Chapter 13
Modern Estimation Methods

Background In this chapter we discuss methods to estimate biased regression coefficients, which lead to better predictions than those obtained with traditional methods. These modern estimation methods include uniform shrinkage methods (heuristic or bootstrap based) and penalized maximum likelihood methods (with various forms of penalty, including the "Lasso"). We illustrate the application of these methods with a data set of 785 patients from the GUSTO-I trial. It appears that rather advanced procedures can now readily be performed with modern software.

13.1 Predictions from Regression and Other Models

In linear regression, we aim to minimize the mean squared error, which is calculated as the square distance between observed outcome Y and prediction $\hat{Y}$. The prediction $\hat{Y}$ can be based on a single predictor, e.g. age predicts blood pressure, or a multivariable combination of predictors, e.g. age, sex, smoking, and salt intake are used to predict blood pressure. As discussed in previous chapters, we can improve predictions from multivariable models for future subjects if the predictions are shrunk towards the average. Statistically speaking, we can reduce the mean squared error for future subjects by using slightly biased regression coefficients.[81,459] This is because predictions will be slightly biased, but have lower variance. The challenge is to find the optimal balance between increasing bias and decreasing variance. This "bias–variance" trade-off underlies the problem of overfitting, and is essential in all predictive modelling (Chap. 5).

In generalized linear regression models, such as logistic or Cox models, maximum likelihood methods are the classical methods for estimation of regression coefficients. Similar to linear regression, the estimated coefficients can be considered as optimal for the sample under study. But again, introducing some bias in the coefficients may lead to better predictions for future subjects.

Neural networks are examples of generalized non-linear models. Their estimation can be done with various techniques. One popular estimation technique is minimizing the Kullback–Leibler divergence, which can be considered as a

E.W. Steyerberg, *Clinical Prediction Models*,
DOI:10.1007/978-0-387-77244-8_13,

distance between two probability densities. One density is provided by the observed outcomes, another by the estimates from the model. Minimizing the Kullback–Leibler divergence is similar to maximizing the likelihood in a generalized linear regression model. Neural networks are quite flexible, and will hence be severely overfitted when they are fully optimized to fit the data. Therefore a common procedure is "early stopping": the model is not fully trained for maximum fit to the data, but training is stopped at the point where predictive ability is expected to be best. Commonly, the optimal number of iterations to train the model is determined from a cross-validation procedure, where the model is trained on part of the data and tested on an independent part. The optimal number of iterations is then used in the full training part to develop the neural network.

13.2 Shrinkage

Shrinkage of regression coefficients towards zero is one way to improve predictions from a regression model.[81,459] We label this method *shrinkage after estimation,* since the shrinkage is applied to regression coefficients after the model has been fitted initially with traditional methods.

Penalized estimation is an alternative method, which uses a penalty factor in the estimation of the regression coefficients: Larger values of regression coefficients are penalized in the fitting procedure, leading to smaller values being preferred. We refer to this method as *shrinkage during estimation.* Although one single penalty factor is used, the degree of shrinkage varies by predictor. A variant of penalized estimation is the Lasso ("least absolute shrinkage and selection operator").[434] This approach penalizes the sum of the absolute values of the regression coefficients. This leads to some coefficients becoming zero. A predictor with a coefficient of zero can be excluded from the model, which means that the Lasso implies *shrinkage for selection* (Table 13.1).

Table 13.1 Characteristics of three shrinkage methods

Name	Label	Characteristics
Uniform shrinkage	Shrinkage after estimation	Application of a shrinkage factor to the regression coefficients. The shrinkage factor is determined with a heuristic formula, or by bootstrapping
Penalized maximum likelihood	Shrinkage during estimation	Regression coefficients are estimated with penalized maximum likelihood. The optimal penalty factor can be determined by AIC
Lasso	Shrinkage for selection	Regression coefficients are estimated with penalized maximum likelihood with a restriction on the sum of the coefficients ("Lasso"). The optimal penalty factor can be determined by a cross-validation procedure, or AIC

13.2.1 Uniform Shrinkage

A simple and straightforward approach is to apply a uniform (or *linear*) shrinkage factor for the regression coefficients. Shrunk regression coefficients are calculated as $s \times \beta_i$, where s is a uniform shrinkage factor, and β_i are the estimated regression coefficients. The shrinkage factor s may be based on a heuristic formula[81,459]:

$$s = (\text{model } \chi^2 - df)/\text{model } \chi^2,$$

where model χ^2 is the likelihood ratio χ^2 of the fitted model (i.e., the difference in -2log likelihood between the model with and without predictors), and *df* indicates the degrees of freedom of the number of candidate predictors considered for the model. The required shrinkage increases when larger numbers of predictors are considered (more *df*), or when the sample size is smaller (smaller model χ^2).

We can also calculate the uniform shrinkage factor s with bootstrapping.[459,174]

1. Take a random bootstrap sample of the same size as the original sample, drawn with replacement.
2. Select the predictors according to the selection procedure and estimate the logistic regression coefficients in the bootstrap sample.
3. Calculate the value of the linear predictor for each patient in the original sample. The linear predictor is the linear combination of the regression coefficients as estimated in the bootstrap sample with the values of the predictors in the original sample.
4. Estimate the slope of the linear predictor, using the outcomes of the patients in the original sample.

Steps 1–4 need to be repeated many times to obtain a stable estimate of the shrinkage factor as the mean of the slopes in step 4. For example, we may use 200 bootstrap samples, although a fully stable estimate may require 500 bootstrap repetitions.[401] The shrinkage factor may take values between 0 and 1.

*13.2.2 Uniform Shrinkage in GUSTO-1

As an example, we consider sample4 from the GUSTO-I study of patients with an acute myocardial infarction (see Chap. 22). The data set consists of 785 patients, of whom 52 had died by 30 days. We consider 2 models for prediction of 30-day mortality after an acute MI: an 8 predictor model, and a 17 predictor model. For estimation of the heuristic shrinkage factor, we need the model χ^2 of each model. These were 62.6 and 73.5. The heuristic shrinkage estimate s was hence $(62.6 - 8) / 62.6 = 0.87$. The larger model required more shrinkage, with $s = (73.5 - 17)/73.5 = 0.77$.

Next, a bootstrap procedure was performance with 200 replications. This resulted in identical estimates of the slope of the linear predictor (0.87 and 0.77, respectively). The regression coefficients are shown in Table 13.2.

Table 13.2 Logistic regression coefficients estimated with standard maximum likelihood ("original"), uniform shrinkage, penalized maximum likelihood, and the Lasso, for sample4 (795 patients with acute MI, 52 deaths by 30 days)

Predictor	Original	Shrunk	Penalized	Lasso
SHO	1.12	0.97	1.17	1.09
A65	1.49	1.30	1.21	1.36
HIG	0.84	0.74	0.72	0.73
DIA	0.43	0.38	0.36	0.35
HYP	0.99	0.86	0.83	0.87
HRT	0.96	0.84	0.84	0.87
TTR	0.59	0.51	0.49	0.46
SEX	0.07	0.06	0.11	0.00
Shrinkage parameter	NA	s=0.87	penalty=8	s=0.88
Effective shrinkage	1	0.87	0.81–1.49	0–0.97

13.3 Penalized Estimation

Penalized maximum likelihood estimation is a generalization of the ridge regression method, which can be used to obtain more stable parameters for linear regression models.[102] Instead of maximizing the log likelihood in generalized linear models, a penalized version of the log likelihood is maximized, in which a penalty factor λ is used:

$$\text{PML} = \log L - 0.5\ \lambda\ \Sigma\ (s_i\ \beta_i)^2,$$

where PML is penalized maximum likelihood, L is the maximum likelihood of the fitted model, λ a penalty factor, β the estimated regression coefficient for each predictor i in the model, and s_i is a scaling factor for each β_i to make $s_i\ \beta_i$ unitless.[174,468] It is convenient to use the standard deviation of each predictor for the scaling factor s_i.[174]

13.3.1 Penalized Maximum Likelihood Estimation

The PML can also be formulated as $\text{PML} = \log L - 0.5\ \lambda\ \beta' P\ \beta$, where λ is a penalty factor, β' denotes the transpose of the vector of estimated regression coefficients (excluding the intercept), and P is a non-negative, symmetric penalty matrix. For penalized estimation, the diagonal of P consists of the variances of the predictors and all other values of P are set to 0.[174] If P is defined as $\text{cov}(\beta)^{-1}$ (i.e., the inverse of the covariance matrix of the regression coefficients β), shrinkage of the regression coefficients is achieved, which is identical to the use of a uniform shrinkage factor as determined by leave-one out cross-validation.[468] If P is equal to the matrix of second derivatives of the likelihood function, PML is similar to applying a uniform shrinkage factor $s = 1/(1 + \lambda)$.

The main problem in penalized estimation is how to choose the optimal penalty factor λ_{opt}. Maximizing a modified Akaike's Information Criterion (AIC) is an

efficient method.[149] Traditionally, the AIC is defined as $-2 \log L + 2p$, where L is the maximum likelihood of the fitted model and p is the degrees of freedom equal to the number of fitted predictors. A more convenient formulation is as

$$\text{AIC}_{\text{model}} = \text{model } \chi^2 - 2p,$$

where model χ^2 is the likelihood ratio χ^2 of the fitted model (i.e., the difference in -2 log likelihood between the model with and without predictors). For penalized maximum likelihood estimation we use a modified AIC, defined as

$$\text{AIC}_{\text{penalized}} = \text{model } \chi^2_{\text{penalized}} - 2\, df_{\text{effective}},$$

where model $\chi^2_{\text{penalized}}$ refers to likelihood ratio χ^2 of the penalized model, and $df_{\text{effective}}$ is the degrees of freedom after penalizing the fitted predictors. In standard logistic regression, the degrees of freedom are equal to the number of predictors in the model; the higher the number of predictors, the higher the degrees of freedom and the more likely the model is overfitted. Because of the penalization, the degrees of freedom effectively used in penalized estimation are lower than the actual number of predictors. More technically, $df_{\text{effective}}$ is derived from the reduction in variance of penalized parameter estimates in comparison to the variance of ordinary parameter estimates[149]:

$$df_{\text{effective}} = \text{trace } [I(\beta)\, \text{cov}(\beta)],$$

where $I(\beta)$ is the information matrix as computed without the penalty function, and $\text{cov}(\beta)$ is the covariance matrix as computed by inverting the information matrix calculated with the penalty function. If both the $I(\beta)$ and $\text{cov}(\beta)$ are estimated without penalty, $I(\beta)\, \text{cov}(\beta)$ is the identity matrix and trace $[I(\beta)\text{cov}(\beta)]$ is equal to the number of estimated parameters in the model (excluding the intercept). With a positive penalty function, the $\text{cov}(\beta)$ becomes smaller and the effective degrees of freedom decrease. With higher penalty values, the model $\chi^2_{\text{penalized}}$ decreases (poorer fit to the data), but so does the $df_{\text{effective}}$. The maximum of $\text{AIC}_{\text{penalized}}$ (model $\chi^2_{\text{penalized}} - 2\, df_{\text{effective}}$) is sought by varying the values of λ in a trial and error process. For example, we may vary λ over a grid such as 0, 1, 2, 4, 6, 8, 12, 16, 24, 32, 48. Larger values of λ are required for more complex models and larger data sets. The optimal penalty factor λ_{opt} is the value of λ that maximizes $\text{AIC}_{\text{penalized}}$. With this optimal λ, the final model is estimated. An alternative is to use cross-validation or bootstrapping to find the optimal λ, which is more computer intensive compared with finding the maximum of $\text{AIC}_{\text{penalized}}$.

*13.3.2 Penalized ML in Sample4

We searched for a penalty factor λ over a grid using the `pentrace` function. The fitting for the 8 predictor model is as follows:

```
# logistic regression model with 8 predictors
full8 <- lrm(DAY30~SHO+A65+HIG+DIA+HYP+HRT+TTR+SEX, data=gustos)
# determine performance over range of penalties
p8 <- pentrace(full8, 0:20)
# fit penalized model
full8.pen <- update(full8, penalty=p8$penalty)
```

The $AIC_{penalized}$ is calculated with the effective degrees of freedom, and is plotted in Fig. 13.1. The optimum penalty factors were 8 for the 8 predictor model, and 24 for the 17 predictor model. The effective degrees of freedom were 6.9 (instead of 8) and 10.8. (instead of 17). Note that the $AIC_{penalized}$ was worse for the 17 predictor model compared with the 8 predictor model, over all penalties considered. The 17 predictor model was hence actually overfitted with only 52 events in the data set.

For comparison we also performed a bootstrap procedure to find the optimal penalty factor λ. We created logistic regression models with a range of penalty factors in bootstrap samples drawn with replacement. The models were tested in the original sample. A linear predictor was calculated with the penalized coefficients from the bootstrap sample and the predictor values in the original sample: $lp = X_{original}$ %×% $coef_{penalized, bootstrap}$. Various performance measures can be calculated for this linear predictor. We focus on the slope of the linear predictor, since the primary objective of shrinkage methods is to improve calibration. As expected, the slope is below 1 when no shrinkage is applied (Fig. 13.2). It appears that the slope is 1 if we apply a penalty factor of 7 for the 8 predictor model, and 12 for the 17 predictor model. These values are slightly lower than those obtained from

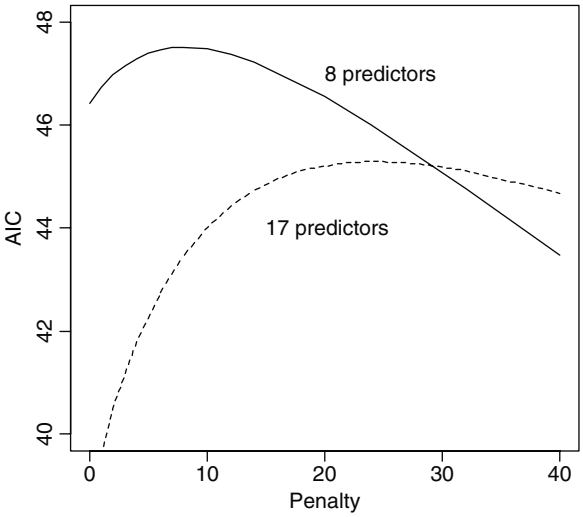

Fig. 13.1 $AIC_{penalized}$ in relation to the penalty factor. Optimum values are 8 and 24 for the 8 and 17 predictor models, respectively

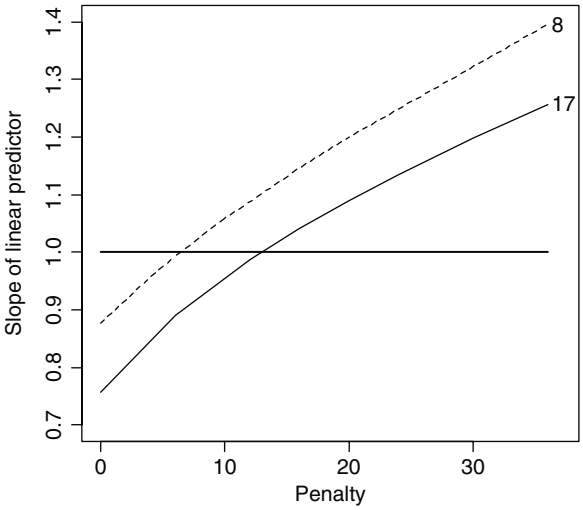

Fig. 13.2 Slope of the linear predictor in relation to the penalty factor according to a bootstrap procedure. Optimum values are 7 and 12 for the 8 and 17 predictor models, respectively

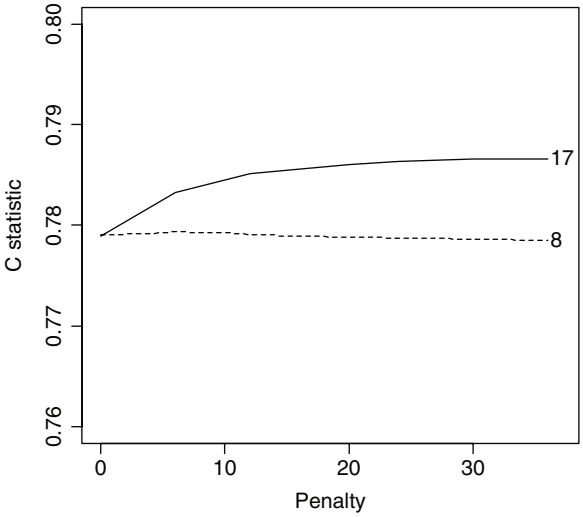

Fig. 13.3 *c* statistic in relation to the penalty factor according to a bootstrap procedure. Optimum values are 8 and 30 for the 8 and 17 predictor models, respectively

maximizing the $AIC_{penalized}$. This is explained by the fact that AIC considers the model χ^2 as criterion rather than the slope of the linear predictor. The model χ^2 also reflects the discriminative ability, which was higher with larger penalty values (Fig. 13.3).

13.3.3 Shrinkage, Penalization, and Model Selection

Uniform shrinkage and penalized estimation methods are defined for pre-specified models. If we apply a selection strategy such as stepwise selection, fewer predictors are included in the selected model, and we might expect less need for shrinkage of coefficients. However, we know that a "testimation" problem arises, i.e. coefficients of selected predictors are overestimated. This selection bias should be taken into account when calculating a shrinkage factor. This may be achieved by considering the number of candidate predictors in the heuristic formula (instead of the number of selected predictors).[459] In a bootstrap procedure, we can include the selection process in step 2.[174] Empirical research suggests that the required shrinkage is more or less similar in pre-specified or selected models.[409] For penalized estimates of the regression coefficients after selection, we can apply the penalty factor that was identified as optimal for the full model, before selection took place.

A specific situation is when a substantial number of interaction terms is tested, and one or more are included in the final model. For shrinkage, we could still use the original *df* of the model with main effects and all interactions considered. A more elegant solution was suggested by Harrell for penalized ML estimation, i.e. to penalize the interaction terms more than the main effects, for example with twice the penalty of the main effects. Similarly, non-linear and nonlinear interaction terms might be penalized by twice and 4 times the penalty of the main effects.[174]

13.4 Lasso

A formal method to achieve model selection through shrinkage is the Lasso (least absolute shrinkage and selection operator).[434] The Lasso can efficiently be applied to linear regression models using "least angle regression."[106] The Lasso can also be used for generalized linear models such as the logistic or Cox model.[435] The Lasso preferentially shrinks some predictors to zero.

13.4.1 Estimation of Lasso Model

The Lasso estimates the regression coefficients of standardized predictors by minimizing the log-likelihood subject to $\Sigma\,|\beta| <= t$, where t determines the shrinkage in the model. We may vary $s = t / |\Sigma\beta_0|$ over a grid between 0 and 1, where β_0 indicates the standard ML regression coefficients and s may be interpreted as a standardized shrinkage factor. We may estimate the final β with the value of t that gives the lowest mean-squared error in a generalized cross-validation procedure.[435] We may also aim to optimize AIC or use bootstrapping to find the optimal value for t.[320]

*13.4.2 Lasso in GUSTO-I

We used the `glmpath` package for *R* to perform lasso analyses, but other packages are nowadays available. This is a path-following algorithm for L1 regularized generalized linear models and Cox proportional hazards model.[320] The logistic regression coefficients were estimated given a bound ("*L*1") to the sum of absolute β, $|\beta|$. The predictors are standardized such that sum $|\beta|$ does not depend on coding of predictors.

```
# make list of predictors in matrix x, outcome in y
gustosd <- list(x=full8$x, y=full8$y)
# fit logistic models over a range of L1
gustopath <- glmpath(data=gustosd)
# plot results: Fig 13.4
plot.glmpath(gustopath, type="coefficients")
plot.glmpath(gustopath, type="aic")
```

With a low *L*1 bound, small coefficients were estimated for the predictors A65 (age>65 years), SHO (Shock), and HRT (Tachycardia). This occurred both in the 8 and 17 predictor models (Fig. 13.4). The other predictors had coefficients set to zero. With larger bounds, non-zero coefficients were estimated for these predictors as well. With a bound over 0.6 (8 predictor model) or over 0.9 (17 predictor model), the original, unshrunk logistic model was estimated.

The optimum penalty can be estimated by studying the AIC (Fig. 13.4). This suggests an optimal selection of seven predictors in the 8 predictor model (*L*1 = 2.1), and a selection of 14 predictors for the 17 predictor model (*L*1 = 3.0). We can validate the selection and estimated coefficients through a bootstrap analysis (see the book's website). The coefficients for the final model are chosen at the lowest AIC value. The effect of SEX was set to zero, and the coefficient of DIA was small (standardized coefficient, 0.10).

```
# coefficients at lowest AIC: Table 13.2
gustopath$b.predictor[gustopath$aic==min(gustopath$aic),]
```

```
Intercept   SHO   A65   HIG   DIA   HYP   HRT   TTR    SEX
-4.55       1.09  1.36  0.73  0.35  0.87  0.87  0.46   0.00
```

```
# linear predictor with Lasso model, step 12 has lowest AIC
predict.glmpath(gustopath, newx=full8$x, newy=full8$y, s=12)
```

13.4.3 Predictions after Shrinkage

Shrinkage leads to a less-extreme distribution of predictions in the GUSTO-I example. The linear predictor is shrunk towards the average compared with standard maximum likelihood, either with uniform shrinkage, penalized maximum likelihood estimation (PMLE), or the Lasso (Fig. 13.5).

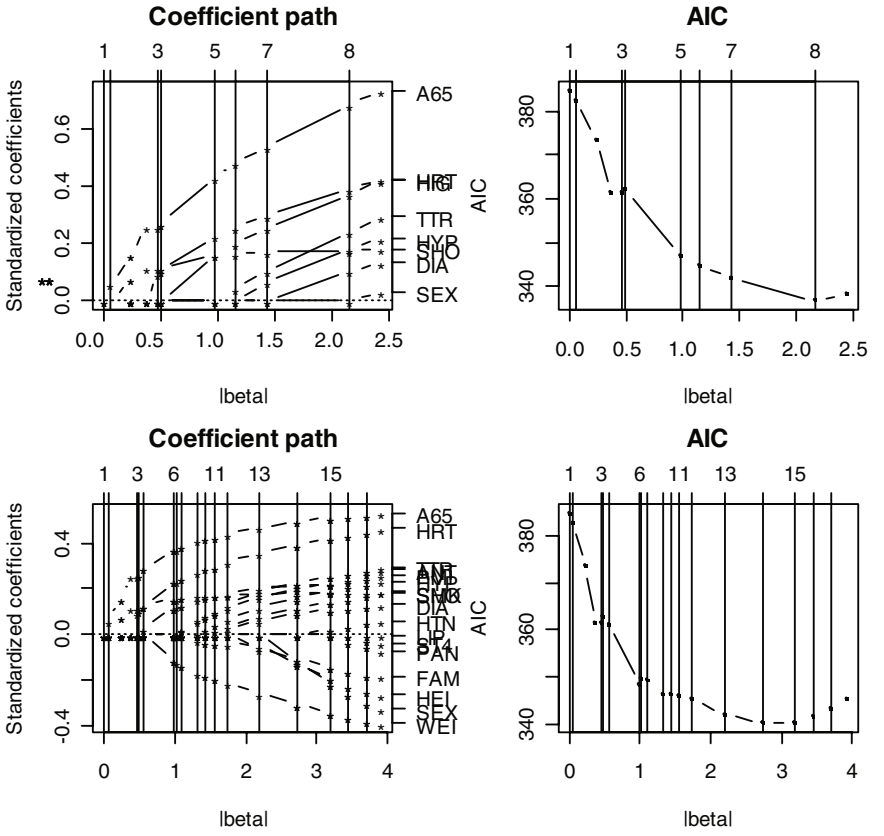

Fig. 13.4 Coefficients and AIC for 8 and 17 predictor models according to the sum of the absolute values of the regression coefficients (|β|) in sample4 from GUSTO-I (n=785, 52 deaths)

13.4.4 Model Performance after Shrinkage

We compared the performance of models constructed in small samples of the GUSTO-I data set in an independent test part (see Chap. 22 for design). Table 13.3 shows the discrimination and calibration with and without shrinkage. As expected, discrimination is not much affected by shrinkage. In contrast, the calibration slope is closer to 1 when shrinkage is applied. Shrinkage hence prevents that too extreme predictions are derived from the development data set.

13.5 Concluding Remarks

Shrinkage of regression coefficients is an important way to battle overfitting; too extreme predictions are prevented. Shrinkage is especially beneficial in small data sets, and/or situations with large numbers of candidate predictors. Using advanced

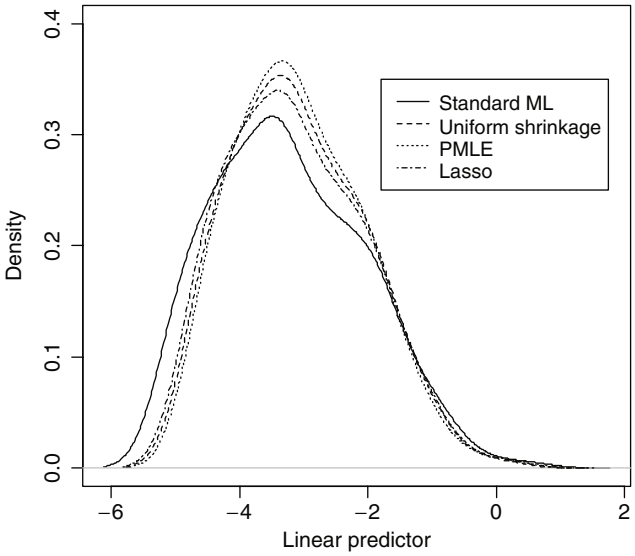

Fig. 13.5 Distribution of the linear predictor in sample4 from GUSTO-I with standard and penalized maximum likelihood, uniform shrinkage, and the Lasso

Table 13.3 Discrimination (*c* statistic) and calibration (calibration slope) of the 8 and 17 predictor models based on small and large subsamples (average, *n*=336 and *n*=892, respectively), and based on the total training part (*n*=20,512), as evaluated in the independent test part of GUSTO-I (*n*=20,318)

		8 predictors		17 predictors	
Training data		C statistic	Slope	C statistic	Slope
Total training (*n*=20,512, 1,423 deaths)	Standard ML	0.789	0.944	0.802	0.959
61 small subsamples (*n*=336, 23 deaths on average)	Standard ML	0.75	0.66		
	Uniform shrinkage	0.75	1.01		
	Penalized ML	0.76	0.93		
	Lasso	0.75	0.83		
23 large subsamples (*n*=892, 62 deaths on average)	Standard ML	0.78	0.86	0.78	0.76
	Uniform shrinkage	0.78	0.97	0.78	0.95
	Penalized ML	0.78	0.96	0.79	0.98
	Lasso	0.78	1.01	0.78	0.93

Mean values are shown for several estimation methods with a fixed selection of predictors

shrinkage procedures is readily possible with modern sofware, implemented in for example *R* (`pentrace` function in `Design` library for penalized estimation, `glmpath` for Lasso). Penalty factors are a general concept in smooth estimation of model parameters; they are also important in curve fitting (e.g. with splines) and generalized additive models.[181] The Lasso currently receives interest for analysis of genomic data.[321]

Shrinkage methods have been applied in a number of case studies. Moons et al. describe penalized maximum likelihood and illustrates the method with a nice case study.[293] Vach et al. compared the empirical behaviour of various shrinkage techniques.[443] The results from simulations in GUSTO-I were presented in more detail in other papers.[408,409,410]

Questions

13.1 Shrinkage and model performance
Explain how shrinkage can influence (a) the predictions from a model, (b) calibration, and (c) discrimination.

13.2 Penalized maximum likelihood
(a) Why might we label PML "shrinkage during estimation" (Table 13.1)
(b) How is it possible that one penalty term leads to differential shrinkage in Table 13.2?
(c) In a recent paper (Smits et al. 2007),[391] we can study the effect of PML on the various coefficients. Which coefficients are penalized most?

13.3 Shrinkage methods and stepwise selection (Sect. 13.3.3)
How can shrinkage and penalization be used when the model is developed with stepwise selection:
(a) Uniform shrinkage with Van Houwelingen's formula or bootstrapping?
(b) Penalized maximum likelihood?