Shrinkage methods and variable selection: Ridge, Lasso, and Elastic Nets

Federico Andreis
Department of Policy Analysis and Public Management
Università Commerciale Luigi Bocconi
federico.andreis@unibocconi.it

McKinsey, Milano
3 November 2017

Course structure

What is this course about?

Today we introduce a few statistical techniques that can be used to solve some commonly encountered problems in data analysis and predictive tasks.

We will:

- take a look at the theoretical framework for shrinkage methods and principal component analysis
- discuss advantages and limitations of commonly used techniques and their generalizations
- present some real-life applications
- learn how to implement these methods in `R`.

Goals for the day

After this course the participants should be able to

- identify the situations where applying regularisation and/or dimensionality reduction methods can improve the quality of the analyses
- devise and implement shrinkage, variable selection, and dimensionality reduction strategies
- correctly interpret the results.

The crew

Course instructor

Federico Andreis, PhD (Bocconi University)

Teaching assistant

Isadora Antoniano-Villalobos, PhD (Bocconi University)

Local organiser

Alberto Oltolini, MSc (McKinsey Milano)

Agenda

The day is structured as follows:

- Shrinkage and variable selection
  - 8.30-10.30 theory [**FA**]
  - **break!**
  - 10.45-12.30 hands-on [**FA**,**IAV**]
- Dimensionality reduction using PCA
  - 13.30-15.15 theory [**FA**]
  - **break!**
  - 15.30-17.15 hands-on [**FA**,**IAV**]
  - **break!**
- Wrap-up
  - 17.30-18.00 brainstorming session [**AO**,**FA**]

Course material

The main reference is The Elements of Statistical Learning, by Hastie, Tibshirani, and Friedman.

The course material includes:

- lecture notes + hands-on, morning session
- lecture notes + hands-on, afternoon session
- `R` workspaces containing the datasets for the practicals
- you! Feel free to raise your hand...

Additional references can be found in the *Bibliography* sections.

Introduction to shrinkage methods

The linear model

The classic linear regression model, expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon,$$

is used to describe the relationship between a response Y and a set of variables $X_1, ..., X_p$; ϵ denotes the non-deterministic component.

The linear model

The classic linear regression model, expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon,$$

is used to describe the relationship between a response Y and a set of variables $X_1, ..., X_p$; ϵ denotes the non-deterministic component.

Notwithstanding its simplicity, the linear model often offers competitive performances in relation to more complex non-linear methods, and has distinctive advantages in terms of inference.

The parameters $\beta_0, ..., \beta_p$ are typically estimated from a sample of n observations using the **ordinary least squares** (OLS) criterion.

Accuracy and interpretability

Prediction accuracy: if the true relationship between Y and $X_1, ..., X_p$ is approximately linear, the OLS estimates will have low bias; if $n \gg p$, they will also tend to have low variance.

Accuracy and interpretability

Prediction accuracy: if the true relationship between Y and $X_1, ..., X_p$ is approximately linear, the OLS estimates will have low bias; if $n \gg p$, they will also tend to have low variance.

- If n is not much larger than p, however, there can be high variability in the estimates $\rightarrow$ overfitting + poor predictive power. Also, if $p > n$, the OLS solution is not even unique.

Accuracy and interpretability

Prediction accuracy: if the true relationship between Y and $X_1, ..., X_p$ is approximately linear, the OLS estimates will have low bias; if $n \gg p$, they will also tend to have low variance.

- If n is not much larger than p, however, there can be high variability in the estimates $\rightarrow$ overfitting + poor predictive power. Also, if $p > n$, the OLS solution is not even unique.

Model interpretability: the OLS estimates have a clear interpretation. However, if some or many of the p variables are not associated with the response, failure to exclude these irrelevant covariates leads to an unnecessary additional complexity.

Accuracy and interpretability

Prediction accuracy: if the true relationship between Y and $X_1, ..., X_p$ is approximately linear, the OLS estimates will have low bias; if $n \gg p$, they will also tend to have low variance.

- If n is not much larger than p, however, there can be high variability in the estimates $\rightarrow$ overfitting + poor predictive power. Also, if $p > n$, the OLS solution is not even unique.

Model interpretability: the OLS estimates have a clear interpretation. However, if some or many of the p variables are not associated with the response, failure to exclude these irrelevant covariates leads to an unnecessary additional complexity.

- Unfortunately, the OLS approach is unlikely to yield any coefficient estimate that is exactly zero, which would lead to removal of the corresponding variables from the model.

Shrinkage

Shrinkage refers to a class of regularisation methods that involve fitting a regression model using all p predictors, under some constraint on the size of their estimated coefficients.

Among the most important features of this regularisation approach, we highlight that shrinking

- tends to reduce the variability of the estimates, hence improving the model's stability
- can go as far as to set some of the coefficients to zero, thus also allowing for variable selection.

Today we discuss the **ridge**, **lasso**, and **elastic net** approaches.

Shrinkage methods: the Ridge regression

The Ridge regression

The solution to the ordinary least squares fitting procedure is the vector $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_p)$ that minimises the Residual Sum of Squares

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

The Ridge regression

The solution to the ordinary least squares fitting procedure is the vector $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_p)$ that minimises the Residual Sum of Squares

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

Ridge regression, similarly, seeks the vector $\hat{\beta}^{\text{ridge}}$ that minimises the *penalized* or *regularised* RSS

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

where $\lambda \geq 0$ is a **complexity parameter**.

The shrinkage effect

The term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called a **shrinkage penalty** and is small when $\beta_1, ..., \beta_p$ are close to zero, hence it has the effect of shrinking the coefficients towards zero.

The shrinkage effect

The term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called a **shrinkage penalty** and is small when $\beta_1, ..., \beta_p$ are close to zero, hence it has the effect of shrinking the coefficients towards zero.

The tuning parameter λ acts as a regulator of the amount of shrinkage on the regression estimates:

- if $\lambda = 0$, then $\hat{\beta}^{\text{ridge}} \equiv \hat{\beta}$
- as $\lambda \to \infty$, the impact of the penalty grows and $\hat{\beta}^{\text{ridge}} \to \mathbf{0}_p$.

The shrinkage effect

The term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called a **shrinkage penalty** and is small when $\beta_1, ..., \beta_p$ are close to zero, hence it has the effect of shrinking the coefficients towards zero.

The tuning parameter λ acts as a regulator of the amount of shrinkage on the regression estimates:

- if $\lambda = 0$, then $\hat{\beta}^{\text{ridge}} \equiv \hat{\beta}$
- as $\lambda \to \infty$, the impact of the penalty grows and $\hat{\beta}^{\text{ridge}} \to \mathbf{0}_p$.

Note: unlike least squares, $\hat{\beta}^{\text{ridge}}$ is not unique, rather a function of λ. Selecting good values for λ is critical, and is usually done numerically via cross-validation.

Standardization of $X_1, \ldots, X_p$

Since the ridge solutions, unlike the standard OLS estimates, are not equivariant under scaling, it is customary to standardize the inputs before estimation.

Moreover, penalization of the model intercept would make the procedure depend on the origin chosen for the dependent variable, hence β_0 is estimated separately as $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$.

Standardization of $X_1, \ldots, X_p$

Since the ridge solutions, unlike the standard OLS estimates, are not equivariant under scaling, it is customary to standardize the inputs before estimation.

Moreover, penalization of the model intercept would make the procedure depend on the origin chosen for the dependent variable, hence β_0 is estimated separately as $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$.

The remaining parameters are then estimated by a ridge regression without intercept, using the standardized covariates. Let $\mathbf{X}$ denote the $p \times p$ (standardized) data matrix. It is easy to show that the solutions take the form

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix.

Ridge coefficients profiles

Let's take a look at the *profiles* or *paths* of the ridge solutions for an example dataset:

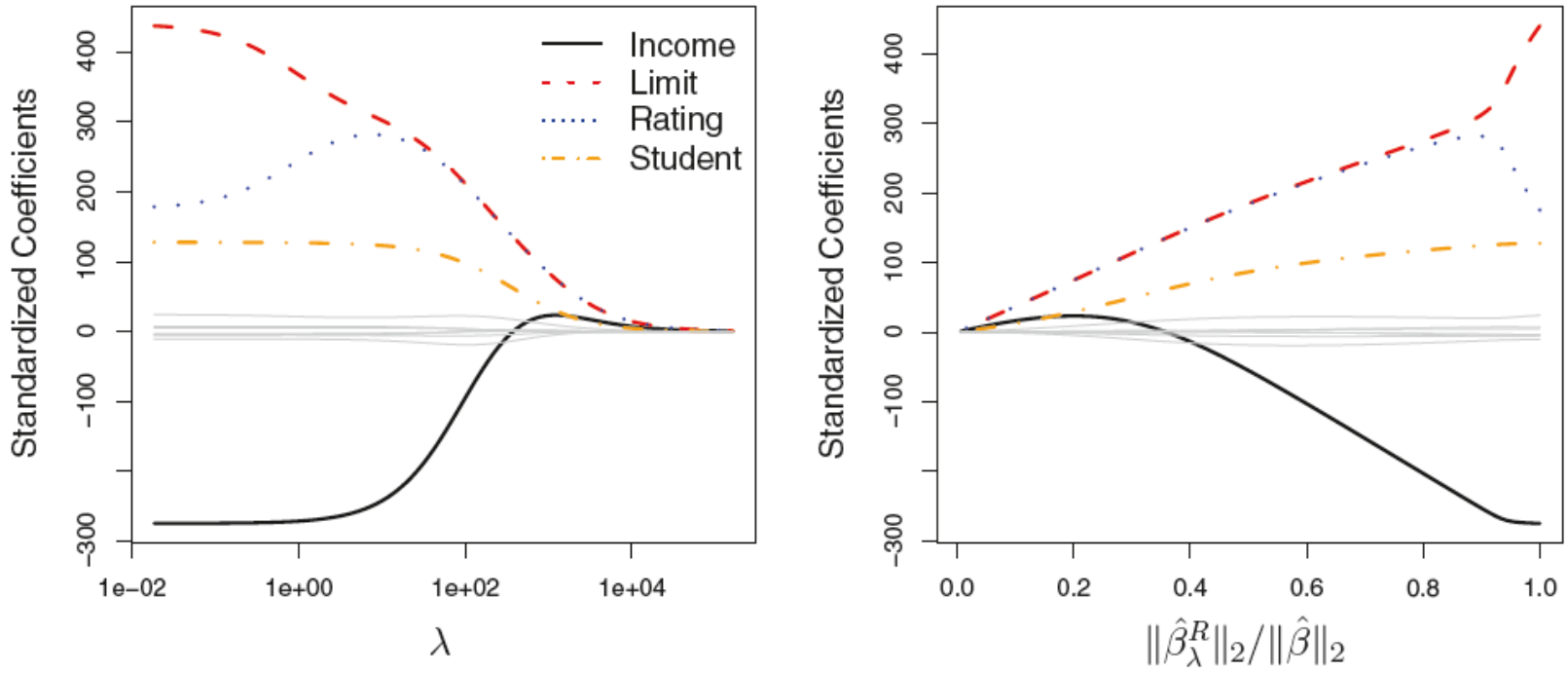


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

Choosing λ by cross-validation

Cross-validation (CV) is a resampling approach used to estimate the test Mean Squared Error (MSE) of a model by repeatedly holding out a subset of the observations, and applying the chosen method to predict the held out outcome.

Choosing λ by cross-validation

Cross-validation (CV) is a resampling approach used to estimate the test Mean Squared Error (MSE) of a model by repeatedly holding out a subset of the observations, and applying the chosen method to predict the held out outcome.

- **Leave-one-out**: take out one observation, fit the model on the remaining $n - 1$. Repeat. Average. *Low bias, high variance*.

Choosing λ by cross-validation

Cross-validation (CV) is a resampling approach used to estimate the test Mean Squared Error (MSE) of a model by repeatedly holding out a subset of the observations, and applying the chosen method to predict the held out outcome.

- **Leave-one-out**: take out one observation, fit the model on the remaining $n - 1$. Repeat. Average. *Low bias, high variance*.
- k-**fold**: split the data in k equal-sized subsets (folds) at random, take out one, fit the model on the remaining $k - 1$. Repeat. Average. *Acceptable bias, low variance*.

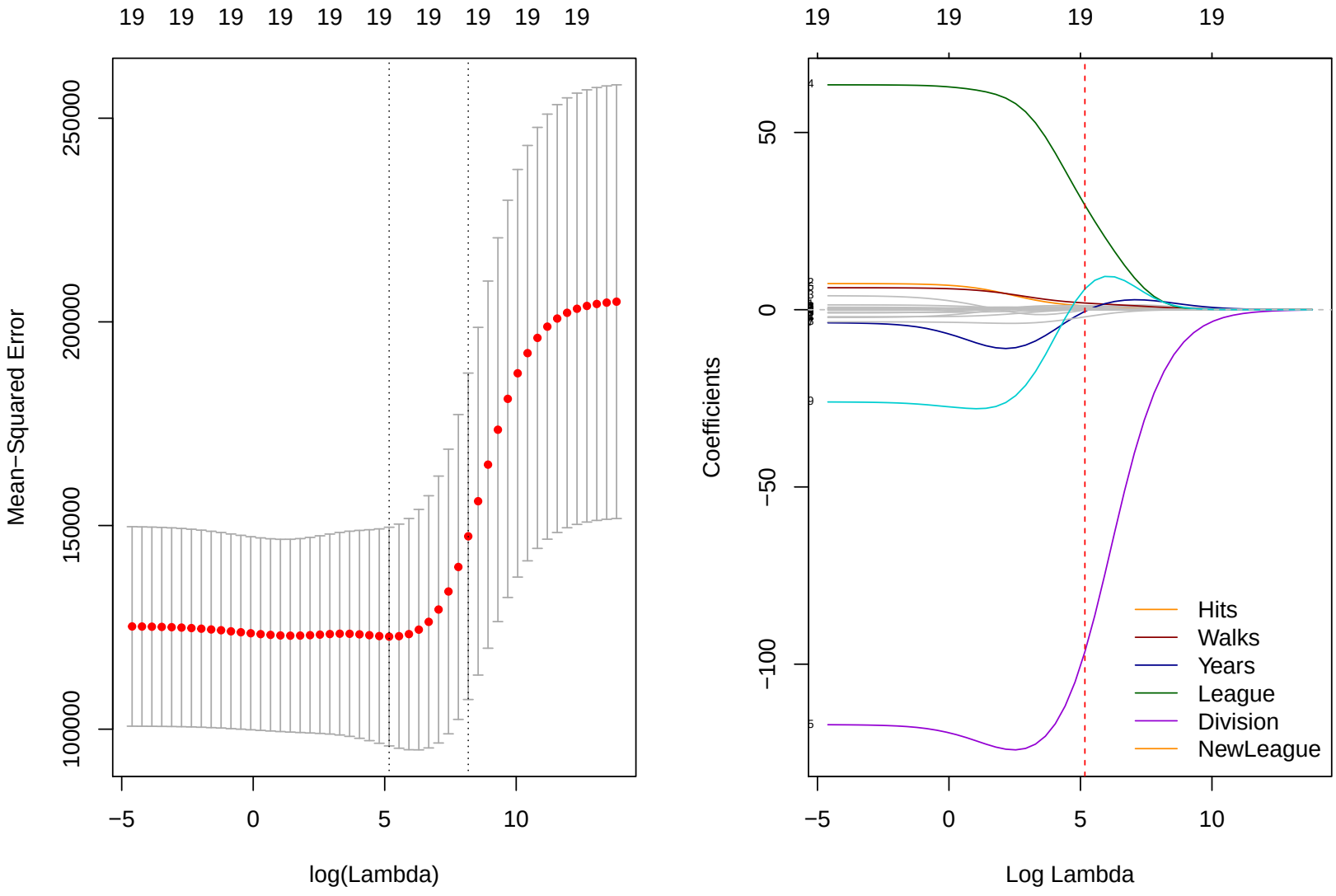
Choosing λ by cross-validation

Cross-validation (CV) is a resampling approach used to estimate the test Mean Squared Error (MSE) of a model by repeatedly holding out a subset of the observations, and applying the chosen method to predict the held out outcome.

- **Leave-one-out**: take out one observation, fit the model on the remaining $n-1$. Repeat. Average. *Low bias, high variance*.
- k-**fold**: split the data in k equal-sized subsets (folds) at random, take out one, fit the model on the remaining $k-1$. Repeat. Average. *Acceptable bias, low variance*.

We shall use k-fold CV to choose the value of λ that minimizes the estimated MSE. Values of k between 5 and 10 are typically indicated as good for computational burden/bias trade-off.

Choosing λ: k-fold cross-validation



The bias-variance trade-off

The key to the improvement of ridge regression over OLS is in the **bias-variance trade-off**: as λ increases, so does the bias, but the variance decreases by virtue of the lower flexibility.

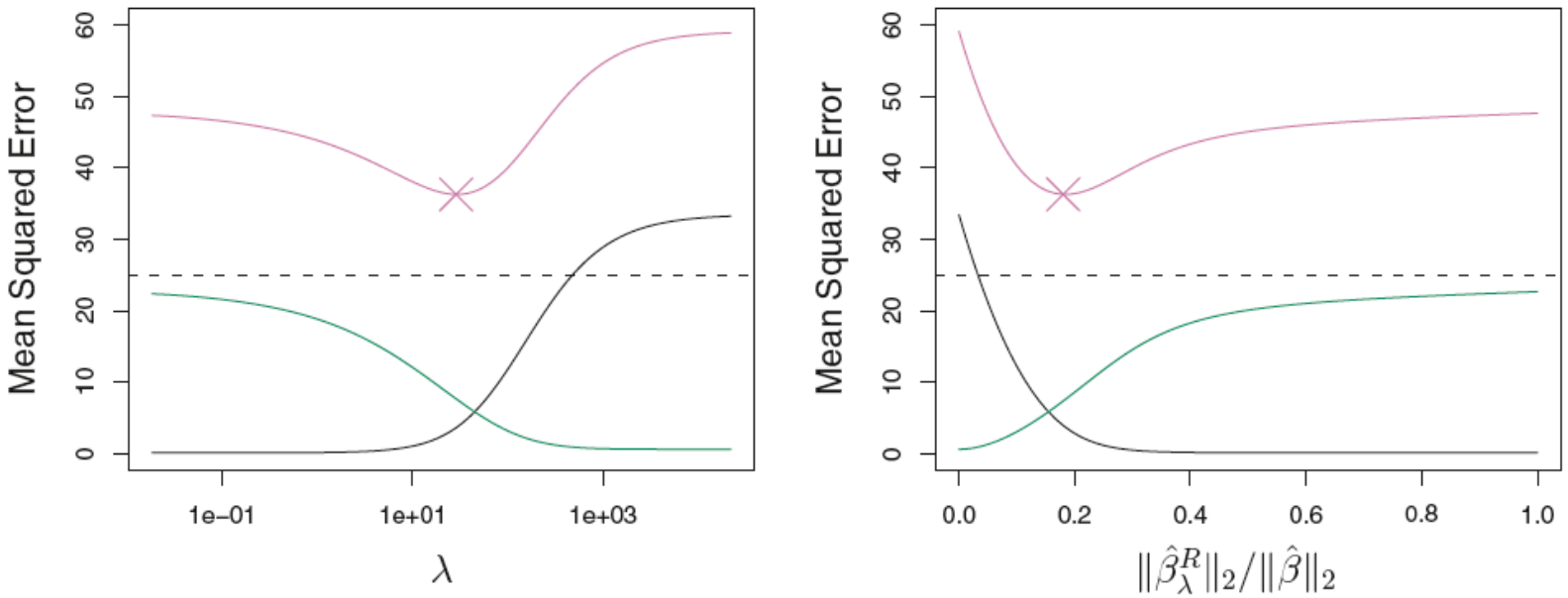
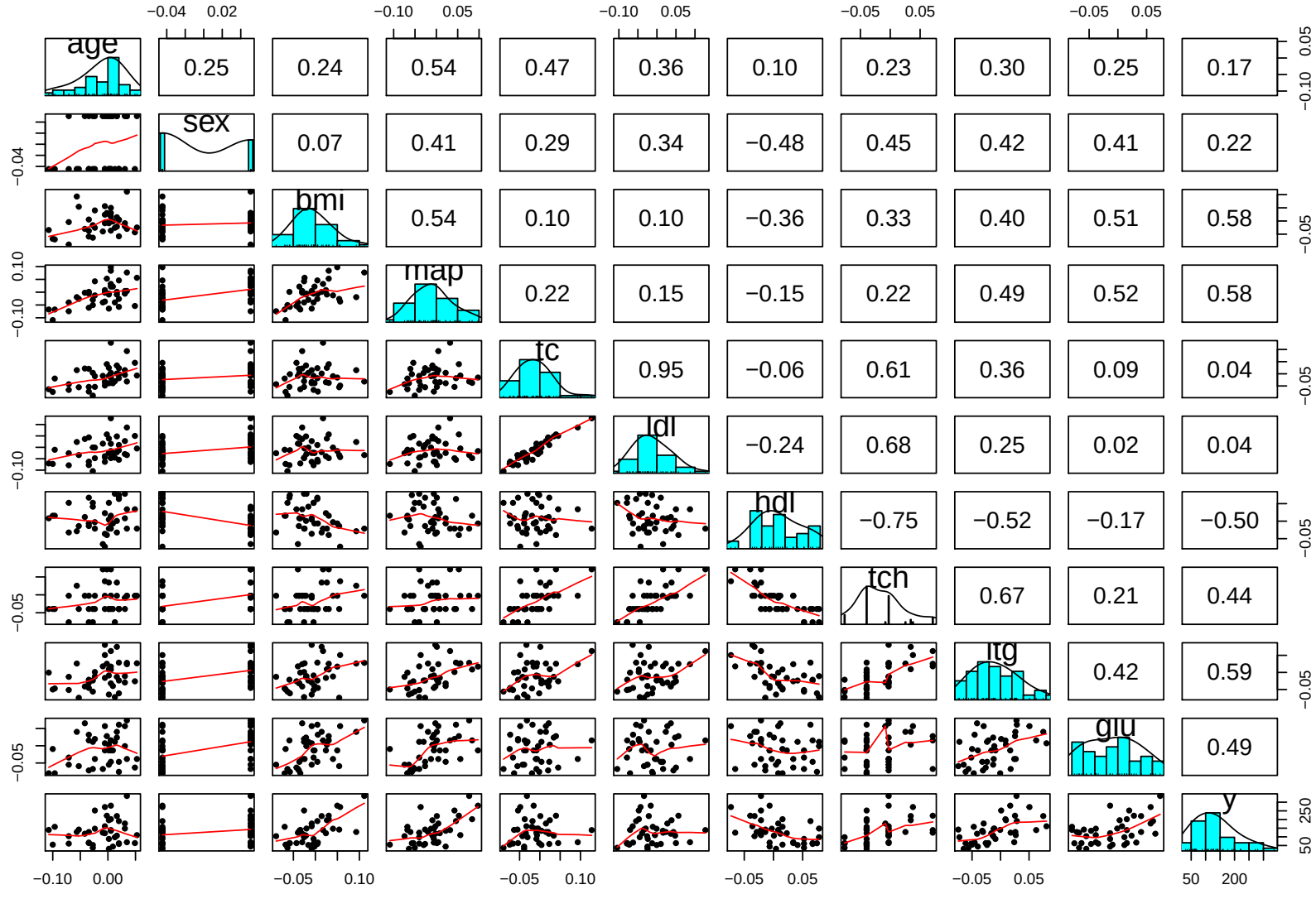
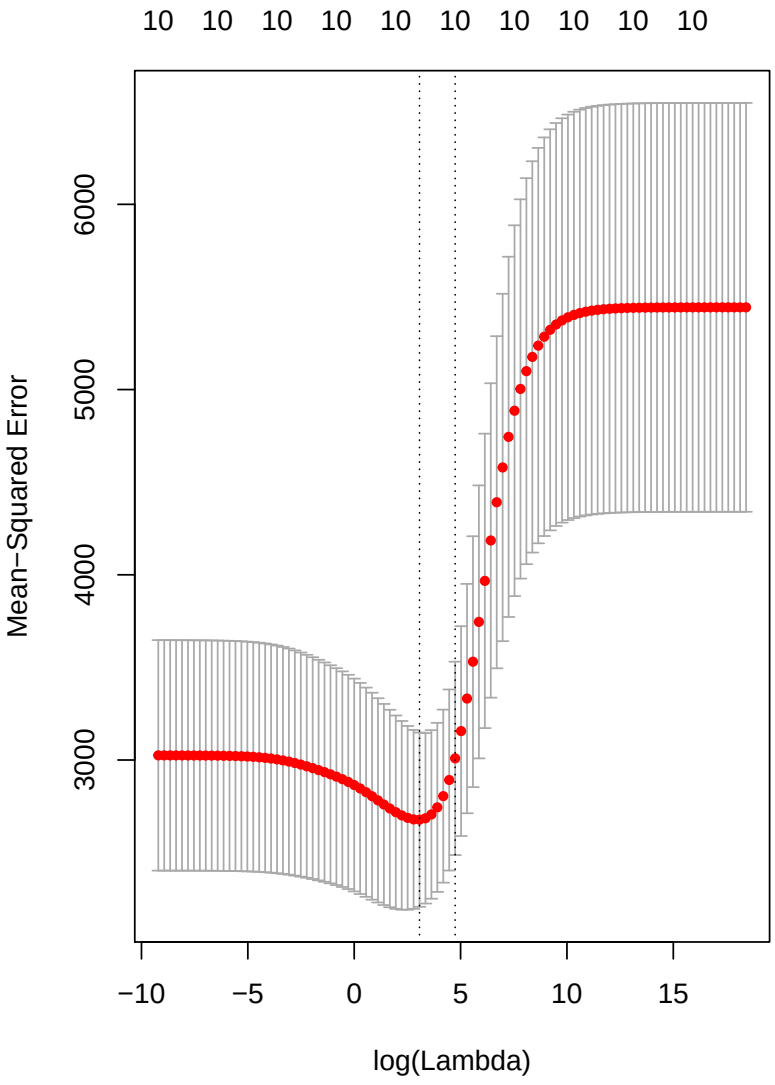


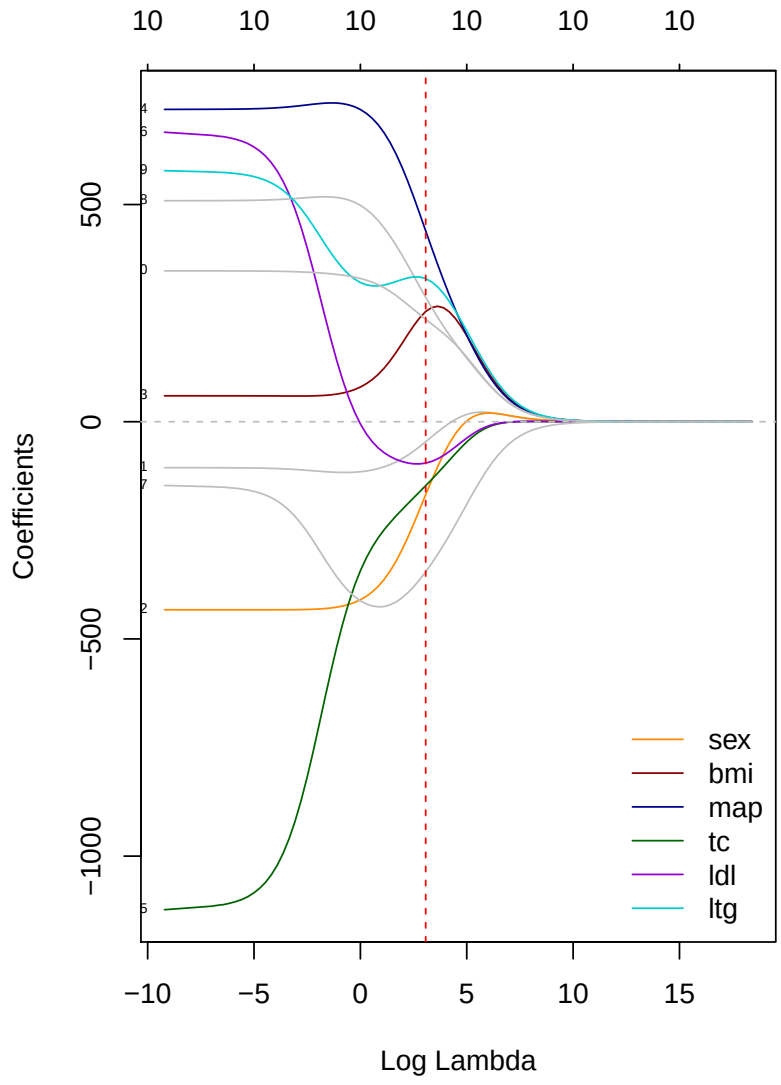
FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$*. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Example: diabetes progression - $n = 40, p = 10$



Example: diabetes progression - ridge



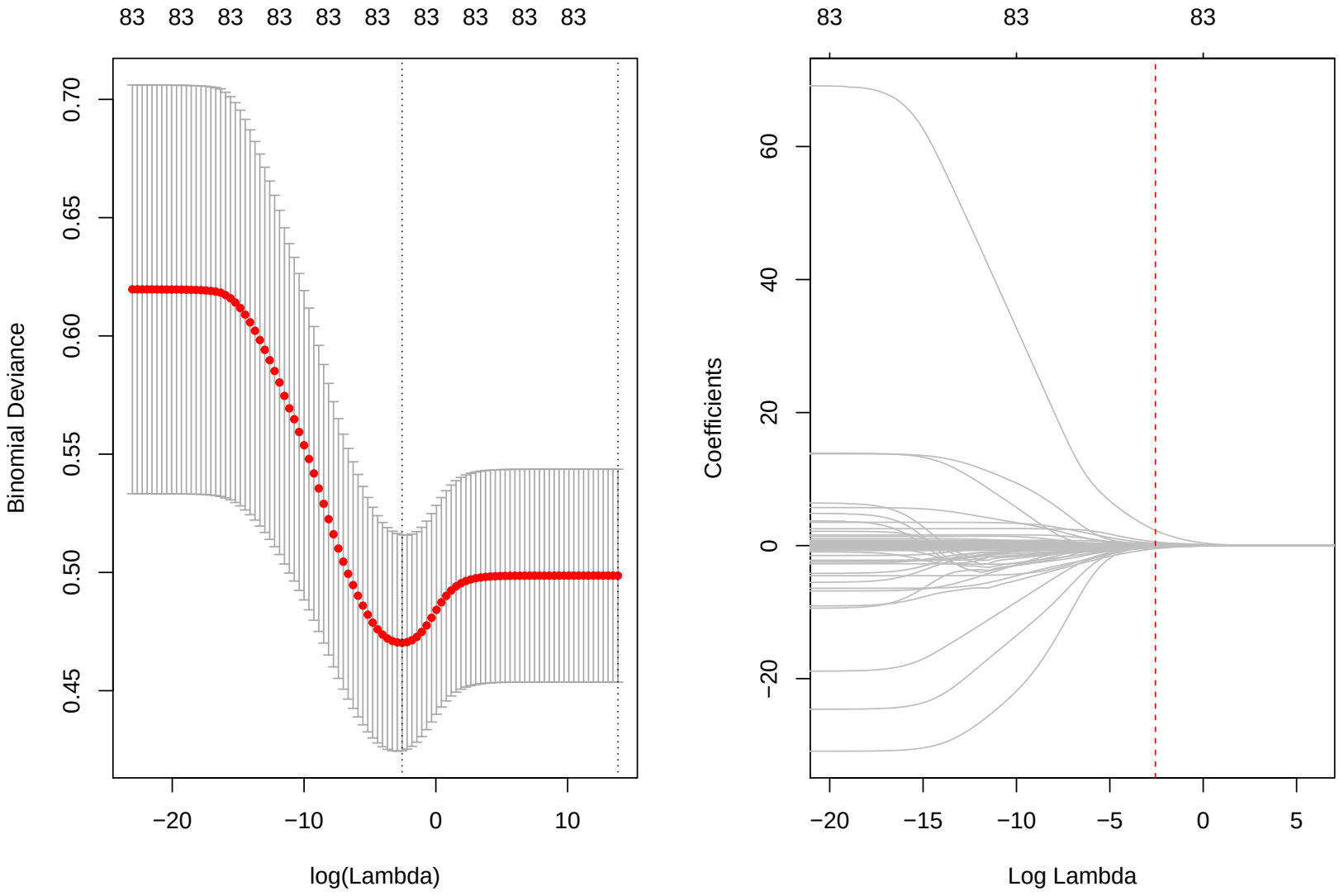


Example: diabetes progression - comparison

Table 1: Coefficient estimates

	OLS	Ridge
intercept	150.2964	147.6457
age	-103.1102	-46.0302
sex	-432.1300	-167.9041
bmi	60.8084	253.7881
map	714.0397	439.3918
tc	-1236.1539	-147.1544
ldl	764.5922	-95.2515
hdl	-106.8334	-345.4657
tch	505.2936	286.4428
ltg	617.6187	329.4664
glu	347.9019	235.6980

Example: caravan insurance - $n = 1164, p = 83$



Example: caravan insurance - prediction

Table 2: Logistic regression

	No	Yes
No	912	22
Yes	62	4

Table 3: Ridge regression

	No	Yes
No	932	2
Yes	66	0

Wrap up on Ridge

Ridge regression is a regularisation method that can be helpful when

- OLS coefficients may be poorly determined because of high correlation between regressors
- extreme variability in the training data is observed, because of low sample size and/or large p relative to n.

Wrap up on Ridge

Ridge regression is a regularisation method that can be helpful when

- OLS coefficients may be poorly determined because of high correlation between regressors
- extreme variability in the training data is observed, because of low sample size and/or large p relative to n.

Ridge improves over OLS by imposing a size constraint on the coefficient estimates; this approach

- typically has lower fit on training data than OLS, but is less prone to overfitting
- usually generalizes better, because of higher robustness to extreme variability
- involves hyperparameter tuning $\rightarrow$ cross-validation
- readily extends to more general models, such as glms.

Shrinkage and variable selection: the Lasso

The Lasso

The idea of constraining the size of the OLS estimates can be extended to consider different kinds of penalizations.

While the ridge penalty encompasses the ℓ_2 norm of the estimates vector, the **lasso** makes use of the ℓ_1 norm:

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

where $\lambda \geq 0$ is again a complexity parameter.

The Lasso

The idea of constraining the size of the OLS estimates can be extended to consider different kinds of penalizations.

While the ridge penalty encompasses the ℓ_2 norm of the estimates vector, the **lasso** makes use of the ℓ_1 norm:

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

where $\lambda \geq 0$ is again a complexity parameter.

The lasso possesses an important property that ridge doesn't have: it allows for automatic **variable selection**.

Lasso coefficients profiles

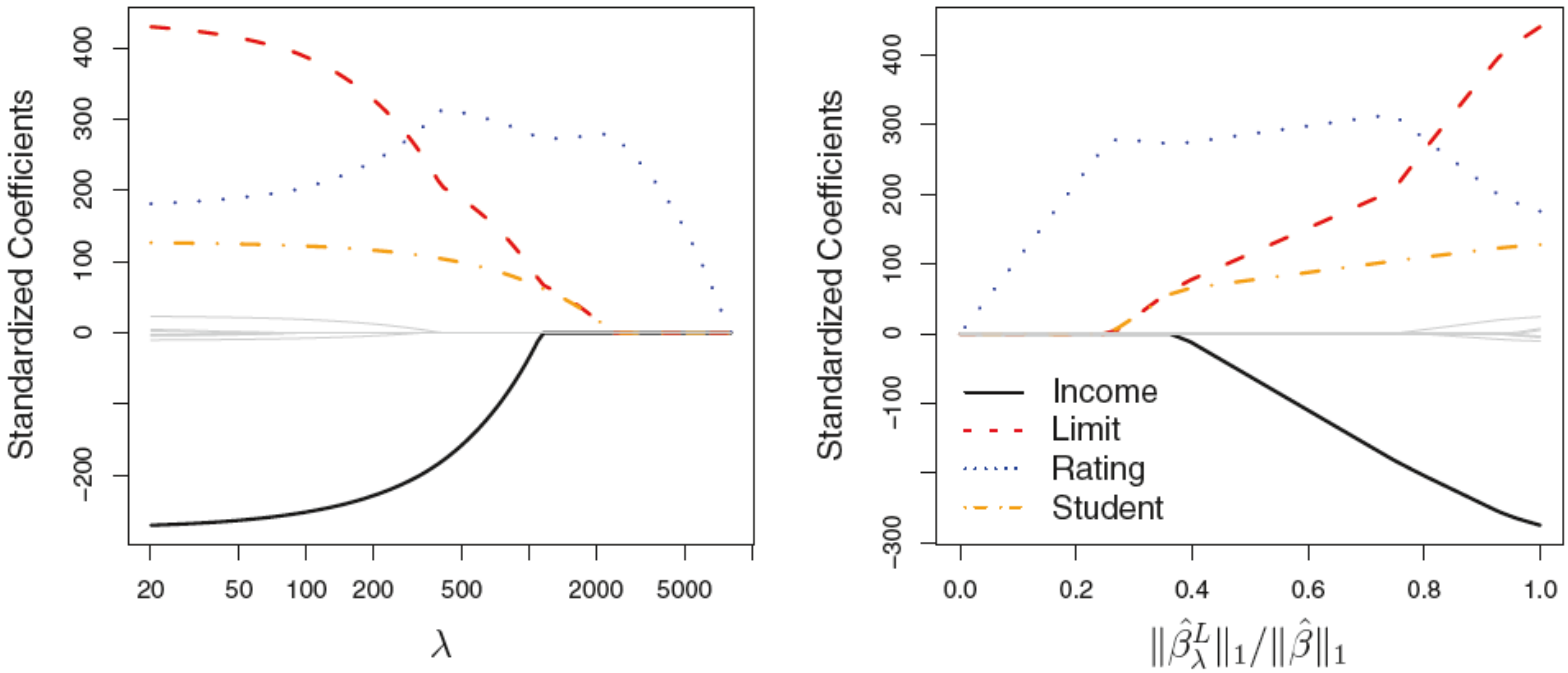
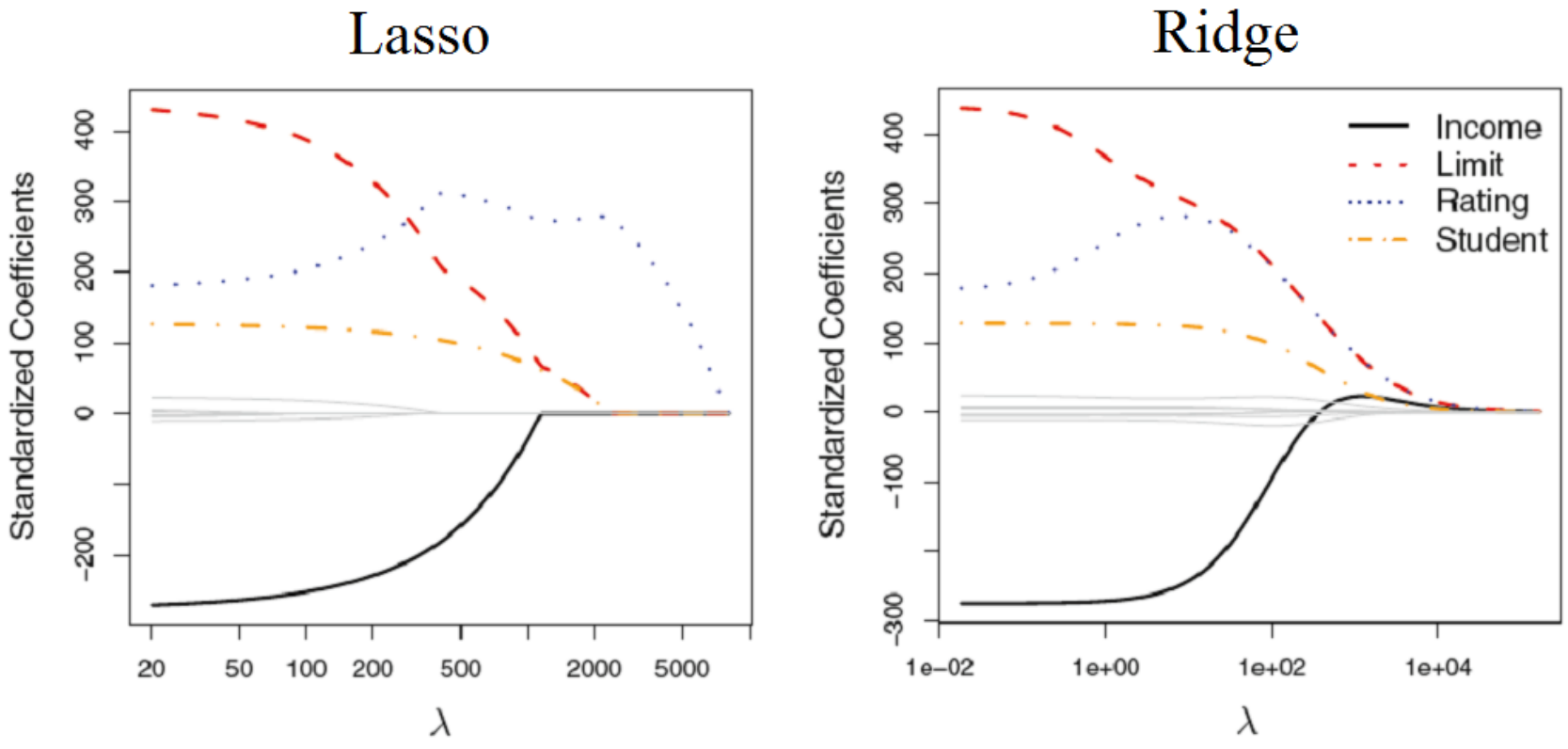


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Comparing lasso and ridge: profiles



Comparing lasso and ridge: constraints

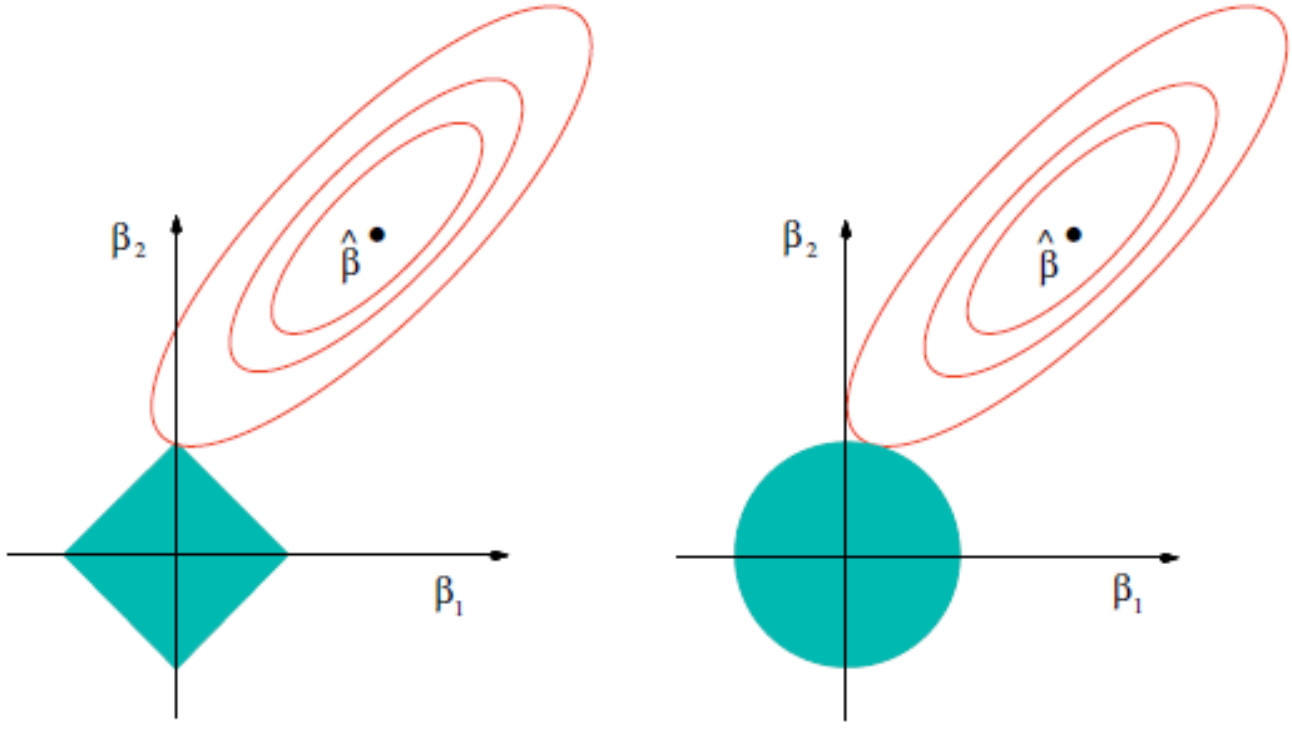
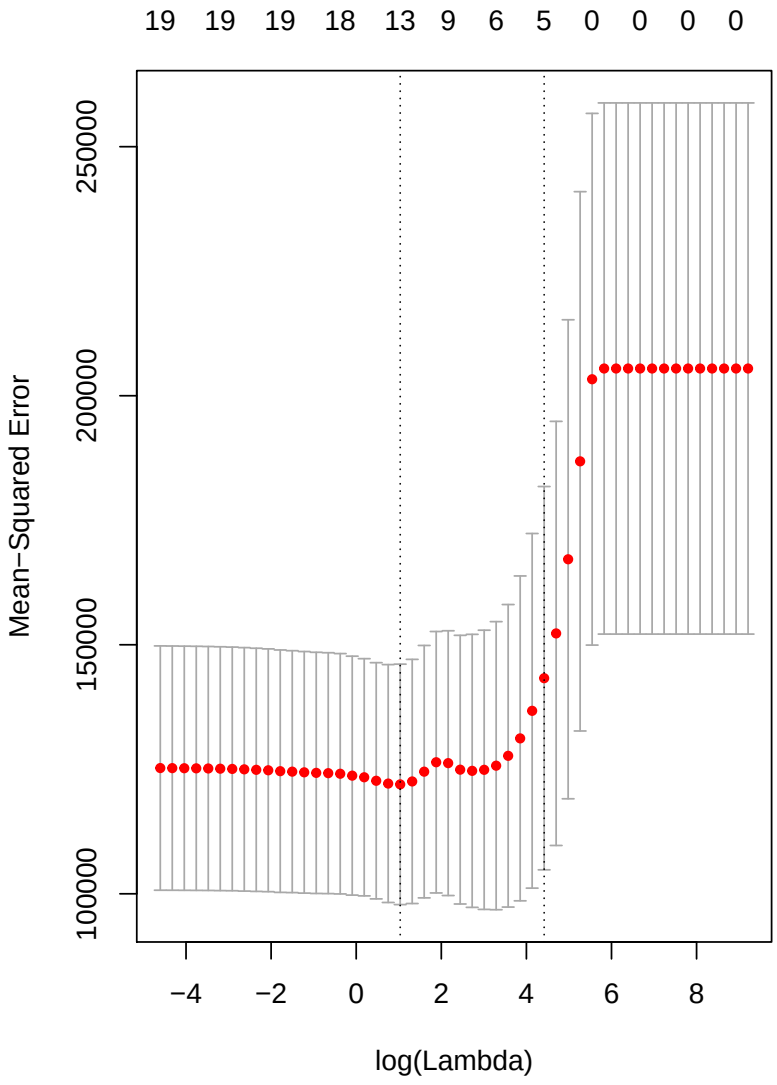
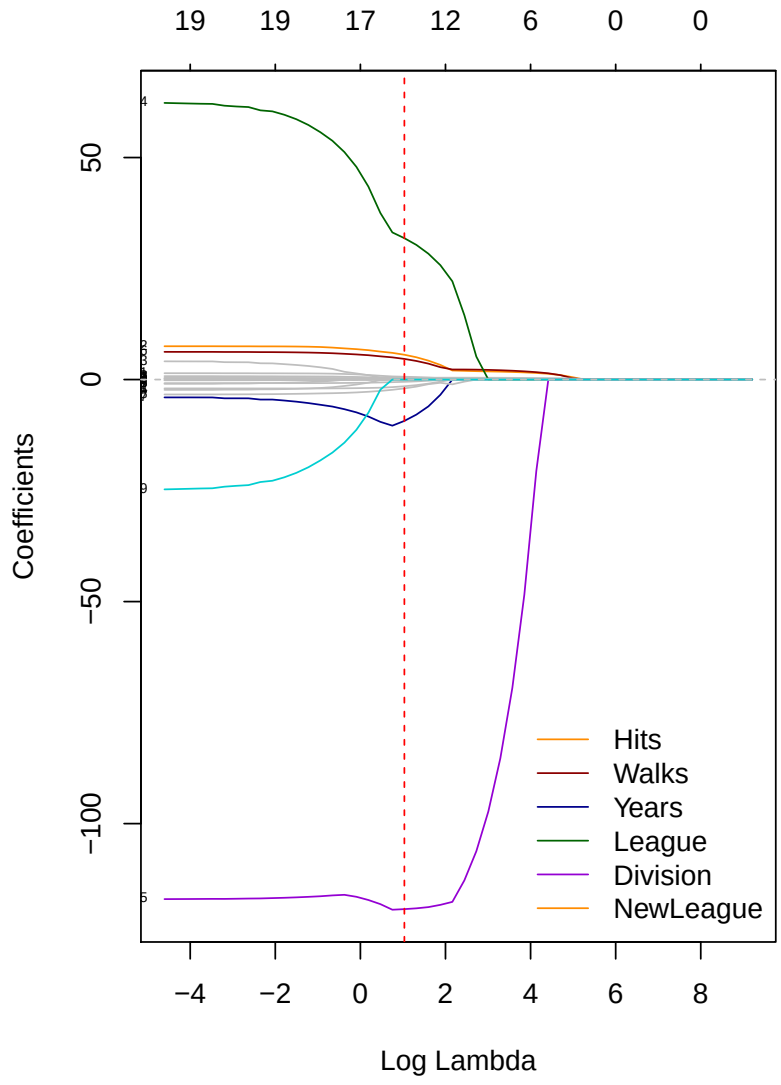


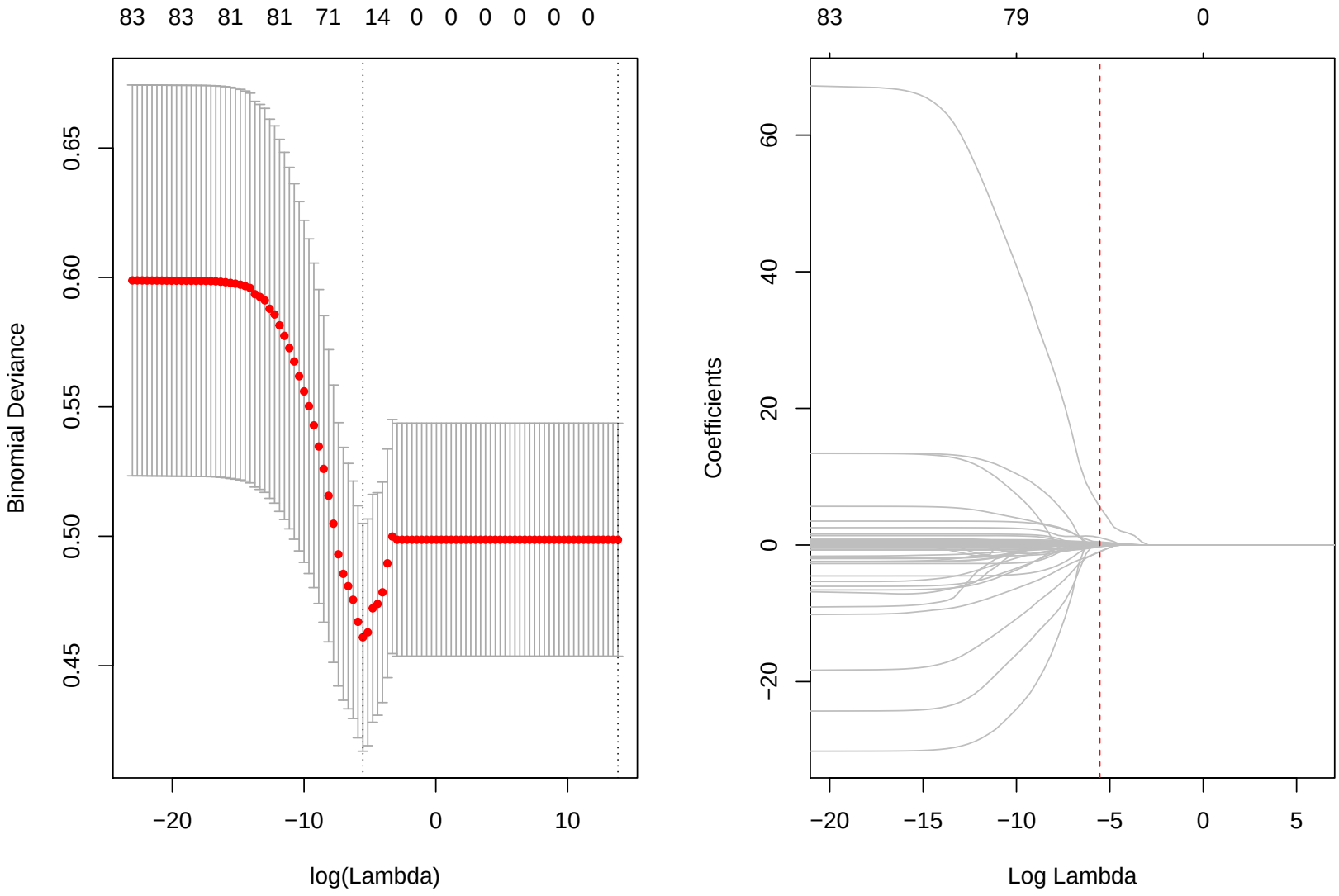
Figure 2.2 *Estimation picture for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions* $|\beta_1|+|\beta_2| \leq t$ *and* $\beta_1^2+\beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the residual-sum-of-squares function. The point* $\widehat{\beta}$ *depicts the usual (unconstrained) least-squares estimate.*

Choosing λ: cross-validation





Example: caravan insurance



Example: caravan insurance - prediction

Table 4: Ridge regression

	No	Yes
No	932	2
Yes	66	0

Table 5: Lasso regression

	No	Yes
No	927	7
Yes	66	0

Wrap-up on lasso

Lasso regression is a regularisation and variable selection method that can be especially helpful

- if variable selection is advisable to improve interpretability of the final model (sparsity)

Wrap-up on lasso

Lasso regression is a regularisation and variable selection method that can be especially helpful

- if variable selection is advisable to improve interpretability of the final model (sparsity)
- when faced with *wide* data, for which $p \gg n$

Wrap-up on lasso

Lasso regression is a regularisation and variable selection method that can be especially helpful

- if variable selection is advisable to improve interpretability of the final model (sparsity)
- when faced with *wide* data, for which $p \gg n$
- for statistical and computational efficiency.

Many extensions already exist: check out the **group-lasso** for dealing with dummy variables, and the **fused lasso** for time series and functional data.

Wrap-up on lasso

Lasso regression is a regularisation and variable selection method that can be especially helpful

- if variable selection is advisable to improve interpretability of the final model (sparsity)
- when faced with *wide* data, for which $p \gg n$
- for statistical and computational efficiency.

Many extensions already exist: check out the **group-lasso** for dealing with dummy variables, and the **fused lasso** for time series and functional data.

Note: under lasso, at most n coefficients can be equal to zero and the solution is nonlinear in $\mathbf{y}$ (no closed form).

More general penalties and the elastic nets

The ℓ_q penalty

The shrinkage penalty term can be readily generalised to $\lambda \sum_{j=1}^{p} |\beta_j|^q$, with $q > 0$.

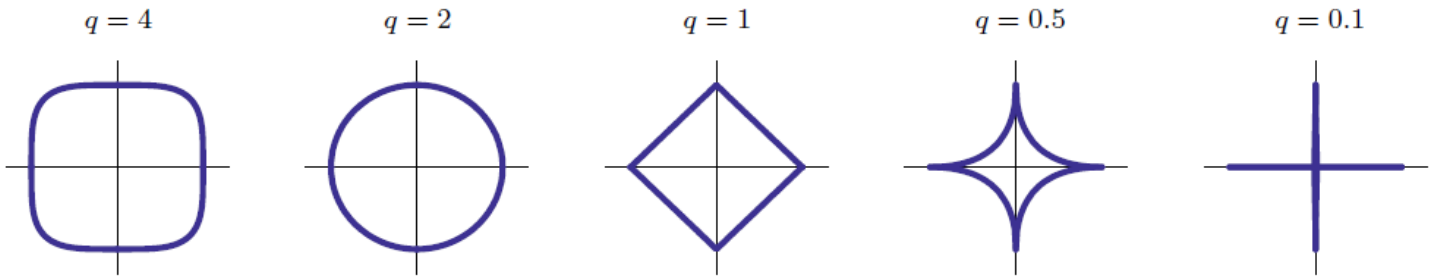


Figure 2.6 *Constraint regions* $\sum_{j=1}^{p} |\beta_j|^q \leq 1$ *for different values of* q. *For* $q < 1$, *the constraint region is nonconvex.*

When $q = 2$, we have the ridge penalty, $q = 1$ is lasso, and as $q \to 0$ we approach the so-called *subset selection* method.

A hybrid penalty: the elastic nets

The lasso sometimes does not perform well with highly correlated variables, and often performs worse than ridge in prediction.

To overcome this limitations, a penalty that combines the ℓ_1 and ℓ_2 constraints has been developed.

A hybrid penalty: the elastic nets

The lasso sometimes does not perform well with highly correlated variables, and often performs worse than ridge in prediction.

To overcome this limitations, a penalty that combines the ℓ_1 and ℓ_2 constraints has been developed.

An **elastic net** is a regularisation and variable selection procedure that makes use of the penalty

$$\lambda \left[\frac{1}{2}(1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right]$$

where $\alpha \in [0, 1]$ is called the **mixing** parameter and λ has the usual interpretation. Lasso and ridge are special cases, respectively for $\alpha = 1$ and $\alpha = 0$.

More on the motivation for elastic nets

Consider the following scenarios:

- in the $p > n$ case, the lasso can select at most n variables before it saturates

More on the motivation for elastic nets

Consider the following scenarios:

- in the $p > n$ case, the lasso can select at most n variables before it saturates
- if there is a group of variables with very high pairwise correlations, the lasso tends to select only one variable from the group, not caring which one

More on the motivation for elastic nets

Consider the following scenarios:

- in the $p > n$ case, the lasso can select at most n variables before it saturates
- if there is a group of variables with very high pairwise correlations, the lasso tends to select only one variable from the group, not caring which one
- for usual $n > p$ situations, if there are high correlations between predictors, the prediction performance of lasso is poor with respect to ridge.

In these situations, a more general approach is advised.

Comparing elastic nets and lasso: constraints

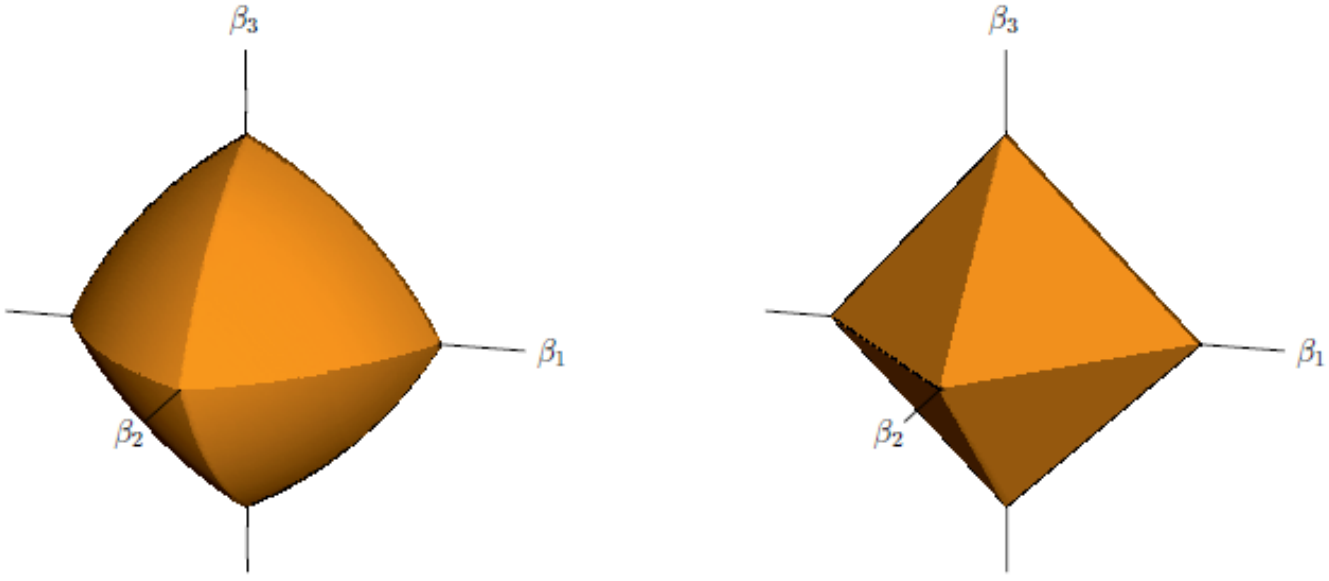


Figure 4.2 *The elastic-net ball with $\alpha = 0.7$ (left panel) in $\mathbb{R}^3$, compared to the ℓ_1 ball (right panel). The curved contours encourage strongly correlated variables to share coefficients (see Exercise 4.2 for details).*

The choice of α and λ

The mixing parameter α governs the extent to which the elastic net behaves as a ridge or a lasso. As $\alpha \to 0$, the ridge penalty gains more weight than the lasso; the opposite happens when $\alpha \to 1$.

The choice of α and λ

The mixing parameter α governs the extent to which the elastic net behaves as a ridge or a lasso. As $\alpha \to 0$, the ridge penalty gains more weight than the lasso; the opposite happens when $\alpha \to 1$.

In practice, one usually constructs a grid of M α values, say $\{\alpha_1, ..., \alpha_M\}$, chooses a folds configuration, and for each $m = 1, ..., M$:

1. k-fold cross-validates λ for given $\alpha = \alpha_m$
2. stores the test MSE profile.

The M MSE profiles are then compared, and the α associated with the preferred one is chosen. The best λ within the selected profile is then used for modelling.

The choice of α and λ

The mixing parameter α governs the extent to which the elastic net behaves as a ridge or a lasso. As $\alpha \to 0$, the ridge penalty gains more weight than the lasso; the opposite happens when $\alpha \to 1$.

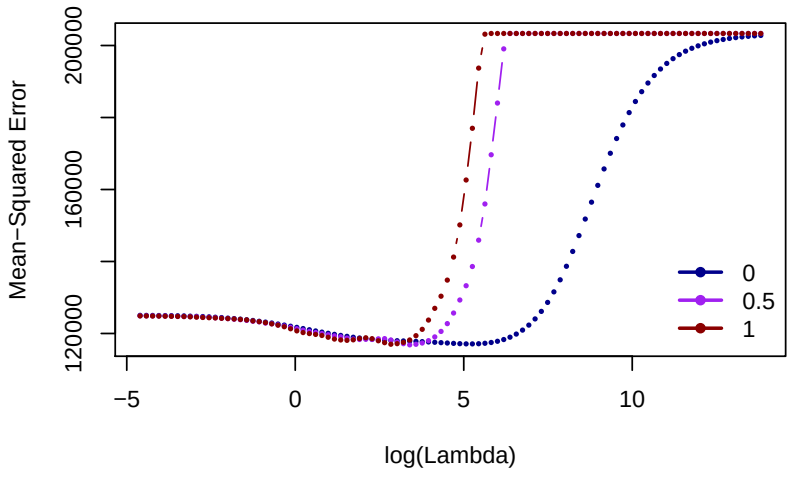
In practice, one usually constructs a grid of M α values, say $\{\alpha_1, ..., \alpha_M\}$, chooses a folds configuration, and for each $m = 1, ..., M$:

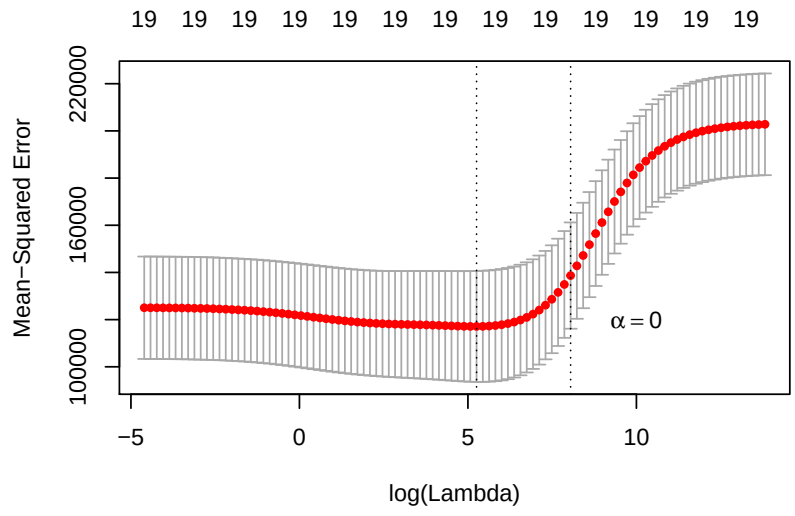
1. k-fold cross-validates λ for given $\alpha = \alpha_m$
2. stores the test MSE profile.

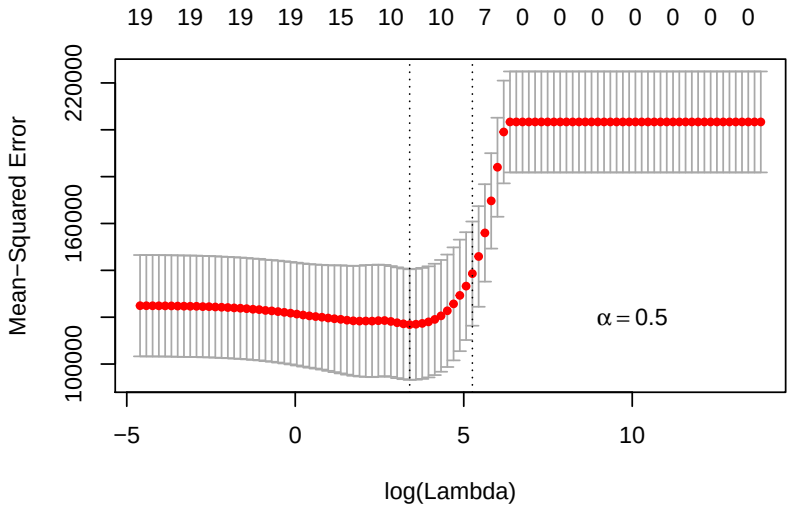
The M MSE profiles are then compared, and the α associated with the preferred one is chosen. The best λ within the selected profile is then used for modelling.

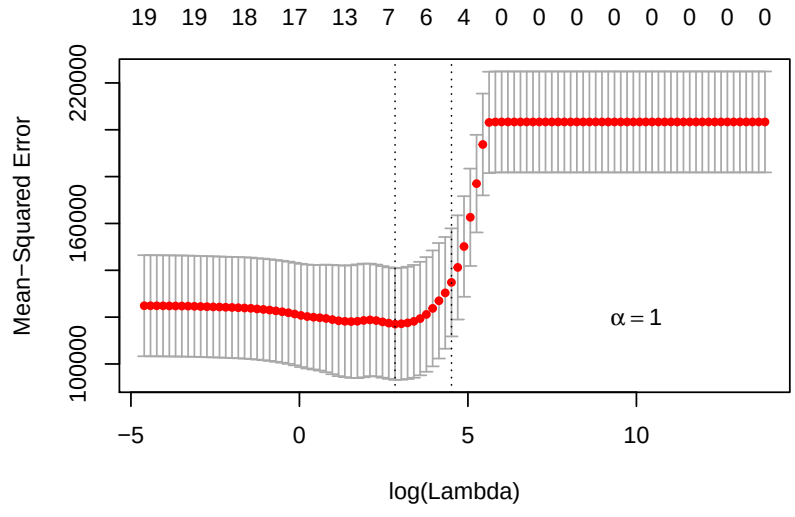
Note: joint cross-validation of α and λ is usually not recommended for computational reasons.

The choice of α and λ

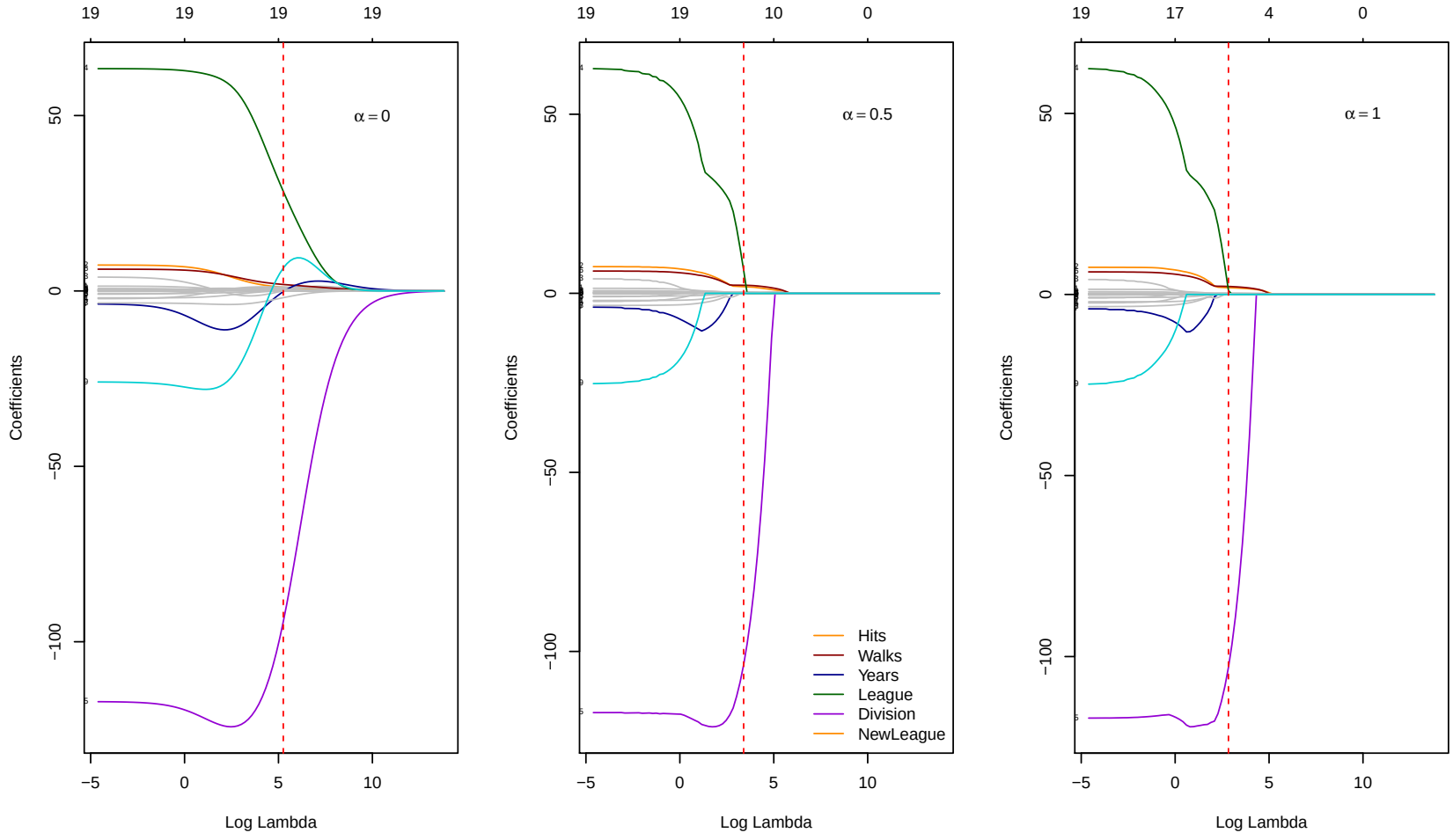




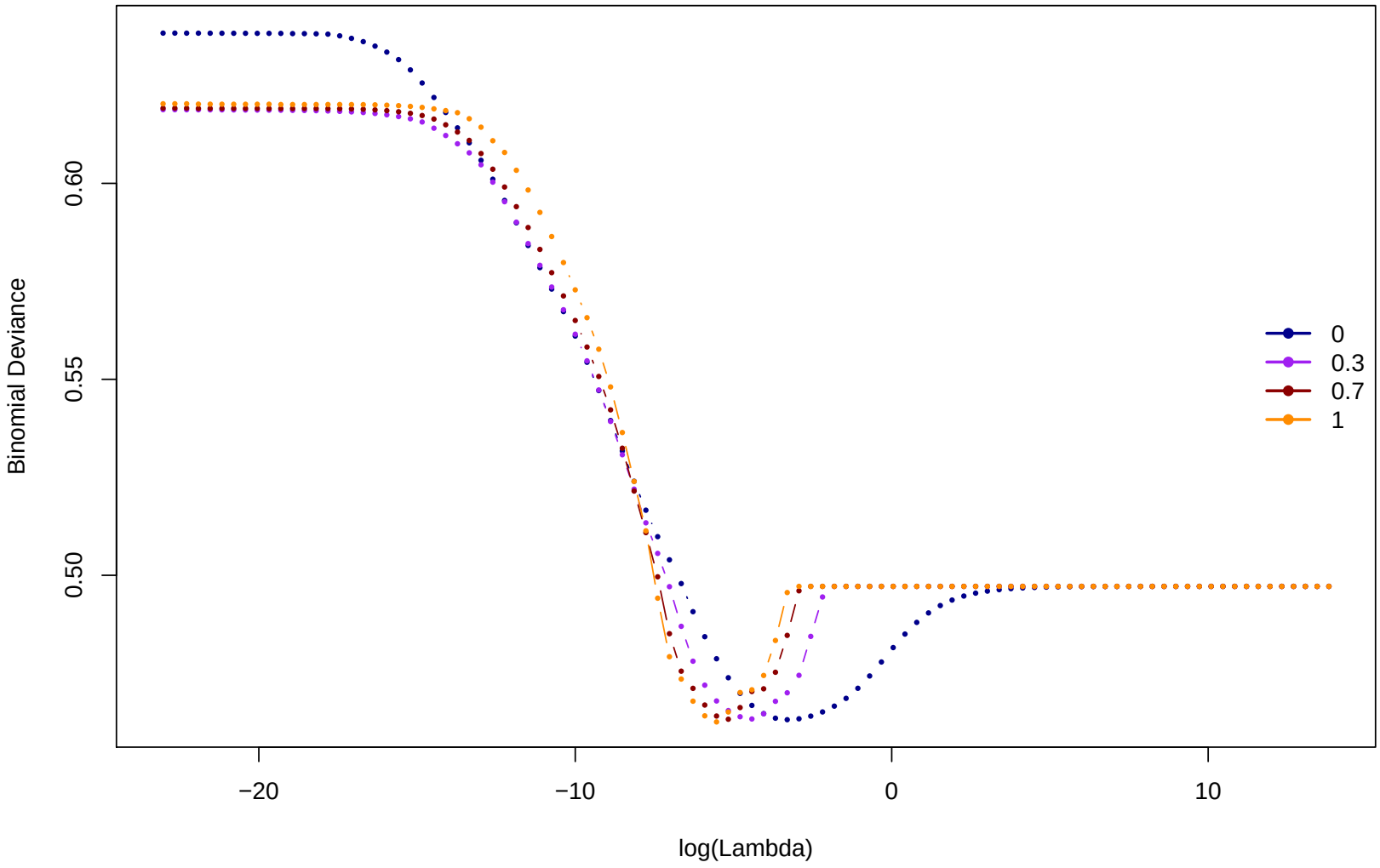




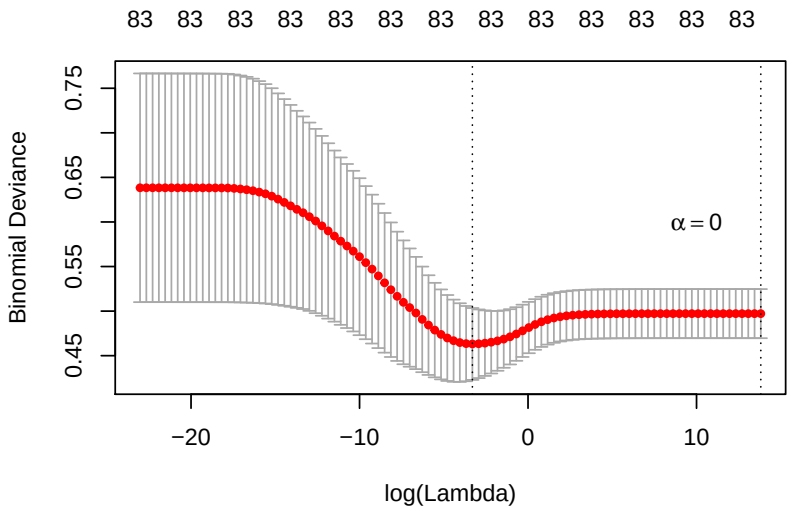
Coefficient profiles

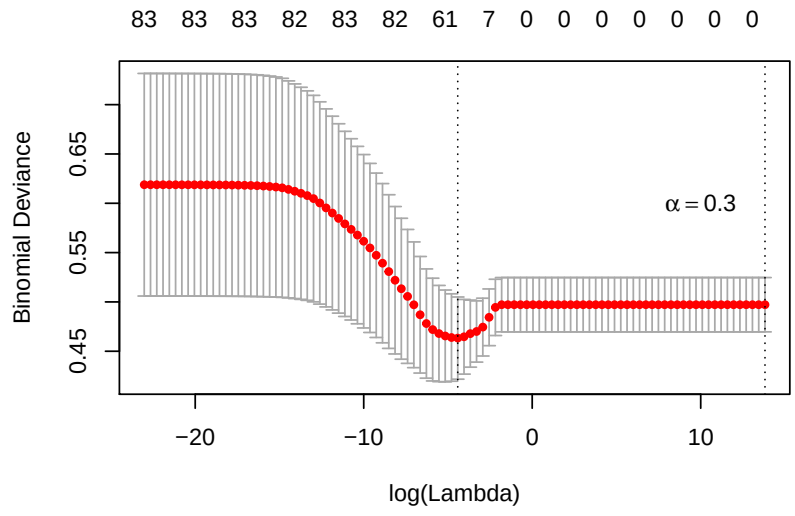


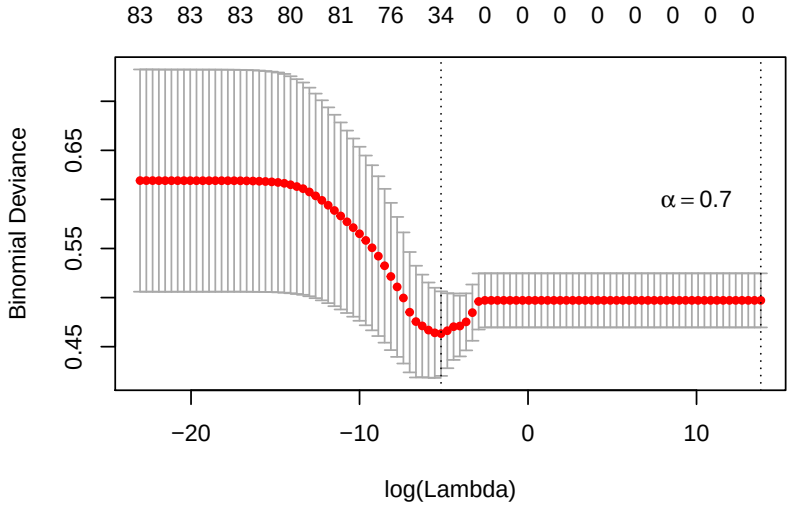
Example: caravan insurance

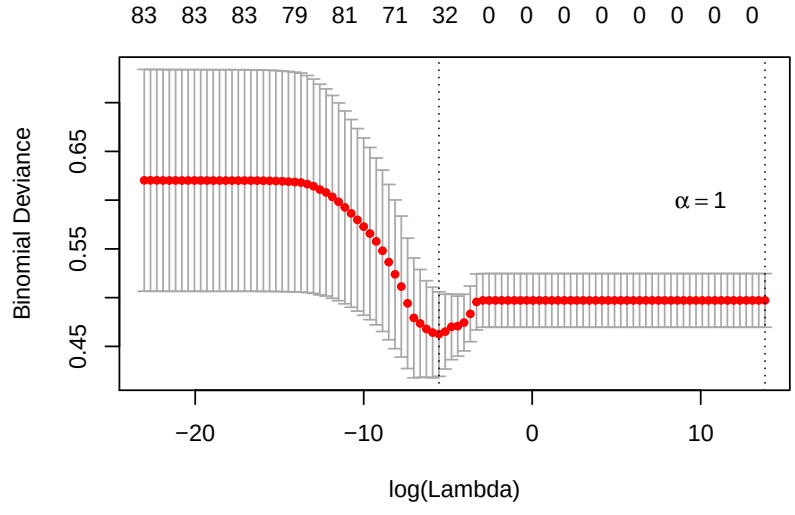


Example: caravan insurance - test MSE

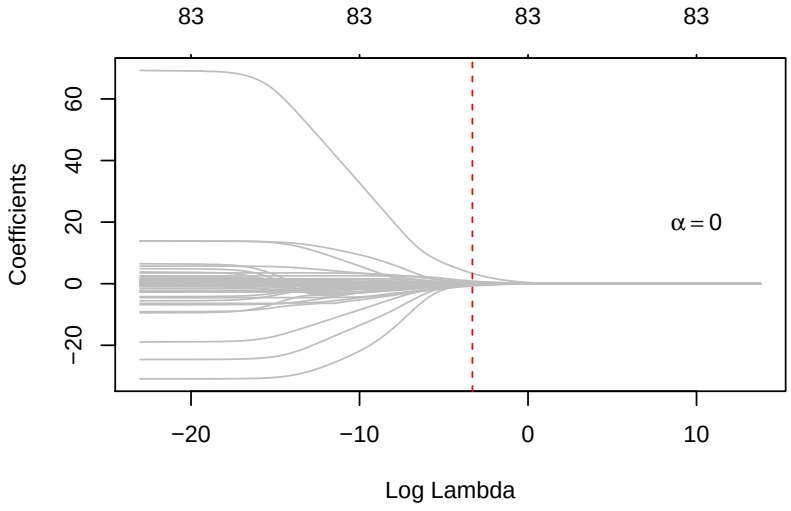


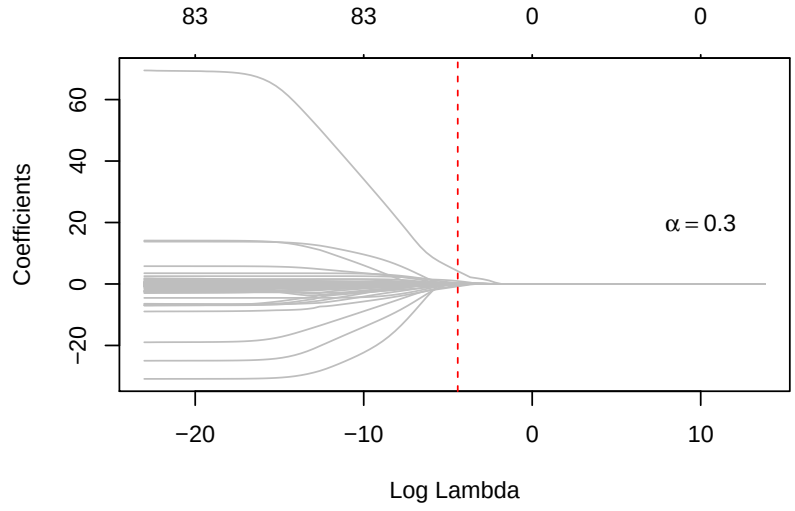


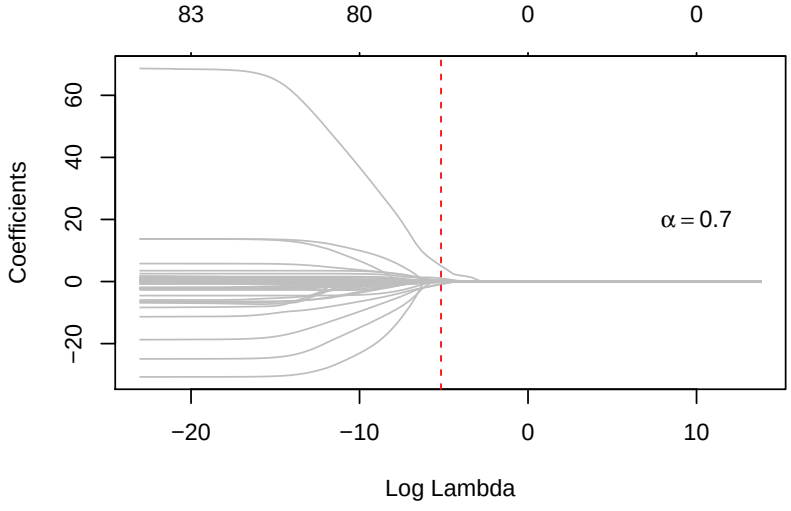


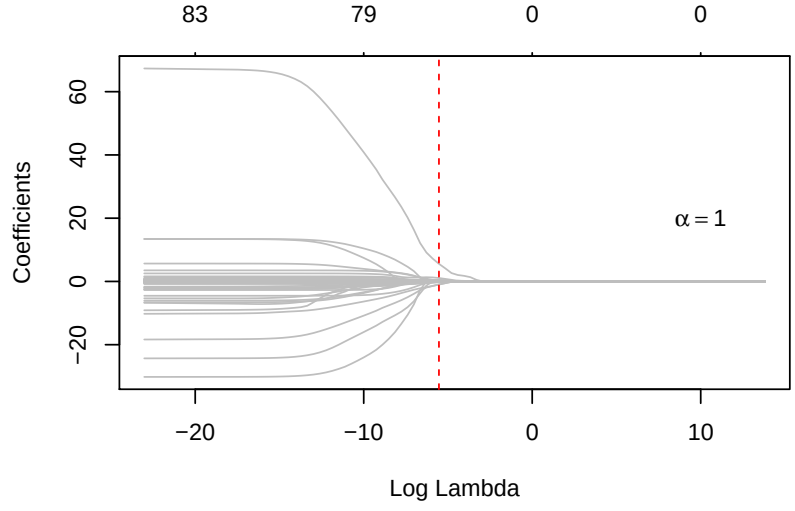


Example: caravan insurace - profiles









Example: caravan insurance - prediction

Table 6: Accuracy in prediction

	% correct
Plain logistic	91.6
Ridge logistic	93.2
Lasso logistic	92.7
EN alpha=0	93.2
EN alpha=0.3	92.9
EN alpha=0.7	92.8
EN alpha=1	92.7

Wrap-up on elastic nets

Elastic net regression is a regularisation and variable selection procedure that overcomes some of the limitations of the lasso by borrowing strength from the ridge. Specifically, it

- allows to select more than n variables

Wrap-up on elastic nets

Elastic net regression is a regularisation and variable selection procedure that overcomes some of the limitations of the lasso by borrowing strength from the ridge. Specifically, it

- allows to select more than n variables
- tends to jointly select or leave out groups of highly correlated variables

Wrap-up on elastic nets

Elastic net regression is a regularisation and variable selection procedure that overcomes some of the limitations of the lasso by borrowing strength from the ridge. Specifically, it

- allows to select more than n variables
- tends to jointly select or leave out groups of highly correlated variables
- improves the predictive performance w.r.t. lasso

Wrap-up on elastic nets

Elastic net regression is a regularisation and variable selection procedure that overcomes some of the limitations of the lasso by borrowing strength from the ridge. Specifically, it

- allows to select more than n variables
- tends to jointly select or leave out groups of highly correlated variables
- improves the predictive performance w.r.t. lasso
- is readily extendable to use with more general methods, such as glm.

Elastic nets are especially useful when a sparse solution is either necessary or desirable (such as in $p \gg n$ problems) and small groups of highly correlated predictors are present.

Bibliography

Bibliography

- Hastie T, Tibshirani R and Friedman J (2017). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. 2n edition, corrected 12th printing. Springer.
- Hastie T, Tibshirani R and Wainwright M (2015). *Statistical Learning with Sparsity - The Lasso and Generalizations*. CRC press, Boca Raton.
- James G, Witten D, Hastie T and Tibshirani R (2017). *An Introduction to Statistical Learning - with Applications in R*. Corrected 8th printing. Springer.
- Tibshirani R (2011). *Regression shrinkage and selection via the lasso: a retrospective*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73:3, 273-282.
- Zou H and Hastie T (2005). *regularisation and variable selection via the elastic net*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67:2, 301-320.