

# STAT5003

Week 8 : Feature Selection

**Jaslene Lin**

*The University of Sydney*



THE UNIVERSITY OF  
**SYDNEY**

# Readings and R functions covered

## ! Important

- **Introduction to Statistical Learning**

- ⇒ Feature selection covered in Chapter 6.1-6.2

- **R** functions

- ⇒ `glmnet::glmnet` (Fit lasso and ridge regression models)

- ⇒ `leaps::regsubsets` (Feature selection through exhaustive searching)

- ⇒ `stats::model.matrix` (Needed to interface with glmnet)

This presentation is based on the [SOLES reveal.js Quarto template](#) and is licensed under a [Creative Commons Attribution 4.0 International License](#).

# Feature Selection - Searching

# Goals of feature selection

- **Prediction accuracy:** especially when  $p > n$ 
  - ⇒  $p$  is the number of features, and  $n$  is the number of observations
- **Model interpretability:**
  - ⇒ Removing irrelevant or poor features (that is, by setting the corresponding coefficient estimates to zero) → we can obtain a model that is more easily interpreted
- Some approaches for feature selection are presented

# Approaches for feature selection

## 1. Subset selection:

- Identify a subset of the  $p$  predictors that we believe to be related to the response or class  $y$ .
- Fit a **classification** or **regression model** on the reduced set of variables.

## 2. Shrinkage:

- Primarily used for **regression models**.
- Fit a model involving all  $p$  predictors.
- Some estimation coefficients are shrunk towards zero.
- This shrinkage (also known as regularisation) has the effect of reducing variance and can also be used for feature selection.

## 3. Dimension reduction:

- We project the  $p$  predictors into  $M$ -dimensional subspace,  $M < p$ .

# Best Subset Selection

Suppose we have a dataset with a response variable  $Y$  and **three predictors**  $(X_1, X_2, X_3)$

We want to use **Best Subset Selection** to find the best model for predicting  $Y$ .

## Step 0: Null Model

This model uses **no predictors**, only the intercept:

$$\mathcal{M}_0 : Y = \beta_0$$

## Step 1: One-Predictor Models ( $k = 1$ )

There are  $\binom{3}{1} = 3$  possible models:

$$\mathcal{M}_{1a} : Y \sim \beta_0 + X_1$$

$$\mathcal{M}_{1b} : Y \sim \beta_0 + X_2$$

$$\mathcal{M}_{1c} : Y \sim \beta_0 + X_3$$

# Best Subset Selection

## Step 2: Two Predictor Models ( $k = 2$ )

There are  $\binom{3}{2} = 3$  models:

- $\mathcal{M}_{2a} : Y \sim \beta_0 + X_1 + X_2$
- $\mathcal{M}_{2b} : Y \sim \beta_0 + X_1 + X_3$
- $\mathcal{M}_{2c} : Y \sim \beta_0 + X_2 + X_3$

Select the best one with the *smallest* RSS or equivalently the *largest*  $R^2$  and denote it  $\mathcal{M}_2$ .

## Step 3: Three Predictors ( $k = 3$ )

Only **one model** to consider:

- $\mathcal{M}_3 : Y \sim \beta_0 + X_1 + X_2 + X_3$

## Summary of Candidate Models

Subset Size ( $k$ )	Number of Models	Best Model ( $\mathcal{M}_k$ )
0	1	Null model ( $\mathcal{M}_0$ )
1	3	Best 1-variable model ( $\mathcal{M}_1$ )
2	3	Best 2-variable model ( $\mathcal{M}_2$ )
3	1	Full model ( $\mathcal{M}_3$ )

# Best Subset Selection

## Final Model Selection

Subset Size ( $k$ )	Number of Models	Best Model ( $\mathcal{M}_k$ )
0	1	Null model ( $\mathcal{M}_0$ )
1	3	Best 1-variable model ( $\mathcal{M}_1$ )
2	3	Best 2-variable model ( $\mathcal{M}_2$ )
3	1	Full model ( $\mathcal{M}_3$ )

Among these four candidate models

- Choose the best  $\mathcal{M}_k$  based on model selection criteria:
  - ⇒ Adjusted  $R^2$
  - ⇒ AIC / BIC
  - ⇒ Cross-validation error

# Best Subset Selection Methods

- It can be too computationally expensive to apply best subset selection when  $p$  is large.
  - ➡ Too many possible feature subsets.
- Statistical problems with large  $p$ .
- Larger search space → increased chance of finding models that overfit.
  - ➡ Perform well on training data.

# Forward Stepwise Selection

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
- In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.

# Forward stepwise selection

## Step 0: Null Model

$$\mathcal{M}_0 : Y = \beta_0$$

## Step 1

Try adding **one predictor at a time**:

- $\mathcal{M}_{1a} : Y \sim \beta_0 + X_1$
- $\mathcal{M}_{1b} : Y \sim \beta_0 + X_2$
- $\mathcal{M}_{1c} : Y \sim \beta_0 + X_3$

Choose the one with **best performance**  $\rightarrow \mathcal{M}_1$ .

## Step 2

Assuming

$$\mathcal{M}_1 = Y \sim \beta_0 + X_1$$

. Now, add a **second predictor** to the chosen  $\mathcal{M}_2$ :

- $\mathcal{M}_{2a} : Y \sim \beta_0 + X_1 + X_2$
- $\mathcal{M}_{2b} : Y \sim \beta_0 + X_1 + X_3$

# Forward Selection

## Step 3

Only one predictor left to add.

$$\mathcal{M}_3 : Y \sim \beta_0 + X_1 + X_2 + X_3$$

## Summary: Forward Selection

Step	Candidate Models	Selected Model
0	Null	$\mathcal{M}_0$
1	Add $X_j$	$\mathcal{M}_1$
2	Add next $X_j$	$\mathcal{M}_2$
3	Full model	$\mathcal{M}_3$

Among these four candidate models

- Choose the best  $\mathcal{M}_k$  based on model selection criteria:
  - ⇒ Adjusted  $R^2$
  - ⇒ AIC / BIC
  - ⇒ Cross-validation error
- Computational advantage over best subset selection is clear
- However
  - ⇒ **Not guaranteed to find the best possible model out of all  $2^p$  models** containing subsets of the  $p$  predictors (why?)

# Backward Stepwise Selection

- Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection.
- However, unlike forward stepwise selection:
  - ➡ begins with the **full model** containing all  $p$  predictors.
  - ➡ iteratively **removes** the **least useful** predictor, one-at-a-time.

# Backward stepwise selection

1. Denote  $\mathcal{M}_p$  to be the **full** model (e.g.  $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$  in linear regression)
  - Contains all predictors
2. For  $k = p, p - 1, \dots, 1$ 
  - Consider all  $k$  models that contain all but one of the predictors in  $\mathcal{M}_k$ , for a total of  $k - 1$  predictors
  - Choose the *best* among these  $k$  models and assign it as  $\mathcal{M}_{k-1}$ 
    - ⇒ Best measured against some metric (RSS or classification error)
3. Select the single best model among the  $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$ 
  - Using cross-validated prediction error,  $C_p$ , BIC, or adjusted  $R^2$  (described later)

# More on Backward Stepwise Selection

- Similarities to forward selection
  - ➡ it searches through only  $1 + p(p + 1)/2$  models (including the null model  $\mathcal{M}_0$ )
  - ➡ can be applied in settings where  $p$  is too large to apply best subset selection
  - ➡ backward selection is **not guaranteed to yield the best model containing a subset of the  $p$  predictors**
- **Note:** for some models such as linear regression, backward selection requires that the number of cases  $n$  is larger than the number of features  $p$  (so that the full model can be fit)
  - ➡ In contrast, forward stepwise can be used even when  $n < p$

# Using R

The `swiss` dataset is a built-in dataset in R, providing standardized fertility measures and socio-economic indicators for 47 French-speaking provinces of Switzerland around 1888.

## ► Code

```
Rows: 47
Columns: 6
$ Fertility      <dbl> 80.2, 83.1, 92.5, 85.8, 76.9, 76.1, 83.8, 92.4, 82.4,...
$ Agriculture    <dbl> 17.0, 45.1, 39.7, 36.5, 43.5, 35.3, 70.2, 67.8, 53.3,...
$ Examination    <int> 15, 6, 5, 12, 17, 9, 16, 14, 12, 16, 14, 21, 14, 19, ...
$ Education      <int> 12, 9, 5, 7, 15, 7, 7, 8, 7, 13, 6, 12, 7, 12, 5, 2, ...
$ Catholic       <dbl> 9.96, 84.84, 93.40, 33.77, 5.16, 90.57, 92.85, 97.16,...
$ Infant.Mortality <dbl> 22.2, 22.2, 20.2, 20.3, 20.6, 26.6, 23.6, 24.9, 21.0,...
```

## ► Code

```
Subset selection object
5 Variables (and intercept)
      Forced in Forced out
Agriculture      FALSE      FALSE
Examination      FALSE      FALSE
Education        FALSE      FALSE
Catholic         FALSE      FALSE
Infant.Mortality FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
      Agriculture Examination Education Catholic Infant.Mortality
1 ( 1 ) " " " " " * " " " "
2 ( 1 ) " " " " " * " " * "
3 ( 1 ) " " " " " * " " * "
4 ( 1 ) " * " " " " * " " * "
5 ( 1 ) " * " " * " " * " " * "
```

## ► Code

```
[1] 35.204895 18.486158 8.178162 5.032800 6.000000
```

## ► Code

```
[1] 4015.236 3054.169 2422.245 2158.069 2105.043
```

# Direct vs Indirect methods

# Linear model (feature) selection

- Recall the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

How to choose the optimal model?

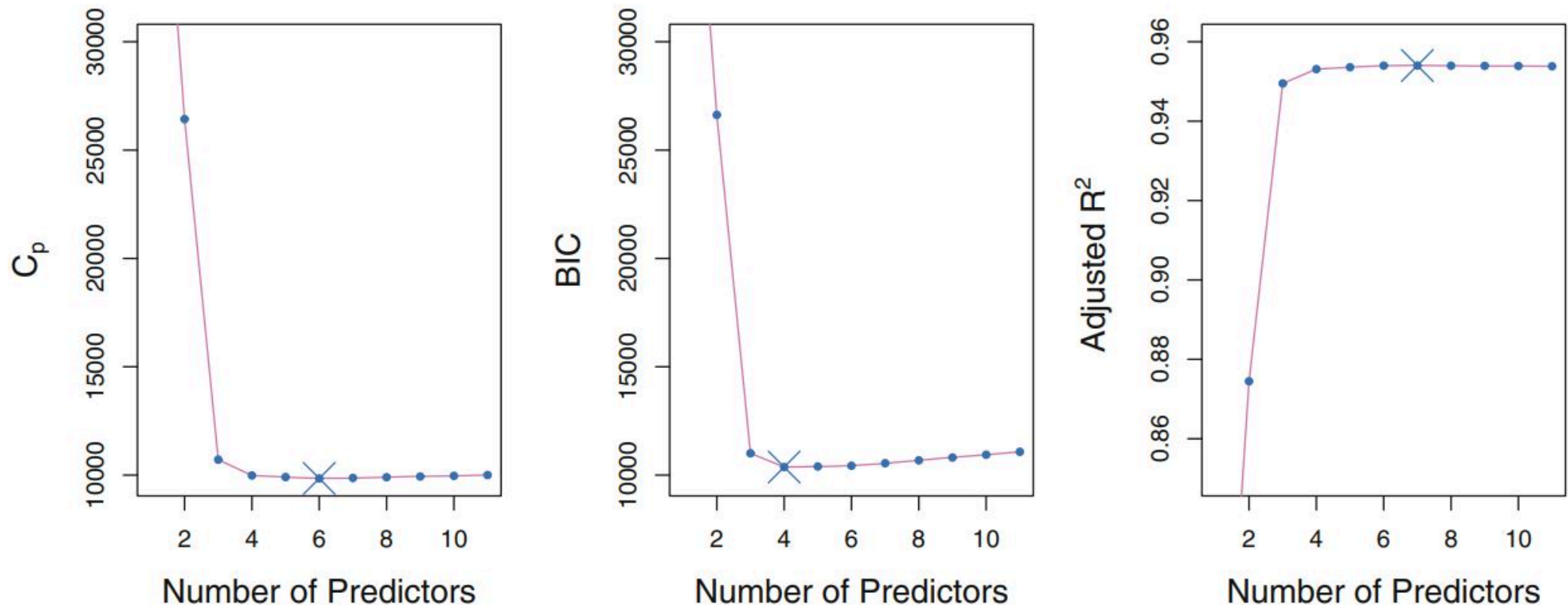
- The model containing all of the predictors will always have the smallest RSS, since these quantities are related to the training error.
- We wish to choose a model with low test error, not a model with low training error.
  - ⇒ Training error is usually a poor estimate of test error.
- Therefore, RSS are not suitable for selecting the best model among a collection of models with different number of predictors.

# Estimating test error - two approaches

- **Indirectly** estimate test error by making an adjustment to the training error.
  - ➡ Account for the bias due to overfitting.
- **Directly** estimate the test error, using either a test set or cross-validation set approach (covered in Week 6).
- Will illustrate the **indirect** approach and also review cross-validation.

# Indirect approaches (e.g. $C_p$ , BIC, adjusted $R^2$ )

- Adjust the training error for the model size (model complexity)
  - ➡ Can be used to select among a set of models with a different number of features
- Figure below displays Mallows's  $C_p$ , the Bayesian Information Criterion (BIC), and adjusted  $R^2$  for the best model produced by best subset selection on the credit data set



# Details of $C_p$ , BIC, adjusted $R^2$ (for standard regression)

- Mallows's  $C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2)$ 
  - ⇒  $n$ : number of observations;
  - ⇒  $d$ : total number of features (not including the intercept)
  - ⇒  $\hat{\sigma}^2$ : an estimate of the variance of  $\varepsilon$
- Bayesian Information Criterion  $\text{BIC} = \frac{1}{n}(\text{RSS} + \log(n)d\hat{\sigma}^2)$
- Like  $C_p$ , the BIC will tend to take on small value for model with a low test error, and so generally we select model that has the lowest BIC value
- Notice that BIC replaces the  $2d\hat{\sigma}^2$  used by  $C_p$  with a  $\log(n)d\hat{\sigma}^2$  term
- Since  $\log 8 > 2$  when  $n \geq 8$ , the BIC statistic typically places a heavier penalty on models with many variables, and hence results in the selection of smaller models than  $C_p$
- $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$  and  $\text{Adj } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$ 
  - ⇒ TSS: total sum of squares

# Direct approach via Validation or Cross-Validation

- We are given a sequence of models  $\mathcal{M}_k$  indexed by model size  $k = 0, 1, 2, \dots, p$ , where each model uses a different number of predictors.
- The goal is to **select the model**  $\mathcal{M}_k$  that yields the lowest test error.

## The Validation Set Approach

- Split the data into training and validation sets (roughly of equal size).
- Train each  $\mathcal{M}_k$  on the training set and then test on the validation set, and record the test error (MSE or equivalent)
- Select  $k$  for which the resulting estimated test error is the smallest.

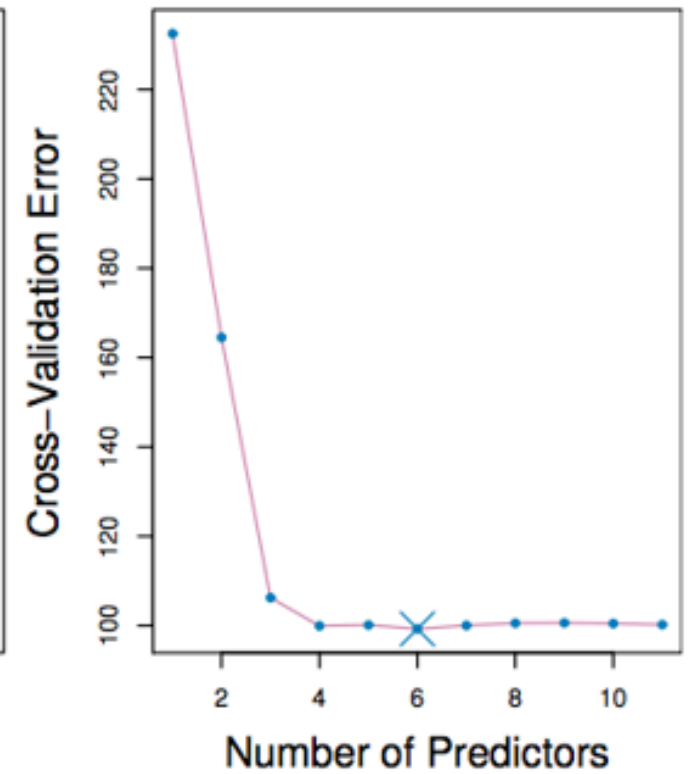
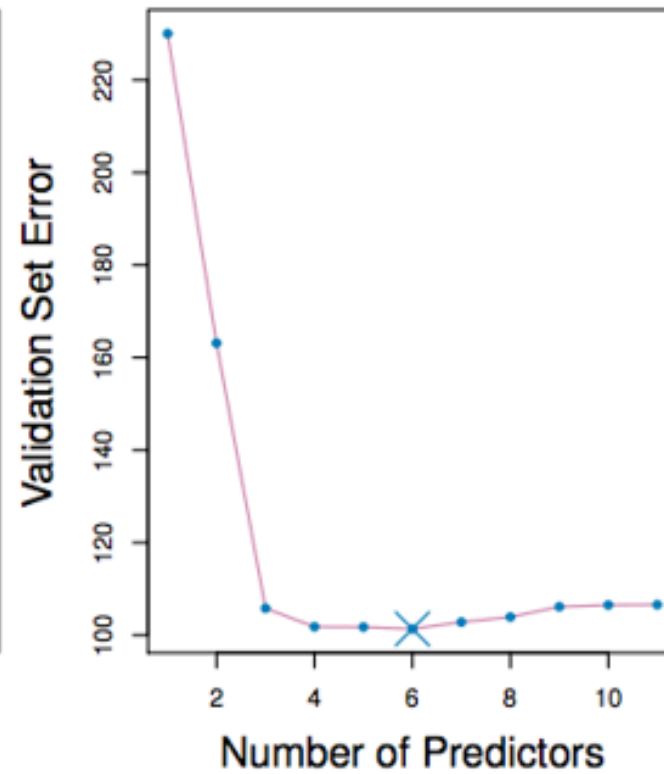
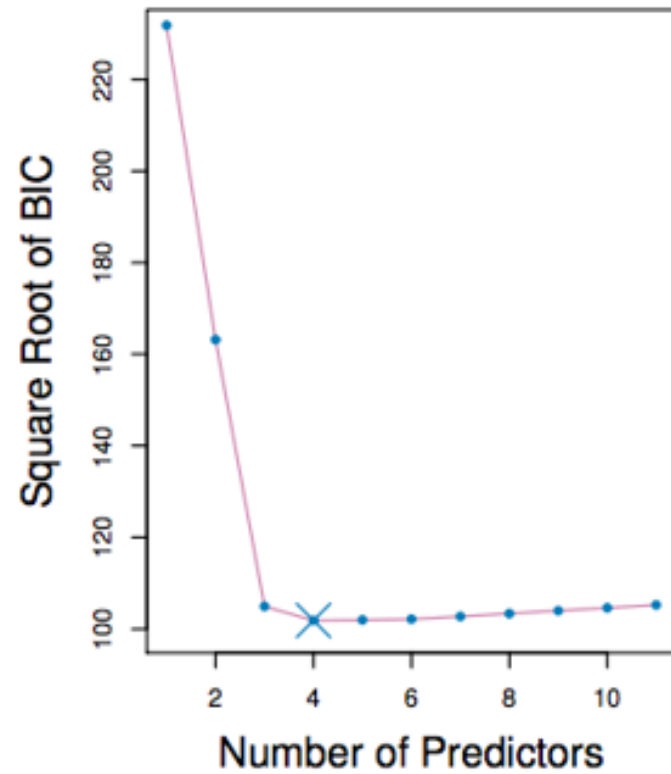
### Drawback:

1. The validation estimate of the test error rate can be highly variable, depending on the random split of the training and the validation set.
2. The validation set error may overestimate the test error as it is trained on a much smaller training set.

# Direct Approach Summary

- These approaches have an advantage relative to  $C_p$  and BIC, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance  $\sigma^2$ .
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance  $\sigma^2$ .

## Credit card example



# Shrinkage methods

- The following two methods specifically designed for **linear regression**: Ridge regression and Lasso.
- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all  $p$  predictors using a technique that constrains or regularizes the coefficient estimates.
- Shrinking the coefficient estimates can significantly **reduce their variance**.

# Ridge regression

- Recall that the least squares fitting procedure estimates  $\beta_0, \beta_1, \dots, \beta_p$  using the values that minimize

$$\text{RSS} = \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 = \|\mathbf{Y} - \mathbf{1}_n \beta_0 - \mathbf{X} \boldsymbol{\beta}\|_2^2$$

⇒  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$

⇒  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$

⇒  $\mathbf{X}$  is an  $n \times p$  matrix with  $(i, j)$ th entry  $X_{ij}$

- The ridge regression coefficient estimates  $\hat{\beta}_R$  are the values that minimize

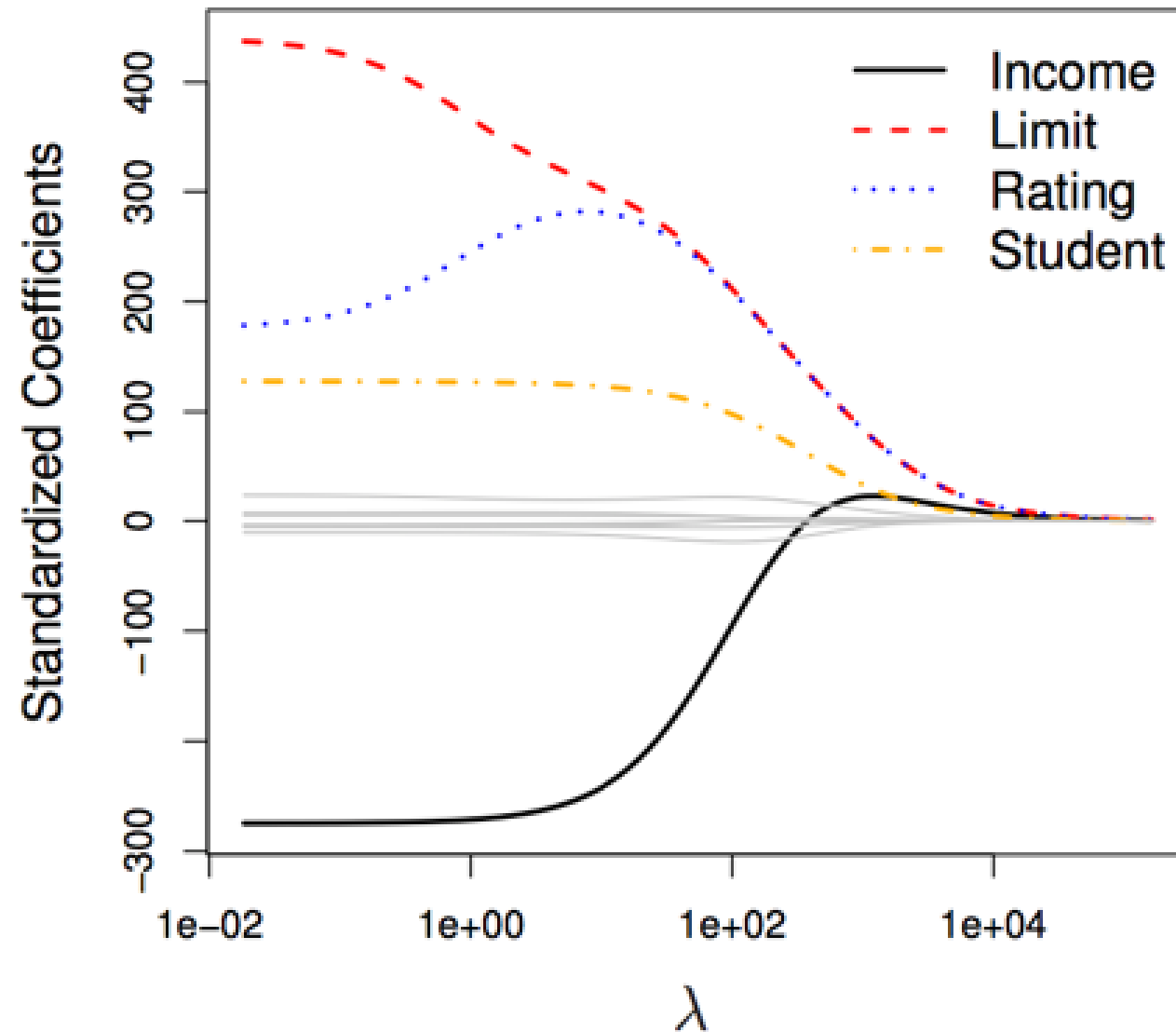
$$\text{RSS} + \lambda \sum_{j=1}^p \beta_j^2$$

⇒  $\lambda \geq 0$  is a **tuning** parameter, to be determined separately

# Ridge regression

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
- However, the second term,  $\lambda \sum_{j=1}^p \beta_j^2$ , is called a shrinkage penalty:
  - is small when  $\beta_1, \dots, \beta_p$  are close to zero, and so it has the effect of shrinking the estimates of  $\beta_j$  towards zero.
- The tuning parameter  $\lambda$  serves to control the relative impact of these two terms on the regression coefficient estimates.
- Selecting a good value for  $\lambda$  is critical; cross-validation is used for this.

## Credit card example



# Ridge regression: scaling of predictors

- The standard least squares coefficient estimates are scale **invariant**:
  - ➡ Multiplying  $X_j$  by a constant  $c$  simply leads to a scaling of the least squares coefficient estimates by a factor of  $1/c$ .
  - ➡ Regardless of how the  $j^{\text{th}}$  predictor  $X_j$  is scaled,  $X_j \hat{\beta}_j$  will remain the same.
- In contrast, the ridge regression coefficients estimates can change **substantially** when multiplying a given predictor by a constant:
  - ➡ Due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after **standardising the predictors**, using a formula such as below:

$$\tilde{X}_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}}.$$

- **Disadvantage:**
  - ➡ Ridge regression will include all  $p$  predictors in the final model
  - ➡ Subset selection will generally select models that involve a subset of the predictors

# The Lasso

- Refers to **l**east **a**bsolute **s**hrinkage and **s**election **o**perator
- The Lasso is a relatively recent alternative to ridge regression that overcomes the disadvantage of retaining all the predictors in the model
  - ➡ The lasso coefficients  $\hat{\beta}_L$  minimise the quantity

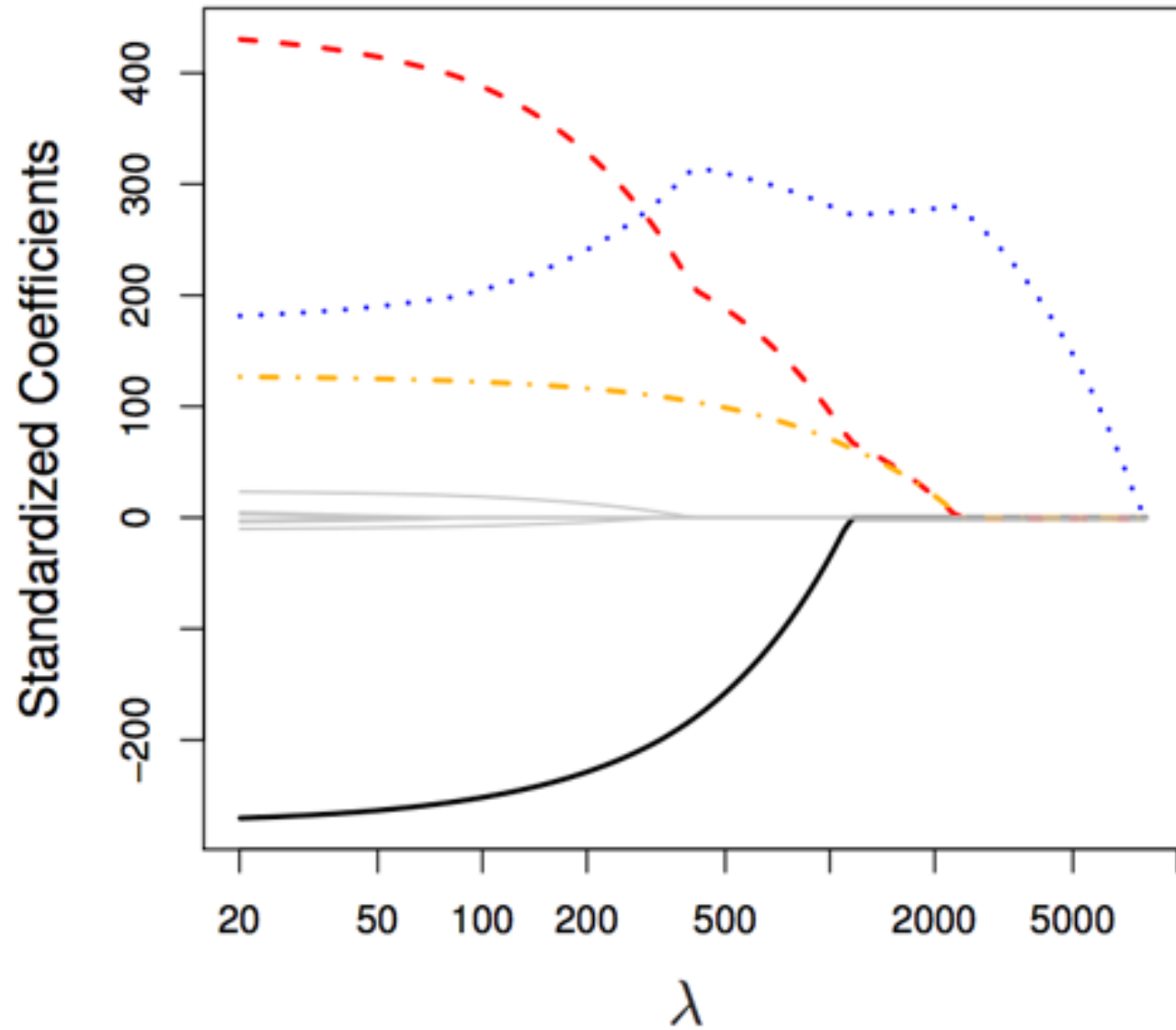
$$\text{RSS} + \lambda \sum_{j=1}^p |\beta_j|$$

- The Lasso uses an  $\ell_1$  penalty instead of the  $\ell_2$  penalty
  - ➡ The  $\ell_1$  norm of a coefficient vector  $\beta$  is  $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$
  - ➡ Ridge regression uses the  $\ell_2$  norm, i.e.,  $\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2$

# The Lasso

- As with ridge regression, the Lasso shrinks the coefficient estimates towards zero.
- However, in the case of the Lasso, the  $\ell_1$  penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter  $\lambda$  is sufficiently large.
- Hence, much like best subset selection, the lasso performs **feature selection** (in an embedded manner).
- We say that the Lasso yields **sparse** models, i.e., models that involve only a subset of variables.
- As in ridge regression, selecting a good value of  $\lambda$  for the Lasso is critical.
  - ⇒ cross-validation is again the method of choice.

## Example: Credit dataset

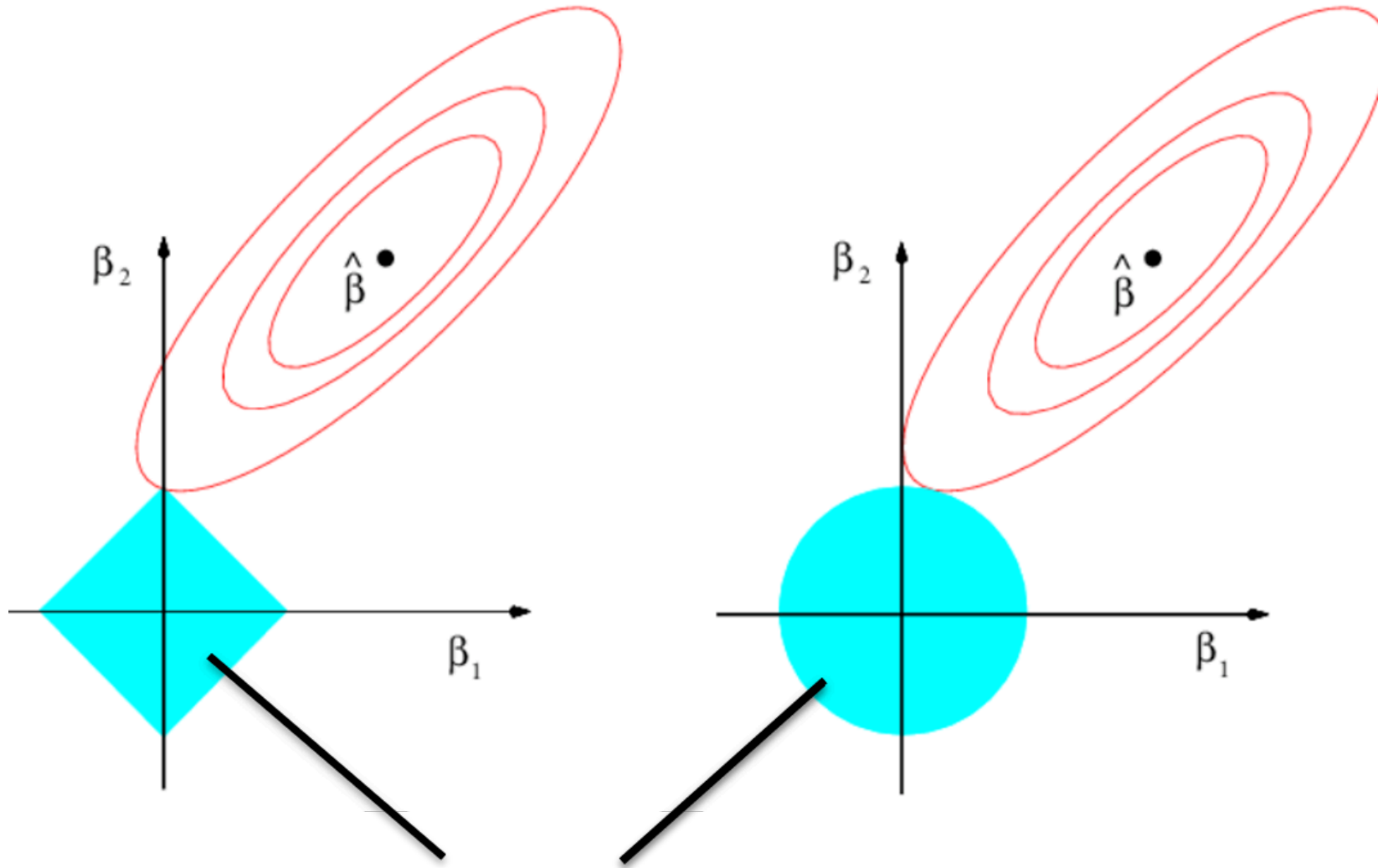


# Variable selection property of the Lasso

- Why is it that the Lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?
- One can show that the Lasso and ridge regression coefficient estimates solve the problems
  - ⇒  $s$  is determined by  $\lambda$  and can be different for ridge and Lasso even under the same  $\lambda$

$$\begin{array}{ll} \min_{\beta} \sum_{i=1}^n (Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij})^2 & \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq s \\ \min_{\beta} \sum_{i=1}^n (Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij})^2 & \text{subject to} \quad \sum_{j=1}^p \beta_j^2 \leq s \end{array}$$

## Comparison of $\ell_1$ and $\ell_2$ constraints

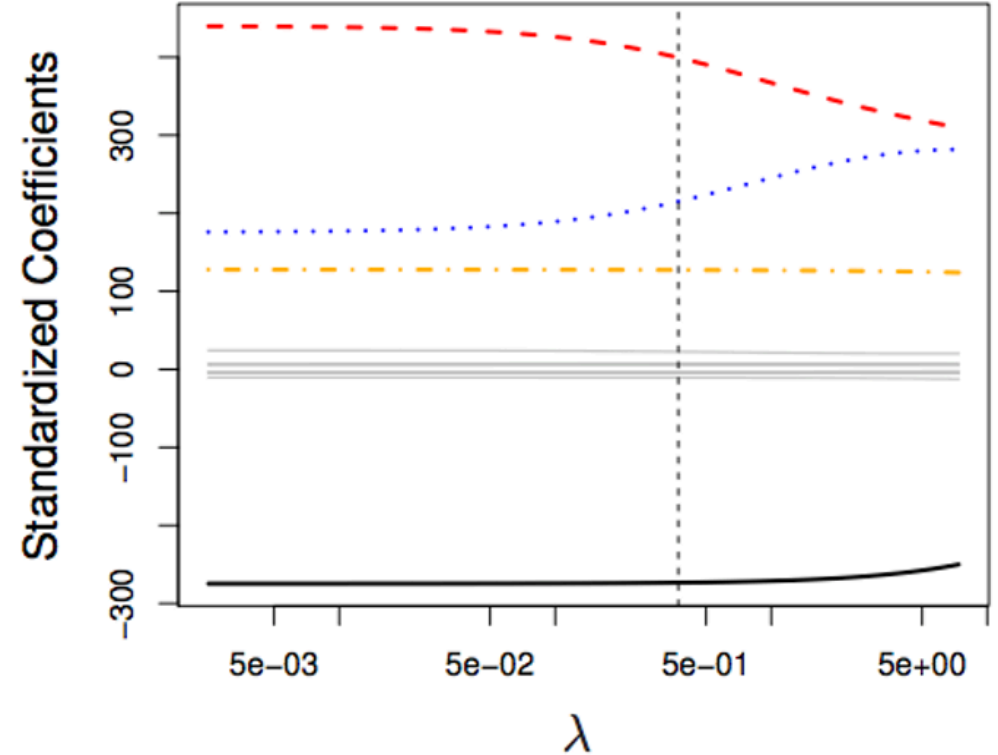
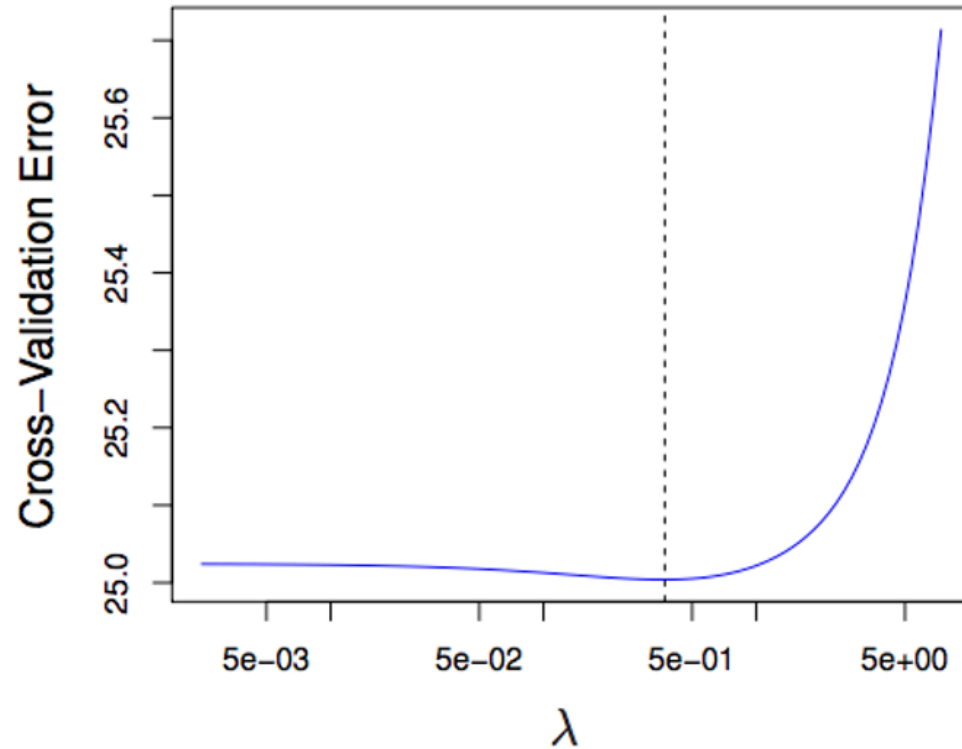


- Solution is feasible if it is within these blue regions for the Lasso (left) and Ridge (right) respectively.

# Selecting the tuning parameter

- As for subset selection, for ridge regression and Lasso, we require a method to determine which of the models under consideration is the best.
- That is, we require a method selecting a value for the tuning parameter  $\lambda$  or equivalently, the value of the constraint  $s$ .
- **Cross-validation** provides a simple way to tackle this problem. We choose a grid of  $\lambda$  values, and compute the cross-validation error rate for each value of  $\lambda$ .
- We then select the tuning parameter value for which the cross-validation error is the smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

# Credit data example



- Left illustrates cross-validation errors that result from applying ridge regression to the Credit data set with a range of  $\lambda$  values.
- Right illustrate the coefficient estimates as a function of  $\lambda$ . The vertical dashed lines indicate the best value of  $\lambda$  selected by cross-validation.

## Elastic net in glmnet

- In glmnet the implementation is actually for a more general model called the Elastic net
- It solves the following penalised minimization problem

$$\arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left( \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right)$$

- Can consider it a weighted combination (mixture) of  $\ell_1$  and  $\ell_2$  penalties
- Elastic net is appropriate when the variables form groups that contain highly correlated independent variables.