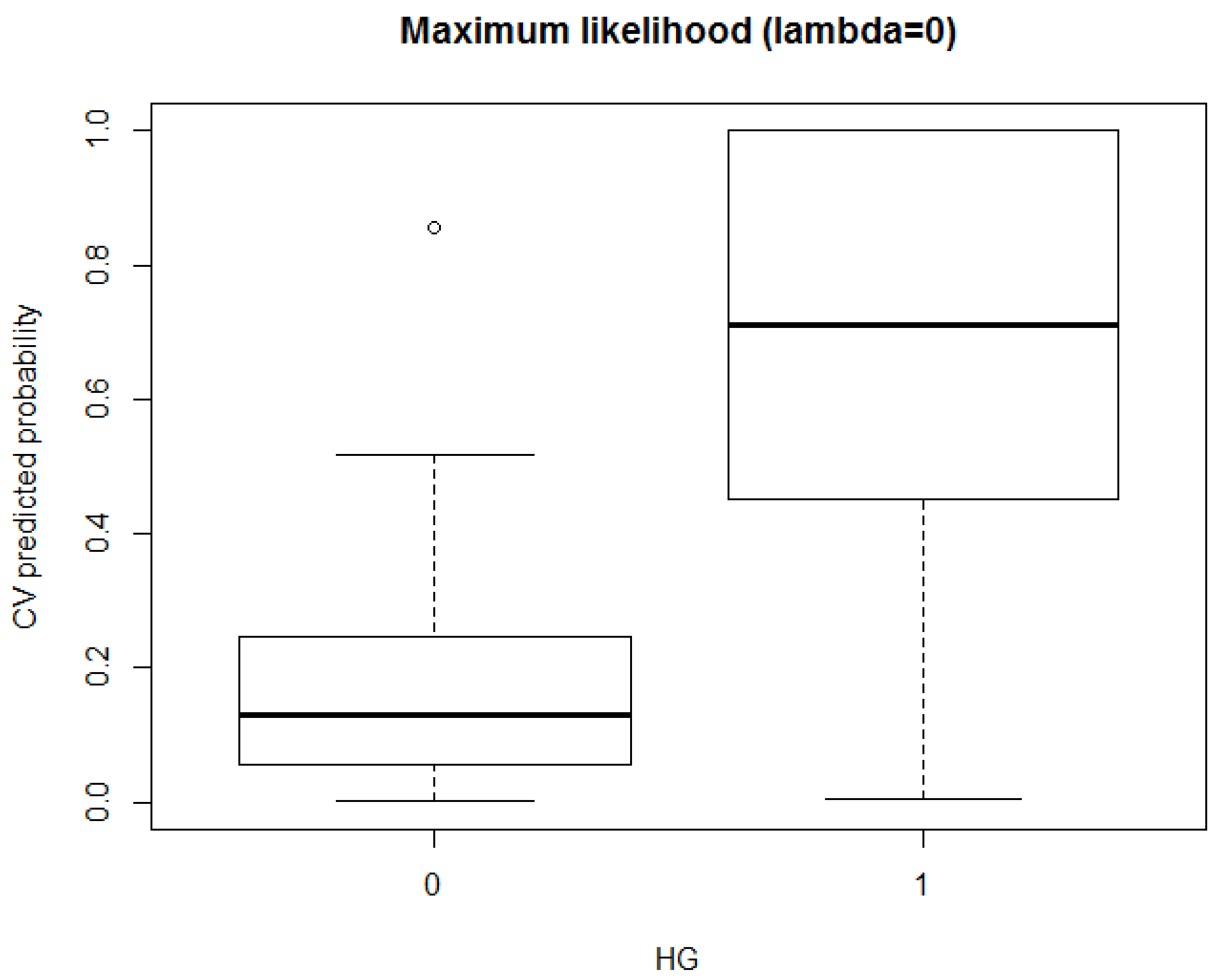


Bivariable model: NV + EH

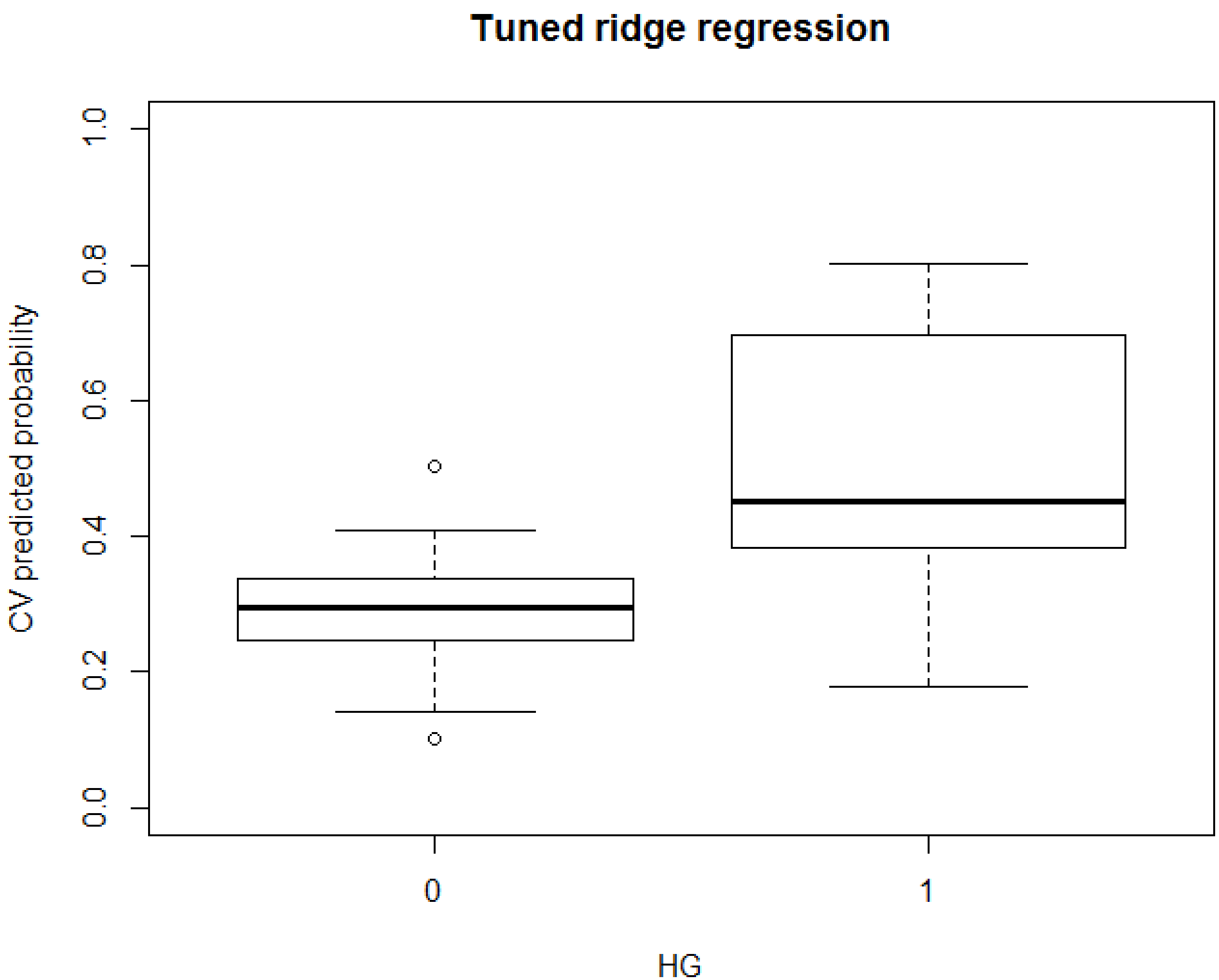


$$\hat{\beta}_{NV}^{\lambda=opt.} = 3.26$$

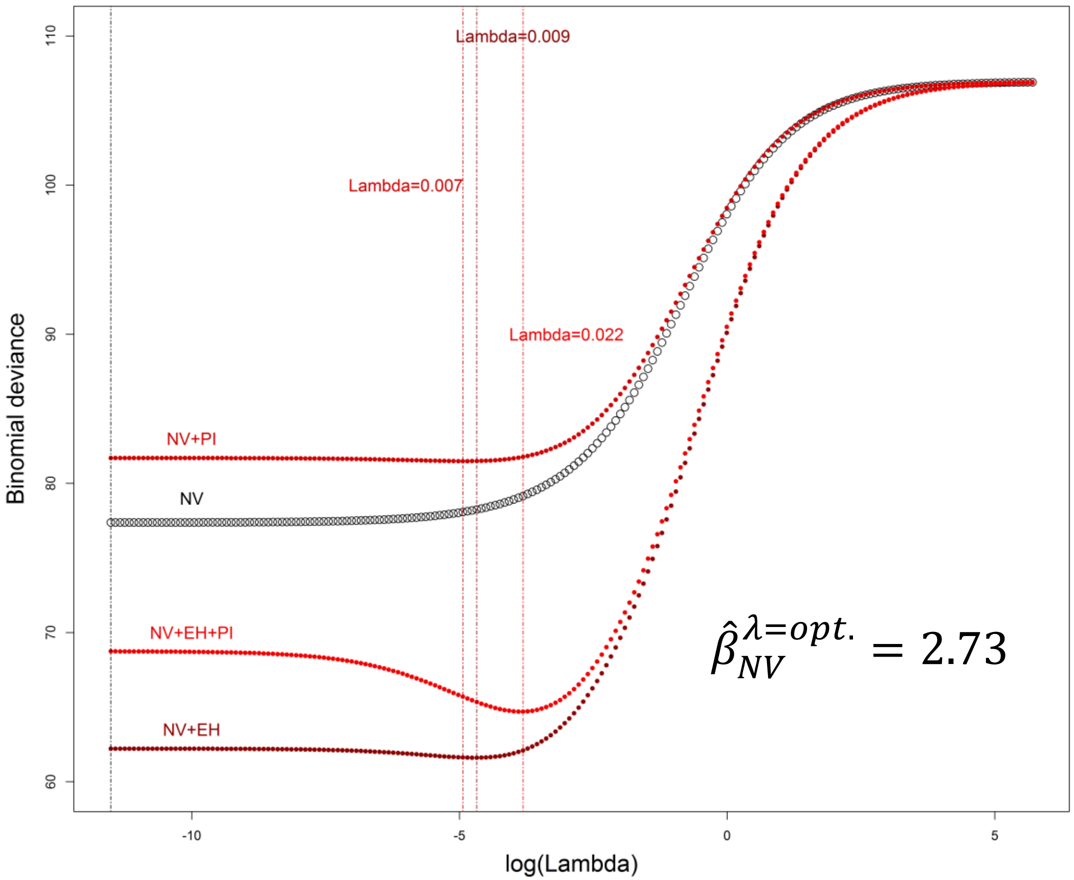
Cross-validating predicted probabilities



Cross-validating predicted probabilities

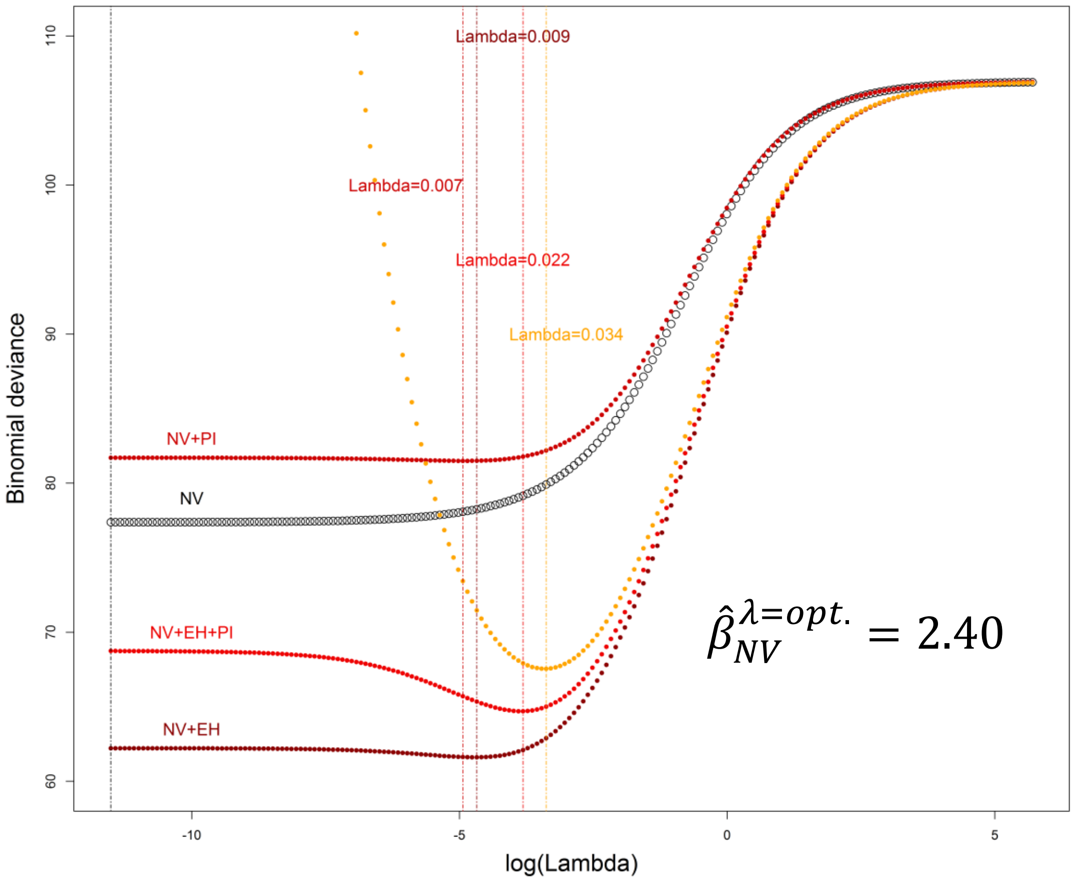


Multivariable model: NV+EH+PI



$$\hat{\beta}_{NV}^{\lambda=opt.} = 2.73$$

Adding 10 noise predictors



Conclusions

- Trouble comes with optimizing λ

- Pre-specifying the value of λ always yields convergence

- ‚*Adding noise -> convergence*‘:
  If you have a perfect predictor,
  and you add noise to it,
  tuned ridge regression will shrink it

- Adding covariates changes λ (and $\hat{\beta}$…)
- Unless there is a lot of noise, the optimized λ is arbitrary

Further work

- For 2x2 table with separation, we have proven that
  $D^{\lambda=0,CV} \leq D^{\lambda=\infty,CV}$,
  with strict inequality in all real examples.
- Still to prove that this holds for any $\lambda > 0$.

- What are the empirical properties of the obtained solution?

References

- Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.
- Heinze, G. and Schemper, M. (2002), A solution to the problem of separation in logistic regression. Statist. Med., 21: 2409–2419.
- Georg Heinze, Meinhard Ploner, Daniela Dunkler and Harry Southworth (2014). logistf: Firth's bias reduced logistic regression. R package version 1.22.
- Mohammad Ali Mansournia, Angelika Geroldinger, Sander Greenland, Georg Heinze; Separation in Logistic Regression – Causes, Consequences, and Control, *American Journal of Epidemiology*, kwx299, https://doi.org/10.1093/aje/kwx299