STA4026S – Honours Analytics
Section A – Theory and Application of Supervised Learning
Lecture 2 – Linear Model Selection & Regularisation

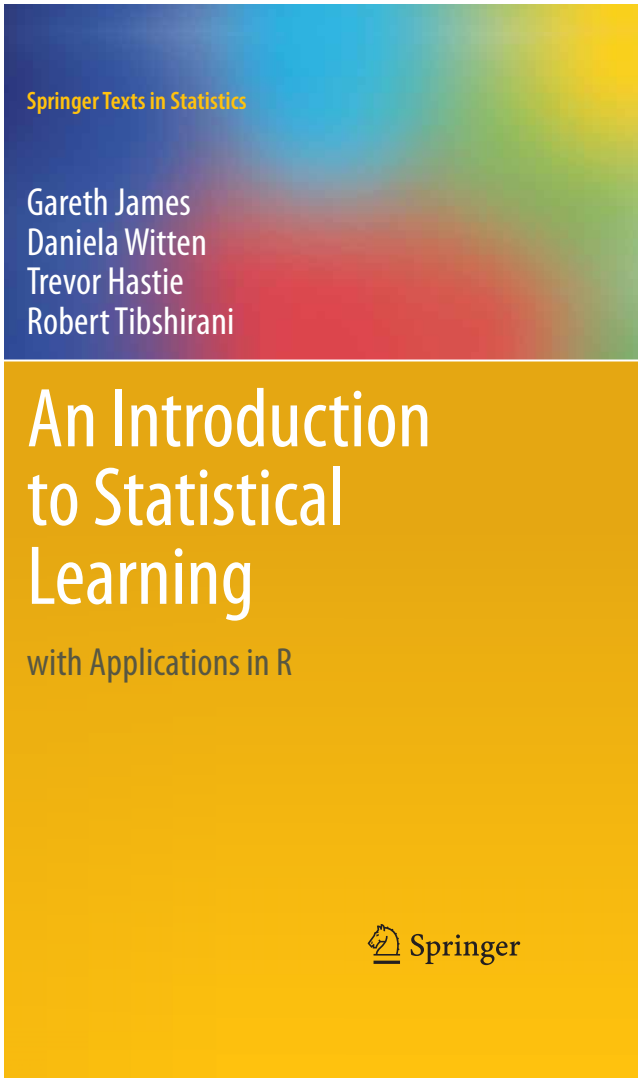
Stefan S. Britz
stefan.britz@uct.ac.za

Department of Statistical Sciences
University of Cape Town

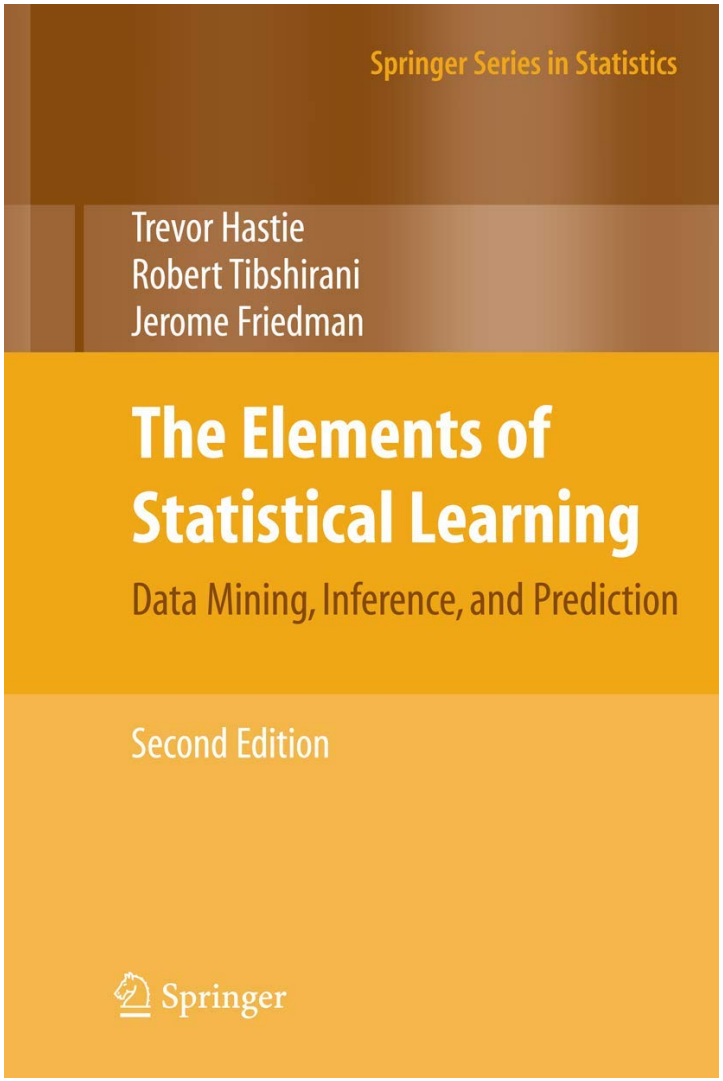
Introduction

- Now that we have explored the dynamics of model complexity and ways of testing model fit, we will turn our attention to **variable selection**
- We will note the existence of **subset selection** procedures
- Although our focus will be on **regularisation** as a method for controlling model complexity
- Note that these methods do not pertain only to linear models, but we will use this class to illustrate the principles

Suggested reading



Chapters 3 & 6



Chapter 3

Linear Regression Models

Model specification

For some real-valued output Y and input vector $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$, the model is defined as:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon, \tag{1}$$

where $\epsilon \sim N(0, \sigma^2)$.

Although few real-world relationships can be considered truly linear, the linear model offers some distinct advantages, most notably in the clear interpretation of features.

Furthermore, they often perform surprisingly well on a range of problems.

Parameter Estimation

The most popular method of estimating the regression parameters based on the training set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$, is **ordinary least squares** (OLS), where we find the coefficients $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]'$ to minimise the residual sum of squares:

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

noting that this does not imply any assumptions on the validity of the model.

Parameter Estimation

To minimise $RSS(\boldsymbol{\beta})$, let us first write Equation 1 in matrix form:

$$ {}_n\boldsymbol{Y}_1 = {}_n\boldsymbol{X}_{(p+1)}\boldsymbol{\beta}_1 + {}_n\boldsymbol{\epsilon}_1, \tag{3} $$

where the first column of $\boldsymbol{X}$ is $\mathbf{1} : n \times 1$

Now we can write

$$ RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{4} $$

which is a quadratic function in the $p + 1$ parameters.

Parameter Estimation

Differentiating Equation 4 with respect to $\boldsymbol{\beta}$ yields

$$\frac{\partial RSS}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{5}$$

$$\frac{\partial^2 RSS}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 2\boldsymbol{X}'\boldsymbol{X} \tag{6}$$

If $\boldsymbol{X}$ is of full column rank – a reasonable assumption when $n \geq p$ – then $\boldsymbol{X}'\boldsymbol{X}$ is positive definite.

When will $\boldsymbol{X}$ be of less than full column rank?

Parameter Estimation

Setting the first derivative equal to $\mathbf{0}$ and solving for $\boldsymbol{\beta}$:

$$\begin{aligned}
\frac{\partial RSS}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{X}'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) = \mathbf{0} \\
\boldsymbol{X}'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) &= \mathbf{0} \\
\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \mathbf{0} \\
\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}'\boldsymbol{y} \\
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad (7)
\end{aligned}$$

These coefficient estimates define a fitted linear regression model.

The focus of this section is on methods for improving predictive accuracy, therefore we will not cover inference on the regression parameters or likelihood ratio tests here.

Variable Selection

- Do we need all the variables in the model?
- Can we improve predictive performance by simplifying the model?
- How would we go about either downweighting variables or removing them entirely?
- Before looking into **regularisation** for this task, first note subset selection

Subset Selection

Although subject to much criticism, there are some specific conditions in which subset selection could yield satisfactory results, for example when p is small and there is little to no multicollinearity.

This selection can generally be done in two ways:

1. Best subset selection

   This approach identifies the best fitting model across all 2^p combinations of predictors, by first identifying the best k-variable model $\mathcal{M}_k$ according to RSS, for all $k = 1, 2, \ldots, p$.

2. Stepwise selection

   Starting with either the null (forward stepwise) or saturated (backward stepwise) model.
   Sequentially add or remove predictors respectively according to some improvement metric.
   One can also apply a hybrid method.

Subset Selection

Although subject to much criticism, there are some specific conditions in which subset selection could yield satisfactory results, for example when p is small and there is little to no multicollinearity.

This selection can generally be done in two ways:

1. Best subset selection

   This approach identifies the best fitting model across all 2^p combinations of predictors, by first identifying the best k-variable model $\mathcal{M}_k$ according to RSS, for all $k = 1, 2, \ldots, p$.

2. Stepwise selection

   Starting with either the null (forward stepwise) or saturated (backward stepwise) model.
   Sequentially add or remove predictors respectively according to some improvement metric.
   One can also apply a hybrid method.

Subset Selection

Although subject to much criticism, there are some specific conditions in which subset selection could yield satisfactory results, for example when p is small and there is little to no multicollinearity.

This selection can generally be done in two ways:

1. Best subset selection

   This approach identifies the best fitting model across all 2^p combinations of predictors, by first identifying the best k-variable model $\mathcal{M}_k$ according to RSS, for all $k = 1, 2, \ldots, p$.

2. Stepwise selection

   Starting with either the null (forward stepwise) or saturated (backward stepwise) model.
   Sequentially add or remove predictors respectively according to some improvement metric.
   One can also apply a hybrid method.

Subset Selection?

- Typically, either Mallow's C_p, AIC, BIC, or adjusted R^2 is used for model comparison in subset selection
- Subset selection is a discrete process, with variables either retained or discarded
- Often exhibits high variance, thereby failing to reduce the test MSE
- **Regularisation** offers a more continuous, general-purpose and usually quicker method of controlling model complexity.
- Can be applied to any parametric model (ANNs – weight decay).

L_1 and L_2 Regularisation

Regularisation

- Alternative to OLS
- Fit a model containing all p predictors, but constrain or **regularise** the coefficient estimates
- Also referred to as **shrinkage methods**, since we shrink the coefficient estimates towards zero by imposing a penalty on their size
- Can significantly reduce the coefficient estimates' variance, thereby reducing the variance component of the total error
- The two best-known regularisation techniques are **ridge regression** and **the lasso**

Ridge regression – L_2 regularisation

Initially developed as a method of dealing with highly correlated predictors in regression analysis.

Instead of finding regression coefficients to minimise 2, the ridge coefficients minimise a penalised residual sum of squares:

$$\hat{\boldsymbol{\beta}}_R = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{8}$$

- $\lambda \geq 0$ is a complexity parameter, controlling the amount of shrinkage
- $(\lambda = 0) \equiv$ OLS
- As λ increases, the coefficients are shrunk towards zero

Ridge regression – L_2 regularisation

- Ridge regression is also referred to as L_2 **regularisation**
- Also stylised as ℓ_2
- The regularisation penalty is based on the L_2 norm (Euclidean norm) of the regression coefficients
- The L_2 norm of a vector $\boldsymbol{\beta}$ is given by $||\boldsymbol{\beta}||_2 = \sqrt{\sum_{i=1}^{p} \beta_i^2}$

Ridge regression – L_2 regularisation

The optimisation problem in 8 can also be written as

$$\hat{\boldsymbol{\beta}}_R = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq \tau, \tag{9}$$

where τ, representing the explicit size constraint on the parameters, has a one-to-one correspondence with λ in 8.

Ridge regression – L_2 regularisation

- Imposing a size constraint on the coefficients addresses the problem of multicollinearity
- Ridge solutions are not equivariant under scaling of the inputs
- Therefore, inputs are generally **standardised** first: x_{ij} replaced with $\frac{x_{ij} - \bar{x}_j}{s_{x_j}}$
- The intercept (β_0) is not penalised, and is estimated by $\bar{y}$

Ridge regression – L_2 regularisation

Assuming this centering has been done, the input matrix $\boldsymbol{X}$ then becomes $n \times p$, such that the penalised RSS, now viewed as a function of λ, can be written as

$$RSS(\lambda) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}'\boldsymbol{\beta}, \tag{10}$$

yielding the ridge regression solution

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}, \tag{11}$$

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

Ridge regression – L_2 regularisation

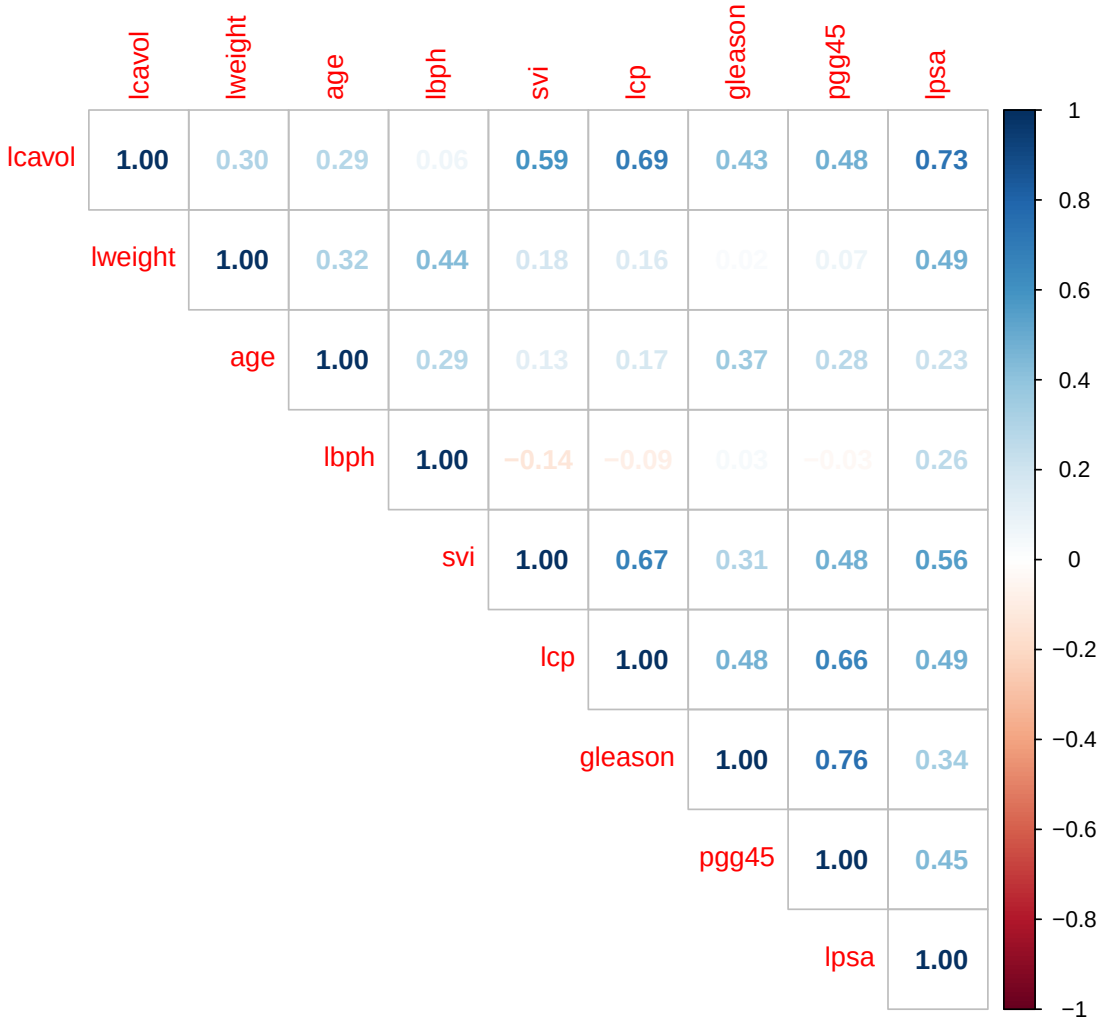
- Equation 11 shows that the ridge regression addresses singularity issues that can arise when the predictor variables are highly correlated
- Even if $\boldsymbol{X}'\boldsymbol{X}$ is singular, the modified matrix $\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}$ is guaranteed to be non-singular
- This allows for stable and well-defined solutions to be obtained
- Let's explore this via an example

Example 2 – Prostate Cancer

- The goal is to model the (log) prostate-specific antigen (`lpsa`) for men who were about to receive a radical prostatectomy
- $n = 97$, $p = 8$
- 30 observations set aside for testing
- `svi` and `gleason` are actually binary and ordinal variables respectively, but we will treat them as numeric for the sake of simplicity in this illustration

Example 2 – Prostate Cancer

Correlation plot:



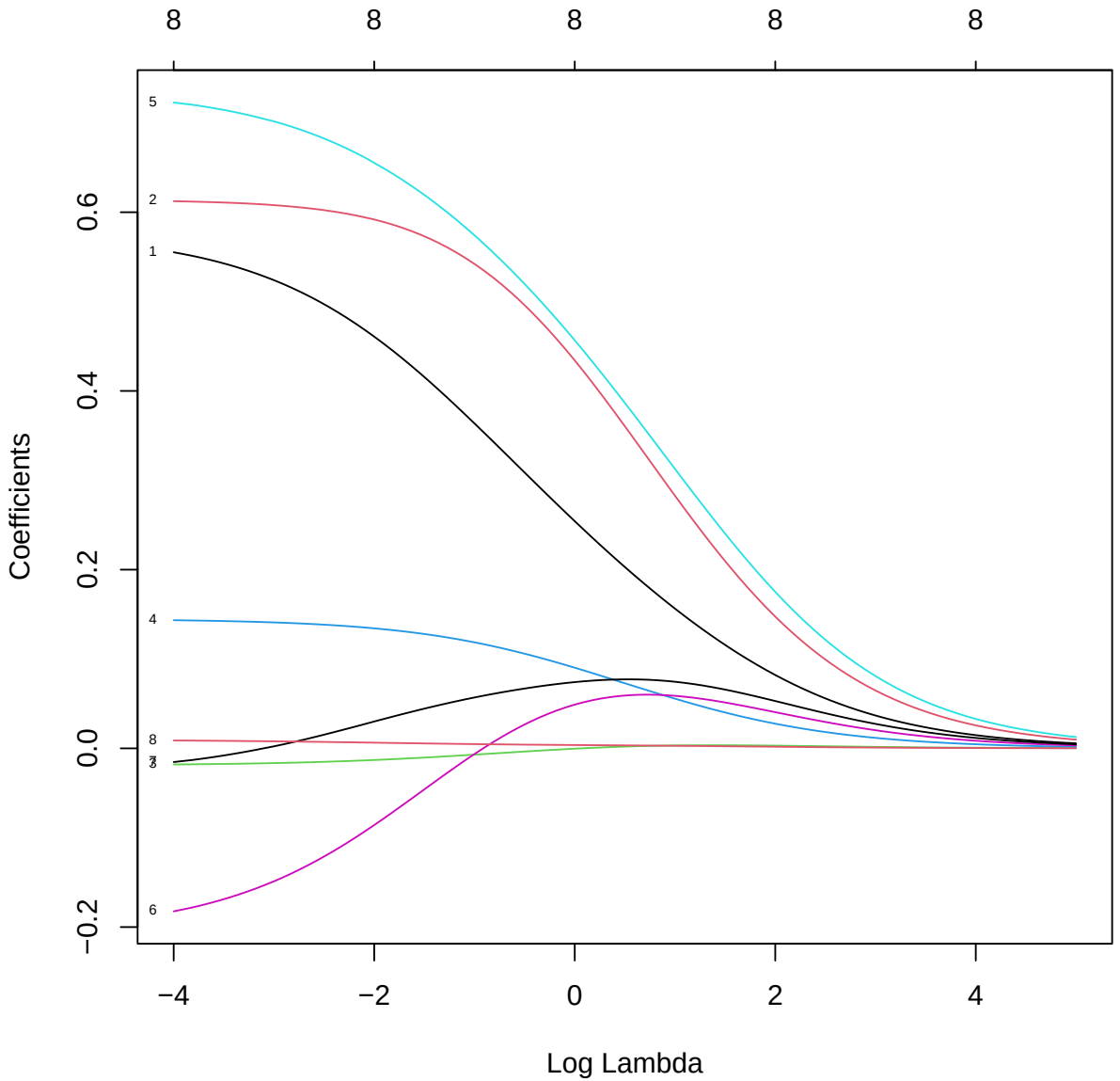
Example 2 – Prostate Cancer

Standardise the predictors and fit a saturated linear model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.45	0.09	28.18	0.00
lcavol	0.72	0.13	5.37	0.00
lweight	0.29	0.11	2.75	0.01
age	-0.14	0.10	-1.40	0.17
lbph	0.21	0.10	2.06	0.04
svi	0.31	0.13	2.47	0.02
lcp	-0.29	0.15	-1.87	0.07
gleason	-0.02	0.14	-0.15	0.88
pgg45	0.28	0.16	1.74	0.09

Example 2 – Prostate Cancer

Now apply L_2 regularisation using the `glmnet` package in R (see notes)

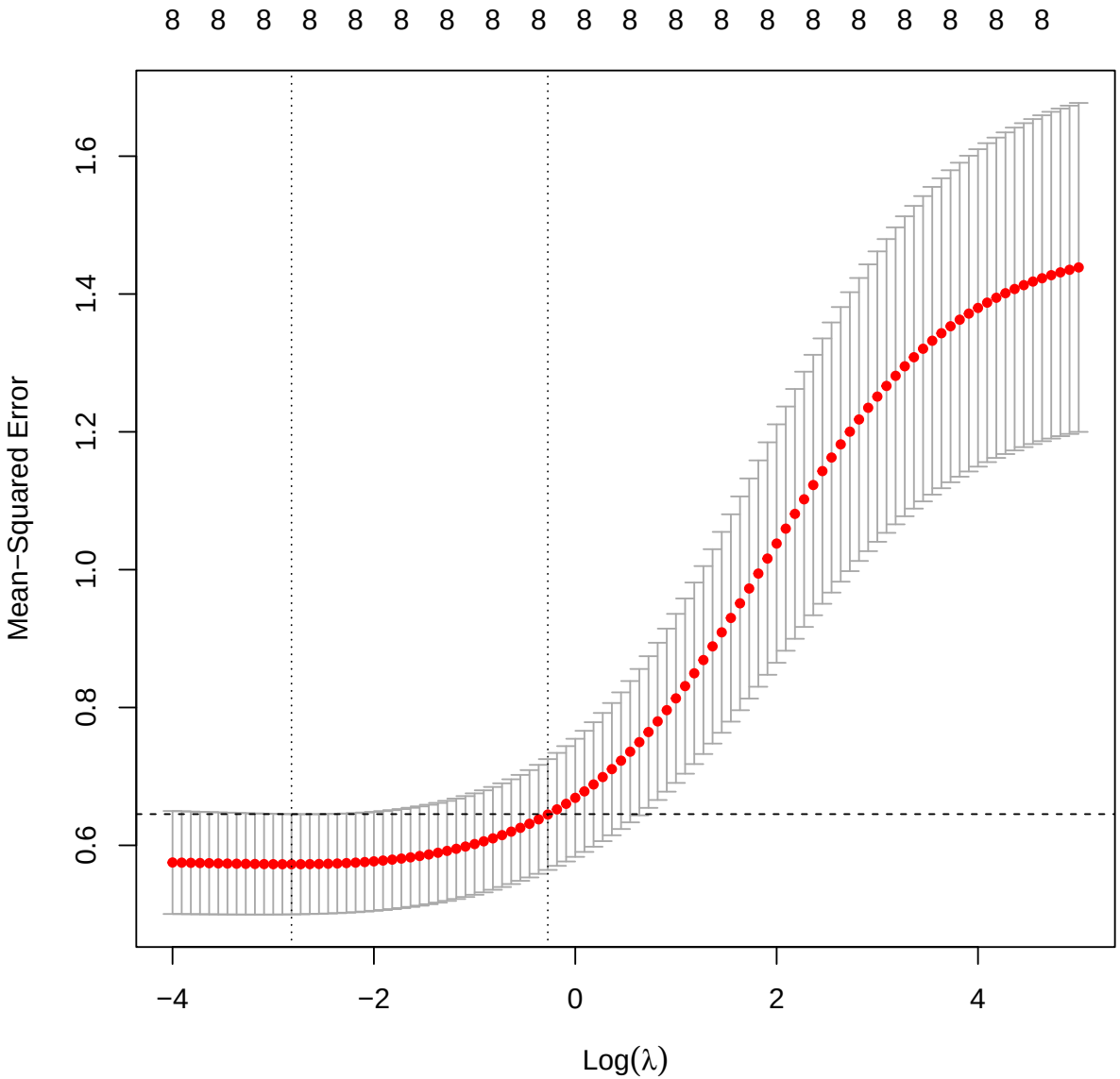


Example 2 – Prostate Cancer

- The initially negative coefficient for `lcp` (β_6) becomes both positive and more significant, relative to the other predictors
- Hence the notion of "coefficients shrinking towards zero" is a slight misnomer, or perhaps an oversimplification of the effect L_2 regularisation has
- However, as $\lambda \to \infty$ (or, equivalently, $\tau \to 0$), all coefficients will indeed be forced towards zero
- The ideal model will usually correspond to a level of λ that allows stronger predictors to be more prominent by diminishing the effect of their correlates

Example 2 – Prostate Cancer

How do we determine the appropriate level of λ? Through CV!



Example 2 – Prostate Cancer

- The best choice of λ is not always immediately apparent
- In the previous plot we saw that a more "reasonable" representation of the coefficients is achieved when $\log(\lambda)$ is closer to zero, rather than at the minimum CV MSE

	$\hat{\beta}_{\text{Min}}$	$\hat{\beta}_{+1\text{SE}}$
(Intercept)	0.173	-0.181
lcavol	0.515	0.284
lweight	0.606	0.470
age	-0.016	-0.002
lbph	0.140	0.099
svi	0.696	0.492
lcp	-0.140	0.038
gleason	0.007	0.071
pgg45	0.008	0.004

Example 2 – Prostate Cancer

- Although some predictors have **almost** been removed, these coefficients are still nonzero
- The ridge regression still includes all p predictors in the final model
- Consider now the lasso as an approach for variable selection

The Lasso – L_1 regularisation

The **least absolute shrinkage and selection operator** (lasso) is another form of regularisation that also minimises a penalised RSS. However, the constraint is slightly different:

$$\hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$
$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq \tau. \tag{12}$$

Or, written in its Lagrangian form:

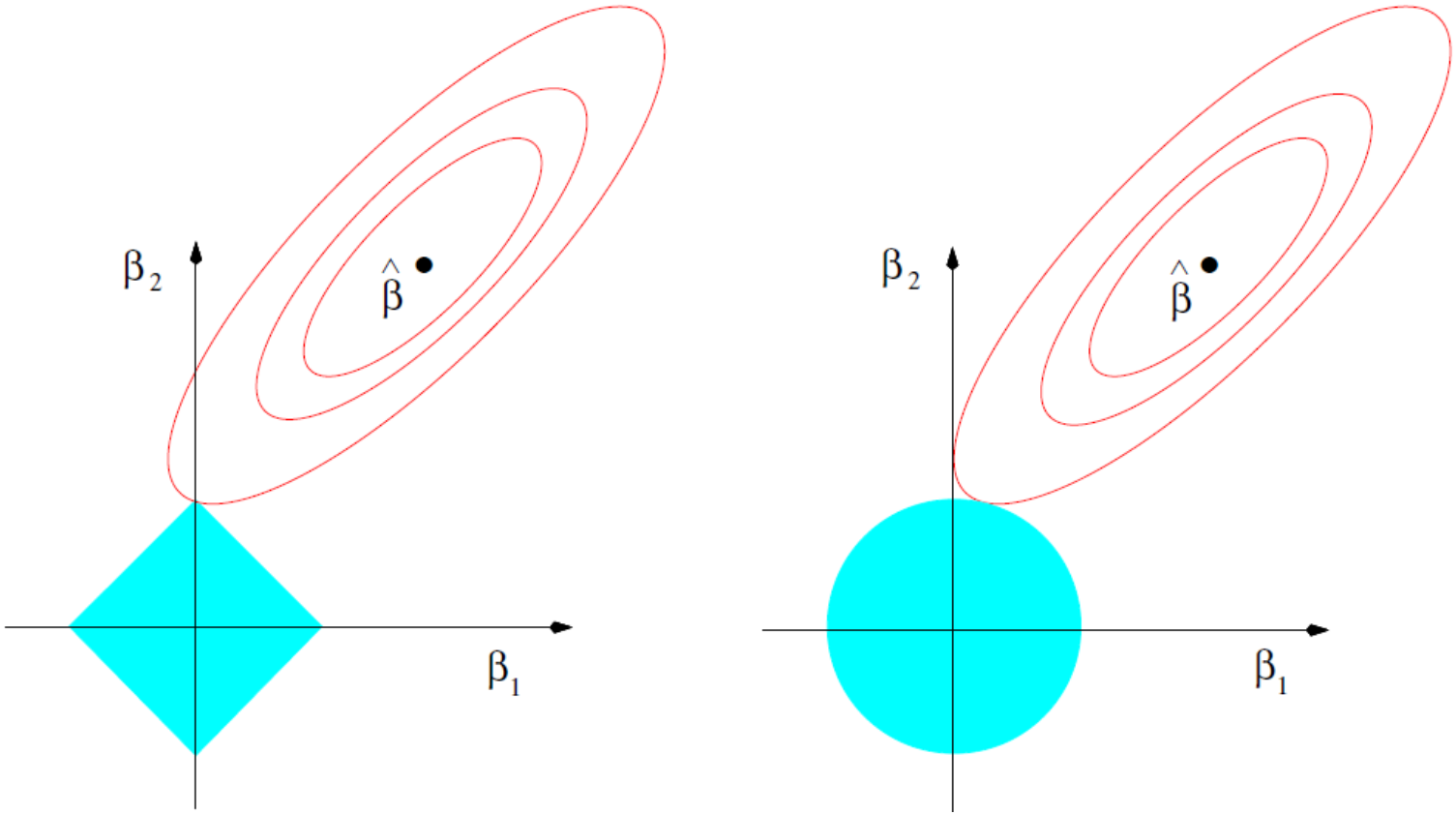
$$\hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{13}$$

The Lasso – L_1 regularisation

- Again the name L_1 **regularisation** arises from the fact that the penalty is based on the L_1 (Manhattan) norm $||\boldsymbol{\beta}||_1 = \sum_{i=1}^{p} |\beta_i|$
- This constraint on the regression parameters makes the solutions nonlinear in the y_i
- Therefore, there is no closed form expression for $\hat{\boldsymbol{\beta}}_L$ like there is for the ridge estimate (except when covariates are orthonormal)
- If we let $\tau > \sum_{j=1}^{p} |\hat{\beta}_j^{LS}|$, then the lasso estimates are exactly equal to the least squares estimates
- If, $\tau = \frac{1}{2} \sum_{j=1}^{p} |\hat{\beta}_j^{LS}|$, then the least squares coefficients are shrunk by 50% **on average**

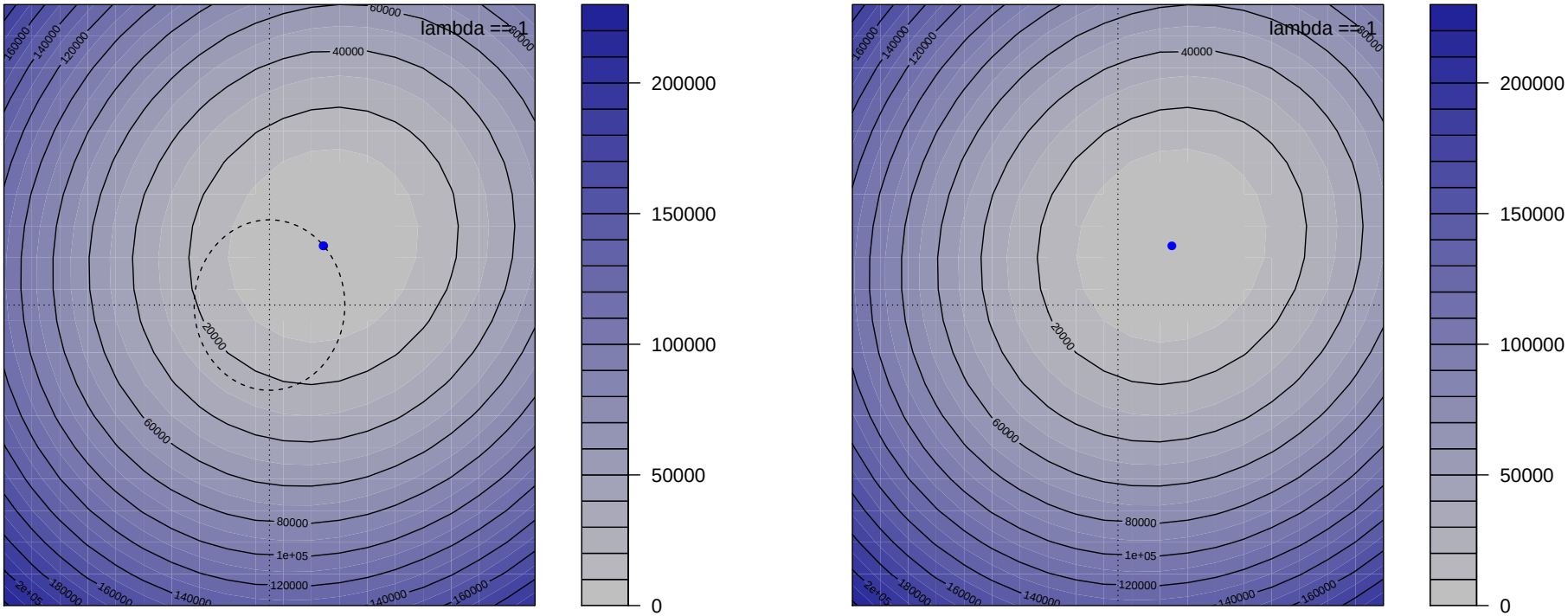
Comparing ridge and lasso

The nature of the constraints yield different trajectories for $\hat{\beta}$ as λ increases/τ decreases:



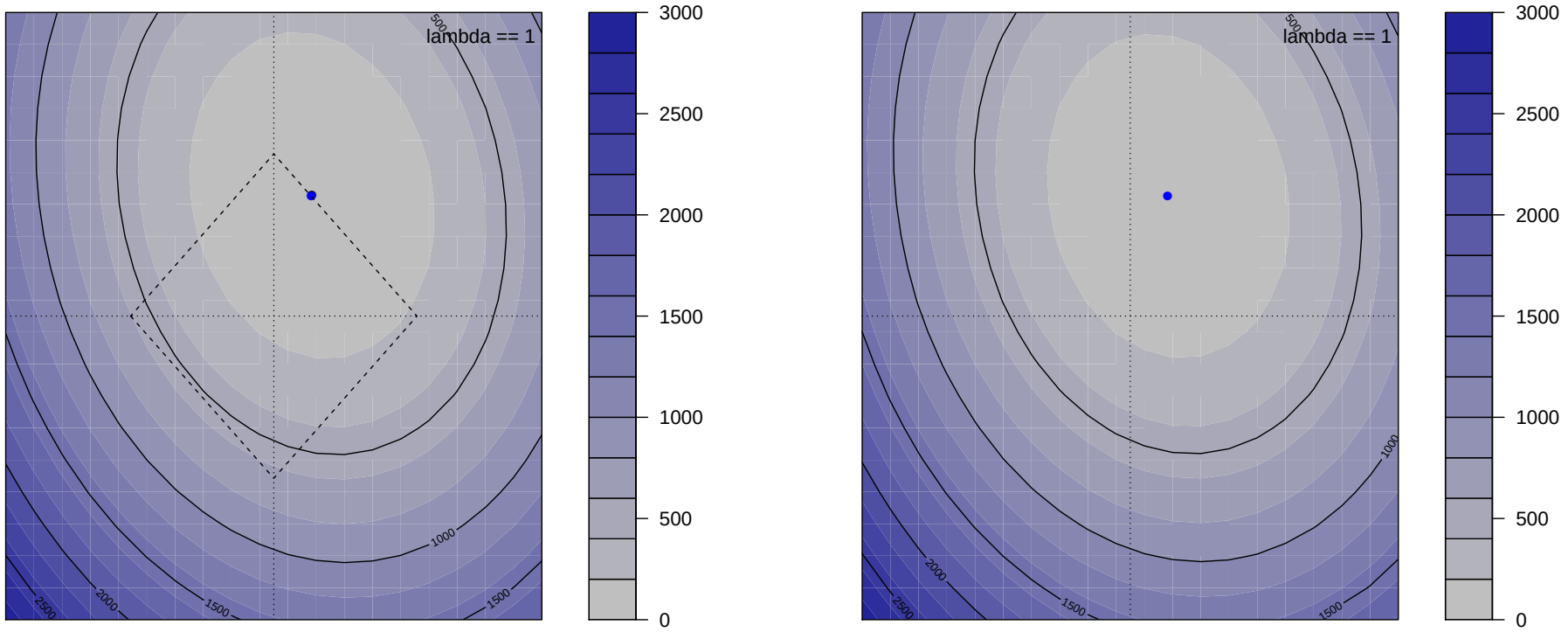
Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq \tau$ and $\beta_1^2 + \beta_2^2 \leq \tau$, while the red ellipses are the contours of the RSS

Ridge Regression - Objectives



Left: original objective s.t. constraint. Right: Penalised objective.

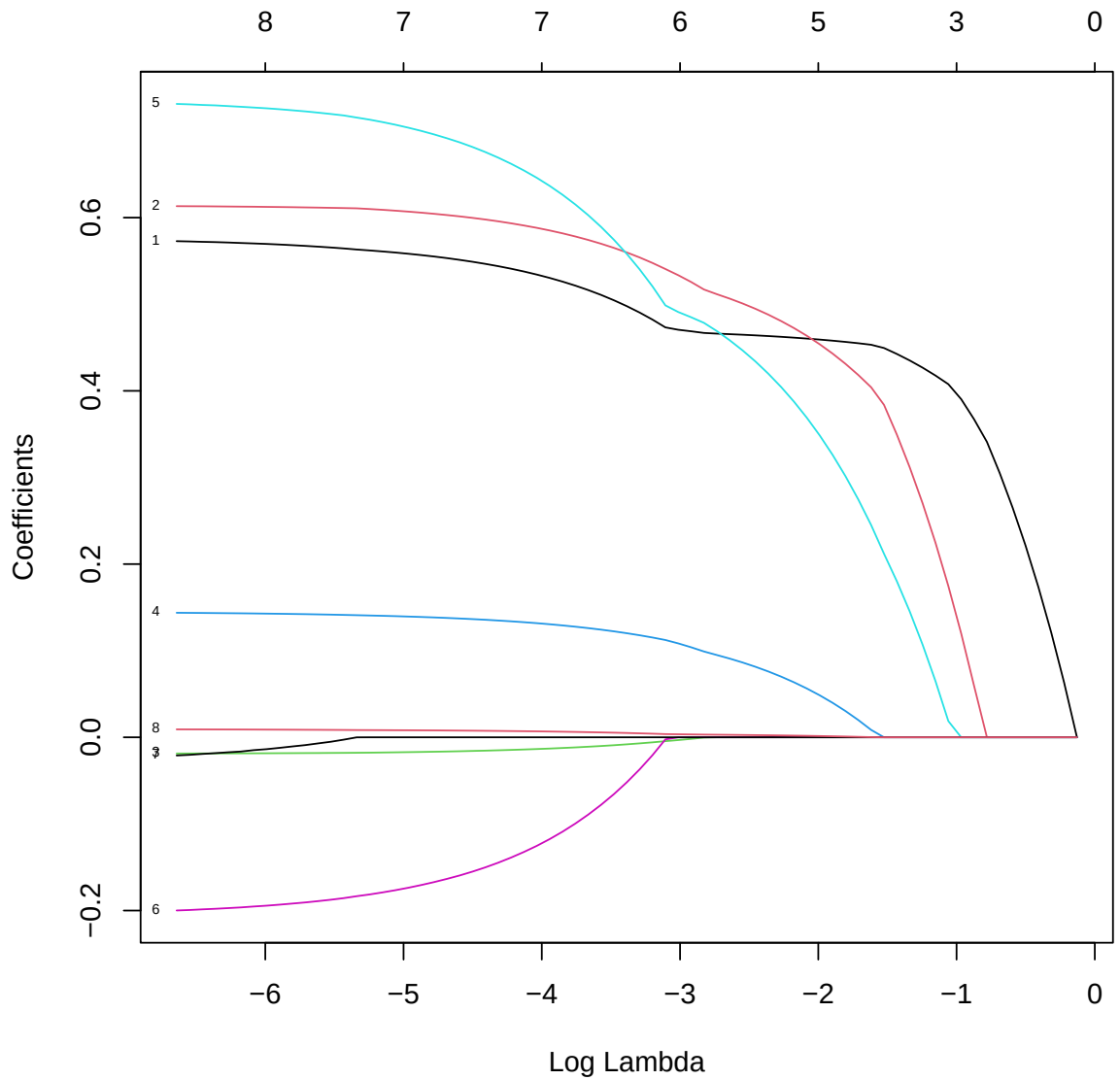
Ridge Regression - Objectives



Left: original objective s.t. constraint. Right: Penalised objective.

Example 2 – Prostate Cancer (continued)

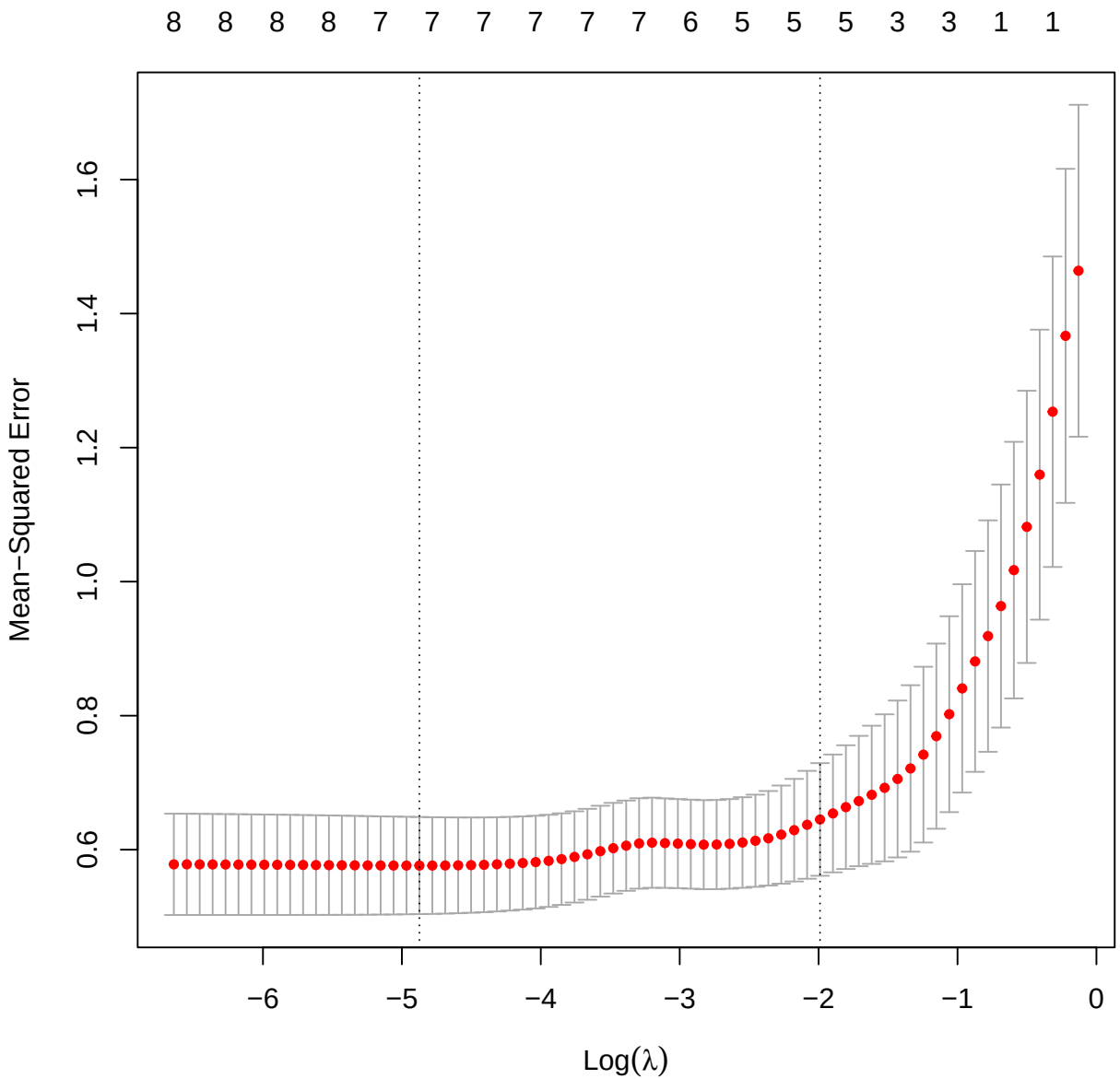
Applying L_1 regularisation via `glmnet`:



Example 2 – Prostate Cancer (continued)

- Here we see the coefficients shrinking and equaling zero as the regularisation penalty increases, as opposed to gradually decaying as in ridge regression
- This shows how the lasso performs **variable selection**
- To determine which variables should be (de)selected, we will again implement CV using the MSE as loss function

Example 2 – Prostate Cancer (continued)



Example 2 – Prostate Cancer (continued)

- If selecting the minimum MSE, the contradictory estimates will clearly still remain
- Selecting the penalty corresponding to the largest MSE within one standard error of the minimum MSE yields a notably simpler model, with three of coefficients shrunk to zero:

	$\hat{\boldsymbol{\beta}}_{+1\text{SE}}$
(Intercept)	0.083
lcavol	0.459
lweight	0.454
age	.
lbph	0.048
svi	0.349
lcp	.
gleason	.
pgg45	0.001

Example 2 – Prostate Cancer (continued)

- At this point, we have once again defined $\hat{f}$
- This is ultimately still a linear model with adjusted coefficient estimates
- Now use these models to determine $\hat{f}(\boldsymbol{x}_0)$ and compare

	OLS	Ridge	Lasso
Test MSE	0.549	0.509	0.454

Always interpret the value of the errors in consultation with the subject experts! (oncologist)

Example 2 – Prostate Cancer (continued)

- At this point, we have once again defined $\hat{f}$
- This is ultimately still a linear model with adjusted coefficient estimates
- Now use these models to determine $\hat{f}(\boldsymbol{x}_0)$ and compare

	OLS	**Ridge**	**Lasso**
Test MSE	0.549	0.509	0.454

Always interpret the value of the errors in consultation with the subject experts! (oncologist)

Elastic-net

Elastic-net

In the code/notes, you will have seen that an α parameter needed to be specified, the value of which determined whether we were applying L_1/L_2 regularisation.

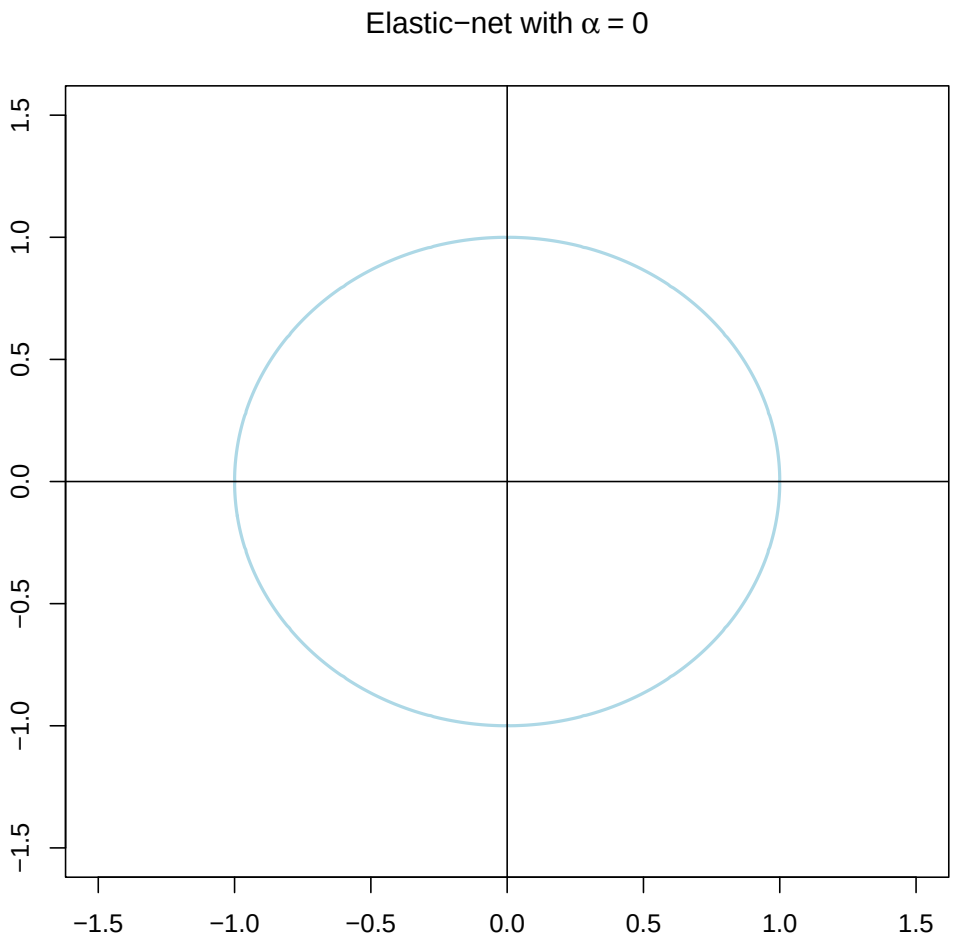
This is because both of these can be seen as special cases (the extremes) of the **elastic-net** penalty, a mixture of the two penalties discussed above:

$$\text{penalty} = \lambda \left[(1-\alpha) \left(\sum_{j=1}^{p} \beta_j^2 \right) + \alpha \left(\sum_{j=1}^{p} |\beta_j| \right) \right] \tag{14}$$

The elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge

Elastic-net

The constraint function for the elastic-net for different values of α:



Note that $\alpha = 0$ corresponds to the ridge penalty, and $\alpha = 1$ to the lasso

Elastic-net

Now there are two parameters to tune simultaneously, and the choice of α influences the range of λ values we should consider searching over.

To my knowledge there is no R package specifically for elastic-net on CRAN that allows one to automatically search over both tuning parameters, or hyperparameters and find the optimal combination. Therefore, there are three options:

1. Manually vary and loop across various α values, each time extracting the optimal λ and identifying the lowest overall CV MSE.
2. Use a non-CRAN package that has already written up exactly this procedure, for example glmnetUtils.
3. Use a wrapper function designed for hyperparameter gridsearches. `caret` is an excellent R package for this purpose, one we will use going forward in this course.

Homework!