

# **Part II: Generalized Linear Models**

## **Chapter II.7**

### **Penalized Regression Models**


## ▶ To be discussed...

- ▶ High-Dimensional Data
- ▶ Penalized Likelihood Methods
- ▶ The Lasso
- ▶ Common Penalty Functions
- ▶ Complete Separation in Logistic Regression

## 7. Penalized Regression Models

### > II.7.1 High-Dimensional Data

Some data sets are high-dimensional, having a very large number  $p$  of explanatory variables and potential model parameters.

 High-dimensional data are common in application areas such as genomics, biomedical imaging, market basket data, and portfolio allocation in finance.

For model fitting, high-dimensional data are not well handled by classical ML methods.

### > II.7.2 Regularization methods

When  $p$  is very large, sometimes even  $p > n$ , a vital issue for model-building is selection of explanatory variables. The ML estimates of GLM parameters can be highly unstable or not even exist. *Regularization methods* are ways of estimating effects *under the assumption that the true model is sparse*, with most of the explanatory variables expected to have no effect or a practically insignificant effect on the response.

### ➤ II.7.3 Corrected AIC and BIC

- With regard to model selection criteria, AIC can lead to variable-selection bias in a high-dimensional setup. Corrected versions are available, such as the AICc function in the `gamlr` R package, which is preferred when  $n < 40p$ .
- Also corrected versions of BIC, such as an extended BIC (EBIC) are more appropriate than BIC. R packages that perform high-dimensional variable selection based on EBIC include `bestglm` and `BeSS`.

# Penalized Likelihood Methods

## ➤ II.7.4 Definition (Penalized Likelihood)

The penalized likelihood method adds a term to the log-likelihood function such that the values that maximize it are smoothings of the ordinary ML estimates, typically shrinking them toward 0. For a model with log-likelihood function  $L(\boldsymbol{\beta})$ , the method maximizes

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - s(\boldsymbol{\beta}),$$

where  $s$  is a function such that  $s(\boldsymbol{\beta})$  decreases as elements of  $\boldsymbol{\beta}$  are smoother in some sense.

## ➤ II.7.5 Remark

The penalized log-likelihood is expressed in terms of standardized versions of the variables, so that the smoothing function treats each variable in the same way and the degree of smoothing does not depend on the choice of scaling.

# Least Absolute Shrinkage and Selection Operator (Lasso)

(Tibshirani, 1995)

## II.7.6 The Lasso

When  $p$  is very large, often only a few  $\{\beta_j\}$  are practically different from 0. For such cases, a popular smoothing function for penalized likelihood methods is

$$s(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p |\beta_j|$$

for some *smoothing parameter*  $\lambda \geq 0$ . This method is called the *lasso*.

The lasso penalizes an effect  $|\hat{\beta}_j|$  for being large, and the method is equivalent to

$$\operatorname{argmax}_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \quad \text{subject to} \quad \sum_j |\hat{\beta}_j| < t.$$

With this method, many  $\{\hat{\beta}_j\}$  may equal 0.

# Tuning

The smoothing parameter  $\lambda$  is also known as *tuning parameter*.

## ➤ II.7.7 The Role of $\lambda$

- The choice of  $\lambda$  reflects the bias/variance tradeoff.
- Ordinary ML estimates result for  $\lambda = 0$ .
- Increasing  $\lambda$  results in greater shrinkage of  $\{\hat{\beta}_j\}$  toward 0, potentially reducing the variance but increasing the bias. A plot of the lasso estimates as  $\lambda$  increases summarizes how  $\{\hat{\beta}_j\}$  reach 0 and corresponding explanatory variables drop out of the linear predictor.

## ➤ II.7.8 Selection of $\lambda$

We choose a grid of  $\lambda$  values and compute the cross-validation error for each of them. We select the  $\lambda$  value that

- 👉 corresponds to the smallest cross-validation error, or
- 👉 gives the most regularized model such that the cross-validated error is within one standard error of the minimum

## ➤ II.7.9 Example (Students Survey)

The Students data file shows responses to a questionnaire by 60 social science graduate students in an introductory Statistics course at the University of Florida. Here we want to predict a subject's opinion about whether abortion should be legal in the first three months of a pregnancy (1 = yes, 0 = no), using the 14 binary and quantitative variables.

```
> Students <- read.table("Students.dat", header=TRUE)
> fit <- glm(abor ~ gender + age + hsgpa + cogpa + dhome + dres + tv + sport + news +
+           aids + veg + ideol + relig + affirm, family=binomial, data=Students)
> summary(fit)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	10.1014	10.8914	0.927	0.3537
gender	1.0022	1.8655	0.537	0.5911
age	-0.0783	0.1275	-0.615	0.5389
hsgpa	-3.7344	2.8093	-1.329	0.1837
cogpa	2.5113	3.7399	0.671	0.5019
dhome	0.0006	0.0007	0.821	0.4116



*(output continues...)*

dres	-0.3388	0.2954	-1.147	0.2514	
tv	0.2660	0.2532	1.051	0.2934	
sport	0.0272	0.2551	0.107	0.9151	
news	1.3869	0.6987	1.985	0.0471	**
aids	0.3967	0.5664	0.700	0.4837	
veg	4.3213	3.8615	1.119	0.2631	
ideol	-1.6378	0.7892	-2.075	0.0380	**
relig	-0.7246	0.7821	-0.926	0.3542	
affirm	-2.7481	2.6899	-1.022	0.3069	

---

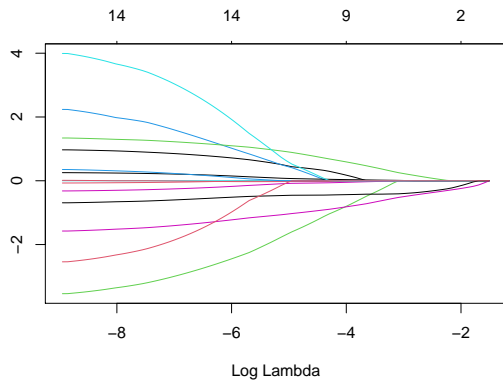
Null deviance: 62.719 on 59 degrees of freedom  
Residual deviance: 21.368 on 45 degrees of freedom

We next use the lasso, implemented in R with the `glmnet` package, which operates on the standardized variables:

```
> x <- cbind(gender, age, hsgpa, cogpa, dhome, dres, tv, sport, news,  
+           aids, veg, ideol, relig, affirm) # explanatory variables for lasso  
> library(glmnet)  
> fit.lasso <- glmnet(x, abor, alpha=1, family="binomial") # alpha=1 selects lasso  
> plot(fit.lasso, "lambda")  
> set.seed(1) # a random seed to implement cross-validation  
> cv <- cv.glmnet(x, abor, alpha=1, family="binomial", type.measure="class")  
cv$lambda.min # best lambda by 10-fold cross-validation  
0.06610251 # a random variable, changes from run to run  
cv$lambda.1se # lambda suggested by one-standard-error rule, a random variable  
0.1267787  
> coef(glmnet(x, abor, alpha=1, family="binomial", lambda=0.1268))  
              s0    # using lambda from one-standard-error rule  
(Intercept)  2.3668 # all 12 lasso estimates that are not shown equal 0  
ideol        -0.2599 # ML estimate is -1.638  
relig        -0.1830 # ML estimate is -0.725
```

```
> coef(glmnet(x, abor, alpha=1, family="binomial", lambda=0.0661))
              s0    # using lambda minimizing CV mean prediction error
(Intercept)  2.6969
news          0.1400  # news effect much less than ML estimate of 1.387
ideol        -0.4239  # ideol and relig effects not shrunk as much toward 0
relig        -0.3603  # as when use one-standard-error rule
```

- The fit using the smoothing parameter value of  $\lambda = 0.1268$  suggested with the one-standard-error rule has only *ideol* (political ideology, higher values being more conservative) and *relig* (how often you attend religious services) as explanatory variables, with estimated negative effects  $-0.260$  and  $-0.183$  on favoring legalized abortion.
- The value  $\lambda = 0.0661$  that has the minimum cross-validated mean prediction error adds *news* as a predictor, and then *ideol* and *relig* are not shrunk quite as much toward 0.



**Figure:** Plot of lasso model parameter estimates for predicting opinion on legalized abortion using student survey data, as function of  $\log(\lambda)$  smoothing parameter. Above the plot are the number of non-zero coefficients for various  $\log(\lambda)$  values.

# Penalty Functions

The smoothing function is also known as *penalty function*.

Popular penalty functions are

➤  **$L_1$  Penalty:**  $s(\beta) = \lambda \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$

↪ Lasso

➤  **$L_2$  Penalty:**  $s(\beta) = \lambda \|\beta\|_2 = \sum_{j=1}^p \beta_j^2$

↪ Ridge regression

👉 Lasso does variable selection and shrinkage, while ridge only shrinkage.

## ➤ II.7.10 Lasso/Ridge Normal Linear Regression

➤ Lasso Estimator  $\hat{\beta}_\lambda^L$ : Minimizes

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\hat{\beta}_j|.$$

➤ Ridge Estimator  $\hat{\beta}_\lambda^R$ : Minimizes

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \hat{\beta}_j^2.$$

# Infinite Estimates in Logistic Regression

## > II.7.11 Firth Method

One of the positive features of penalized likelihood methods is the derivation of sensible estimates when the ML estimate may not exist. They are especially effective for logistic regression models in case of infinite estimates (s. Remark II.3.30). The ML estimator in logistic regression is biased away from 0, and a penalized likelihood correction<sup>a</sup> reduces the bias.

---

<sup>a</sup>Known as the Firth method, since it was proposed by David Firth (Biometrika, 1993).

## II.7.12 Example (Endometrial Cancer)

A study about endometrial cancer with 79 patients analyzed how a histology grade response variable ( $HG = 0$ , low;  $HG = 1$ , high) relates to three risk factors:  $NV$  = neovasculation (1 = present, 0 = absent),  $PI$  = pulsatility index of arteria uterina (ranging from 0 to 49), and  $EH$  = endometrium height (ranging from 0.27 to 3.61).

```
> Endo <- read.table("Endometrial.dat", header=TRUE)
> Endo
NV PI EH HG # HG is histology grade binary response variable
1 0 13 1.64 0
...
79 0 33 0.85 1
> xtabs(~ NV + HG, data=Endo) # contingency table for NV and HG
      HG
NV     0     1
0    49    17    # quasi-complete separation:
1      0    13    # when NV=1, no HG=0 cases occur (MLE of beta_1: infinite)
> fit.pen <- logistf(HG ~ NV + PI + EH, family=binomial, data=Endo); summary(fit.pen)
```

```

> library(logistf) # implements Firth's penalized likelihood method
> fit.pen <- logistf(HG ~ NV + PI + EH, family=binomial, data=Endo)
> summary(fit.pen)
Confidence intervals and p-values by Profile Likelihood # penalized likelihood

```

	coef	se(coef)	lower 0.95	upper 0.95	Chisq	p
(Intercept)	3.7746	1.4887	1.0825	7.2093	8.20	4.194e-03
NV	2.9293	1.5508	0.6097	7.8546	6.80	9.12e-03
PI	-0.0348	0.0396	-0.1245	0.0405	0.75	3.87e-01
EH	-2.6042	0.7760	-4.3652	-1.2327	17.76	2.51e-05

Compare the penalized-likelihood estimates with the MLEs. The penalized likelihood estimate for  $\beta_1$  of 2.93 and the 95% profile penalized likelihood CI of (0.61, 7.85) shrink the ML estimate  $\hat{\beta}_1 = \infty$  and the ordinary profile likelihood interval of (1.28,  $\infty$ ) considerably toward 0, but still suggest that the NV effect may be very strong.



**THE END**