STAT5003 – Week 4

Still need to cover:

- Exhaustive searches in R
- Regularisation in R

Deadline: 11/05/23

Introduction to Feature and Model Selection

Model Selection

Motivating Questions

- How to cope with an ever-increasing number of observed variables?
- How to practically select the **most useful** features from a set of candidate features?

We want to identify a parsimonious model, which is a model that gives us still a adequate predictive ability.

Goals of Feature Selection

- **Prediction accuracy**: especially when $p > n$
  - Where p is the number of features and n denotes number of observations
- **Model interpretability**
  - Removing irrelevant or poor features (that is, by setting the corresponding coefficient estimates to zero) $\rightsquigarrow$ we can obtain a model that is more easily interpreted

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**
   - It is primarily used for regression models.
   - Fit a model involving all p predictors.
   - Some 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, $M < p$.

We project the data to its smallest space, whilst still trying to preseve all the information.

Linear Model (Feature) Selection

- Recall the linear model:

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

  - 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.
- RSS is not suitable for selecting the best model among a collection of models with different number of predictors

Estimating Test Error: Two Approaches

1. **Indirectly** estimate test error by making an adjustment to the training error.
    - Account for the bias due to overfitting.
2. **Directly** estimate the test error, using either a test set or cross-validation approach.

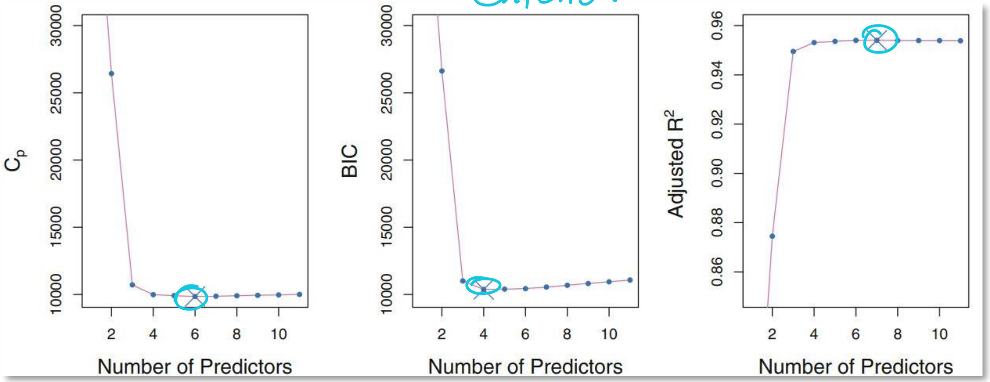
Indirect Approaches

- 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
- Mallow's C_p and the Bayesian information criterion (BIC) and adjusted R^2 for the best model produced by best subset selection on the credit data set.

Mallow's cp

Bayesian Information Criterion

Adjusted R²



Pick the lowest

Pick the lowest

Pick the highest.

Indirect: Mallows C_p and BIC

Recall MATH2831

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

$$BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$$

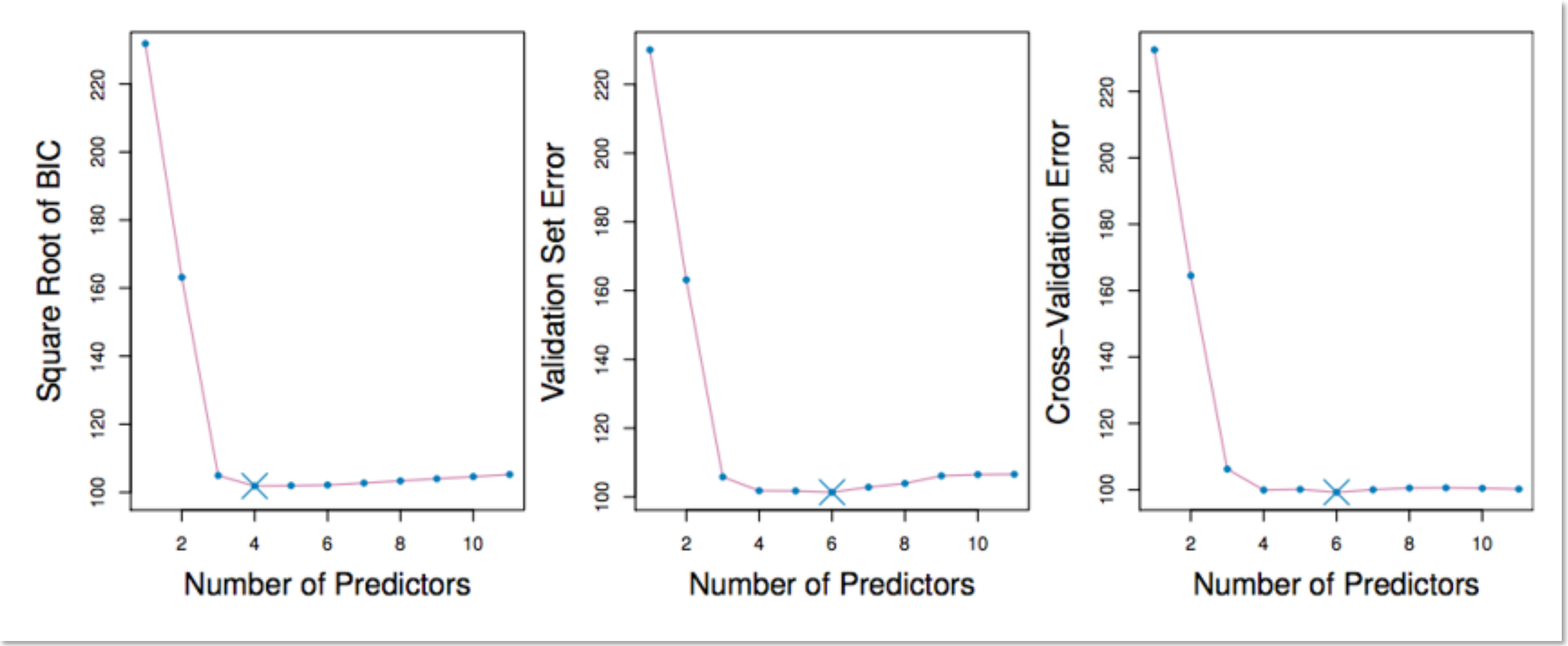
- Mallow's $\boxed{C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)}$
  - d is the total number of predictors
  - $\hat{\sigma}^2$ is an estimate of the variance of ε
- Bayesian information criterion: $\boxed{\text{BIC} = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)}$
- Like C_p, the BIC will tend to take on small 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, where n is the number of samples
- Since $\log n > 2$ when $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p

Direct: Test Set and Cross-Validation

- Each of the procedures returns a sequence of models $\mathcal{M}_k$ indexed by model size $k = 0{,}1{,}2, \dots$. Our job here is to select k. Once selected, we will return model $\mathcal{M}_k$.
- We compute the validation set error or the cross-validation error for each model $\mathcal{M}_k$.
    - Select the k for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to C_p and BIC, in that it provides direct estimate of the test error and doesn't require an estimate of the error variance σ^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 σ^2.

Credit Card Example

Square Root of BIC   Validation Set Error   Cross Validation Error



THE UNIVERSITY OF
SYDNEY

Recall MATH2831

Stepwise Methods

Stepwise Variable Selection Process

1. Start with **some** model, typically null model (with no explanatory variables) or full model (with all variables).
2. For each variable in the current model, **investigate** effect of **removing** it.
3. **Remove** the least informative variable, unless this variable is nonetheless supplying significant information about the response.
4. For each variable not in the current model, **investigate** effect of **including** it.
5. **Include** the most statistically significant variable not currently in model (unless no significant variable exists).
6. Go to step 2. Stop only if no change in steps 2–5.

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 (cont.)

1. Denote $\mathcal{M}_0$ to be the null model (e.g., $Y = \beta_0 + \varepsilon$ in linear regression)
   - Contains no predictors
2. For $k = 0,1,2, \dots, p-1$
   - Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor
   - Choose the *best* among these $p-k$ 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 Cp, BIC, or cross-validated prediction error, etc.

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

Forward
Null model
↓
Full model

Backward
Full model
↓
Null model

Backward Stepwise Selection (cont.)

1. Denote $\mathcal{M}_p$ to be the **full** model (e.g., $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ in linear reg)
   - Contains all predictors
2. For $k = p, p-1, \ldots, 1$
   - 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 $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$
   - Using Cp, BIC, or cross-validated prediction error, etc.

Criticism of Stepwise Procedures

- *“Stepwise regression is probably the most abused computerized statistical technique ever devised. If you think you need stepwise regression to solve a particular problem you have, it is almost certain you do not. Professional statisticians rarely use automated stepwise regression.”*

  —Wilkinson (1998)

- See Wiegand (2010) for a more recent simulation study and review of the performance of stepwise procedures.

Considerations

- Stepwise methods only look at searches through only $\boxed{1 + p(p+1)/2}$ models so they have a computational advantage over exhaustively searching the model space.
- However, they are not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.
- 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 $n < p$.

THE UNIVERSITY OF
SYDNEY

Exhaustive Searches

Exhaustive Search

2^p models for large p

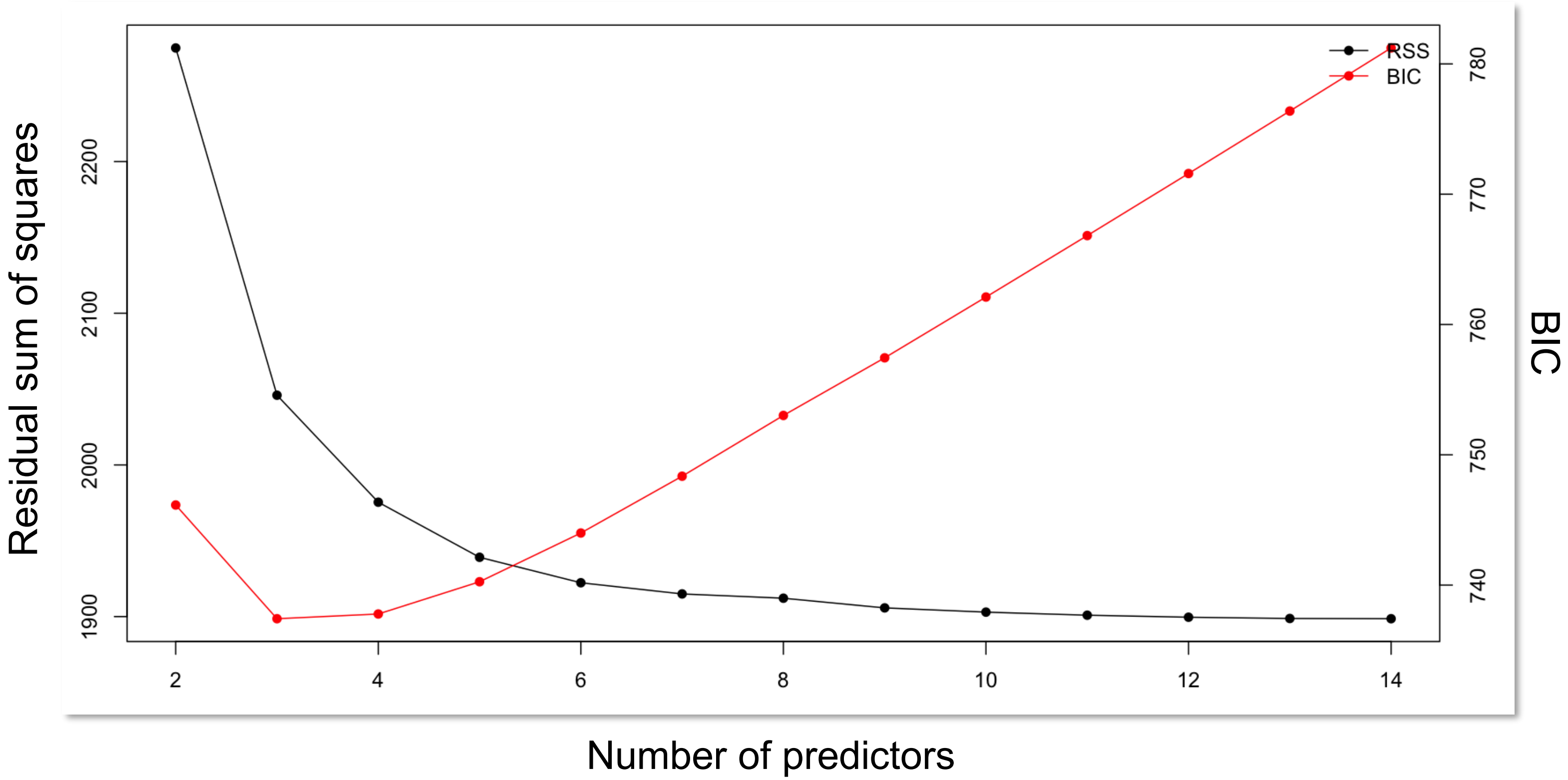
- **Exhaustive search** is the only technique **guaranteed** to find the predictor variable subset with the best evaluation criterion
- Since we look over the whole model space, we can identify the best model(s) at each model size
- Sometimes known as **best subsets** model selection
- **Loss** component is (typically) the residual sum of squares
- Main drawback: exhaustive searching is **computationally intensive**

Best Subset Selection

- Consider as an example linear regression
  - $\mathcal{M}_0: Y = \beta_0 + \varepsilon$ nullmodel
  - $\mathcal{M}_p: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ fullmodel

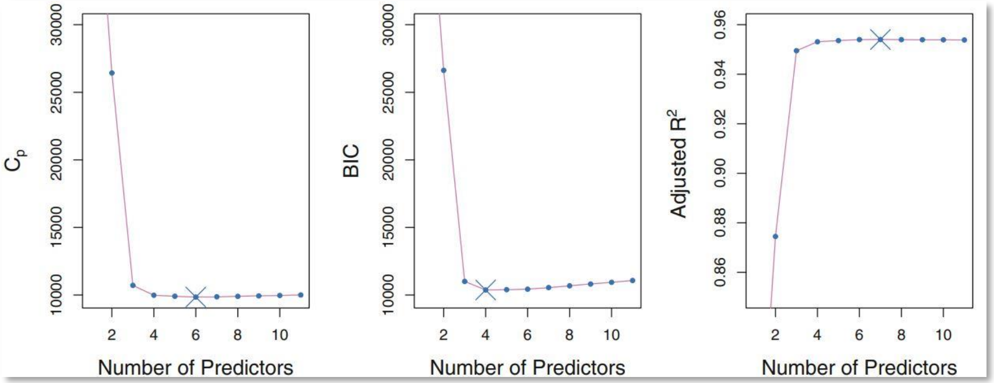
1. Denote $\mathcal{M}_0$ to be the null model
   - Contains no predictors
2. For $k = 1,2, \dots, p$
   - Fit all $\binom{p}{k}$ models that contain exactly k predictors
   - Denote $\mathcal{M}_k$ the best among the $\binom{p}{k}$ models
     - Measured as best against some metric (smallest residual sum of squares or highest accuracy etc.)
3. Select the single best model among the $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$
   - Using cross-validated prediction error or residual sum of squares etc.

Evaluating “Best”



Indirect Approaches

- 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
- Mallow's C_p and the Bayesian information criterion (BIC) and adjusted R^2 for the best model produced by best subset selection on the credit data set.



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 $\rightsquigarrow$ increased chance of findings models that overfit
  - Perform well on training data

THE UNIVERSITY OF
SYDNEY

Feature and Model Selection

Regularisation

Introduction to Regularisation

We cover:

- Ridge
- LASSO
- ElasticNet (Ridge + LASSO)

Shrinkage or Regularisation Methods

- Regularisation methods **shrink** estimated regression coefficients by imposing a penalty on their sizes
- There are different choices for the **penalty**
- The penalty choice drives the **properties** of the method
- Helpful to find solutions for ill-posed problems or to prevent overfitting

* particularly useful when all our variables are highly correlated or in situations where we have a larger number of variables than observations

Penalty

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

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

- Shrinkage methods seek to minimise

$$RSS + \lambda\, R(\boldsymbol{\beta})$$

for some tuning parameter λ and **penalty** function R

Shrinkage Methods

Feature selection
↓
Sparse model
Then using AIC/BIC and cp to select the best ideal model with the predictors.

- The subset selection methods use least squares to fit linear models 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 *regularises* the coefficient estimates. ✓✓
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance. ✓✓

Shrinkage methods allow us to get coefficients closer to 0.
When these coefs are zero, we effectively get feature selection,
since these predictors are dropped.

THE UNIVERSITY OF
SYDNEY

* This method doesn't
do feature selection

Ridge Regression

Shrinkage or Regularisation Methods

- Regularisation methods **shrink** estimated regression coefficients by imposing a penalty on their sizes
- There are different choices for the **penalty**
- The penalty choice drives the **properties** of the method
- Helpful to find solutions for ill-posed problems or to prevent overfitting

Ridge Regression

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

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

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

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

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

Ridge Regression

λ = 0 means we are using Linear Regression

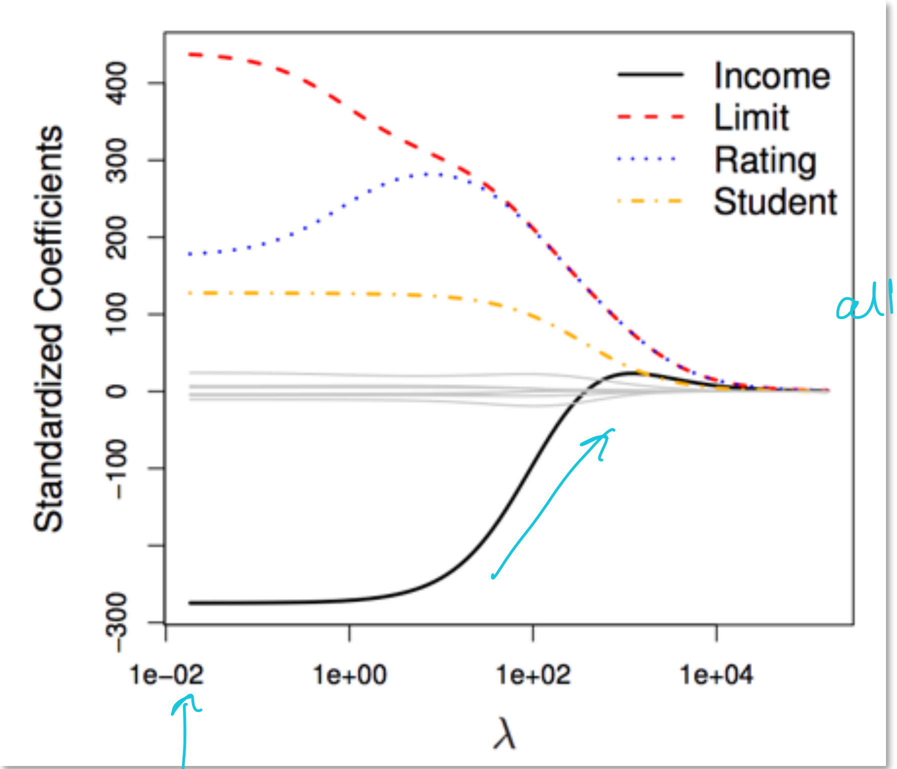
Large λ means coefficients get close to 0.

- 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 β_j towards zero
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.
- Selecting a good value for λ is critical; cross-validation can be used for this.

Over a range of λ values

Credit Card Example

- Note the rate at which each variable goes to 0.



all variables go to 0 as λ increases.

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 $1/c$
  - In other words, regardless of how the j^{th} predictor 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:

$$\boxed{\widetilde{X_{ij}} = \frac{X_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{ij}-\overline{X}_j\right)^2}}}$$

sd of that variables

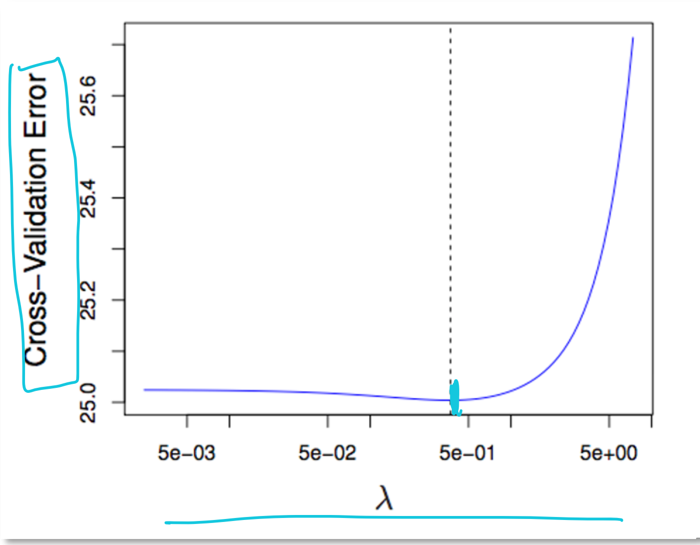
→ effectively standardising our variables.

Selecting the Tuning Parameters

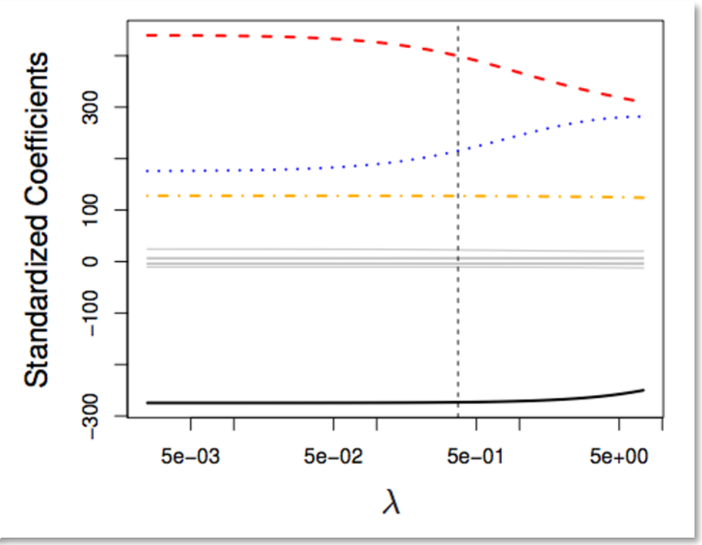
λ

- As for subset selection 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 λ or equivalently.
- **Cross-validation** provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ.
- We then select the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

Credit Data Example



Cross-validation errors that result from applying ridge regression to the credit data set with a range of λ values.



Coefficient estimates as a function of λ. The vertical dashed lines indicate the best value of λ selected by cross-validation.

Ridge Penalty

- Trades some bias in the parameter estimates for reduced variance ✓
- Particularly useful with highly correlated predictors ✓
- Does not perform feature selection: all variables are still included in the model ✓

THE UNIVERSITY OF
SYDNEY

Has a diff penalty than Ridge

Lasso

Shrinkage or Regularisation Methods

- Regularisation methods **shrink** estimated regression coefficients by imposing a penalty on their sizes
- There are different choices for the **penalty**
- The penalty choice drives the **properties** of the method
- Helpful to find solutions for ill-posed problems or to prevent overfitting

The Lasso

- The Lasso is a relatively recent alternative to ridge regression that can be used for feature selection
- The lasso coefficients, $\hat{\beta}_L$, minimise the quantity

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

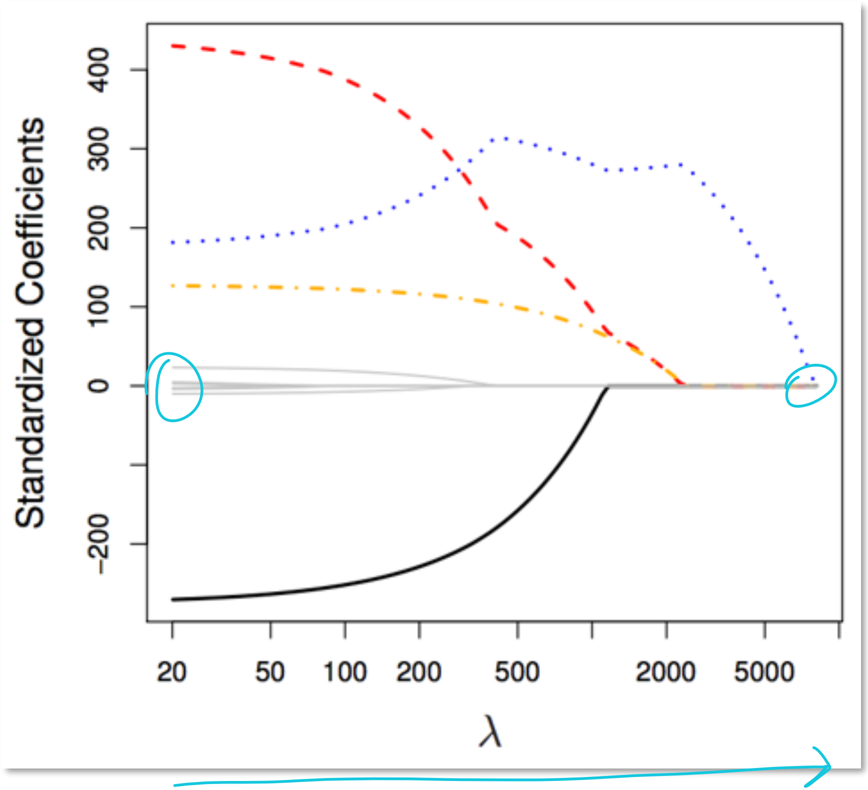
- The lasso uses an ℓ_1 penalty instead of the ℓ_2 penalty used for ridge regression
  - The ℓ_1 norm of a coefficient vector $\boxed{\boldsymbol{\beta} \text{ is } ||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|}$.

The Lasso (cont.)

* Does do feature selection

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero
- However, in the case of the lasso, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large
- Hence, much like best subset selection, the lasso performs **feature selection**
- We say that the lasso yields **sparse** models—that is, models that involve only a subset of variables
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice

Example: Credit Data Set



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?
- The lasso and ridge regression coefficient estimates solve the problems:

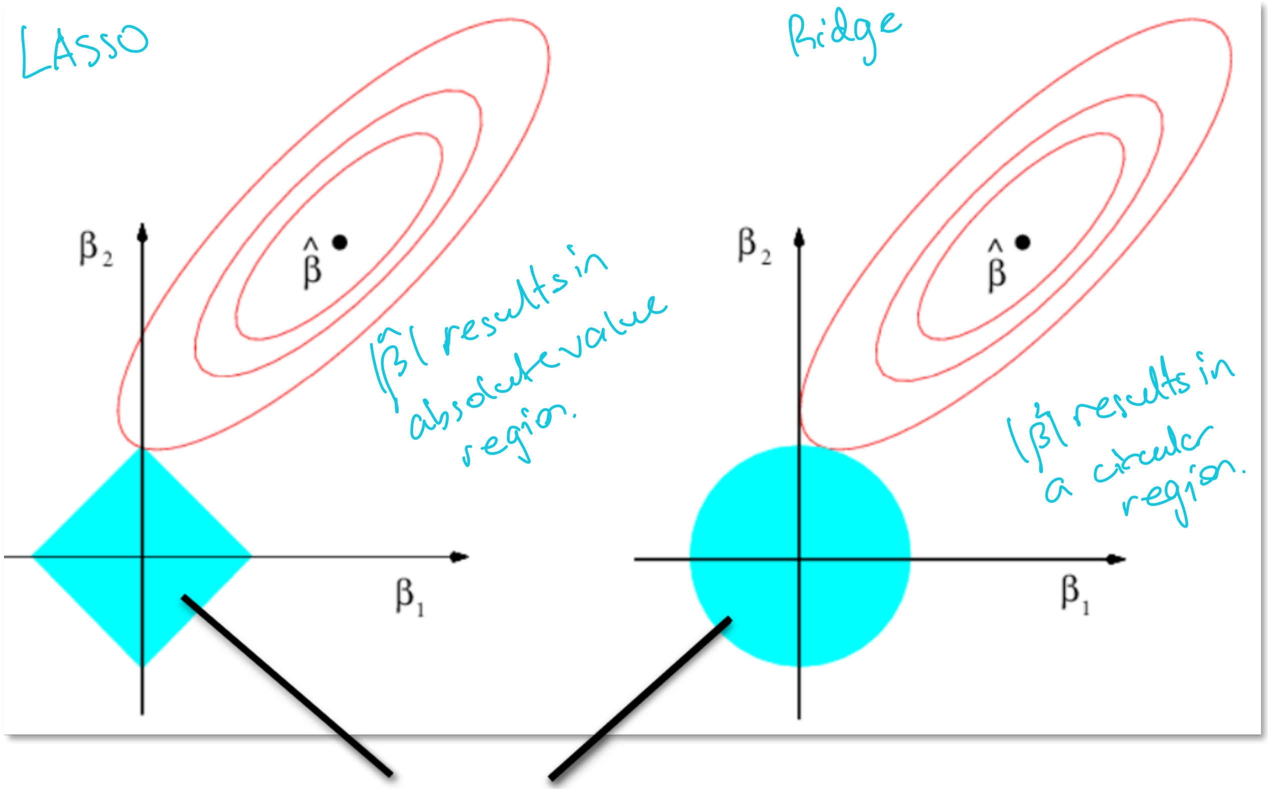
$$\min_{\textbf{beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s.$$

$$\min_{\textbf{beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s.$$

Ridge: β_j^2 (square of the coefficients)

LASSO: $|\beta_j|$ (abs value of coefficients)

Comparison of Constraints



LASSO

$|\hat{\beta}|$ results in absolute value region.

Ridge

$|\hat{\beta}|$ results in a circular region.

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

Selecting the Tuning Parameters

- 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 λ or equivalently, the value of the constraint s. ✓
- **Cross-validation** provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ. ✓
- We then select the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

Extension: Elastic Net

- A more general model, the Elastic net, combines the ridge and the lasso penalties (Friedman, Hastie, and Tibshirani, 2010) into one minimisation problem
- It solves the following penalised minimisation problem:

$$\text{argmin}_{\boldsymbol{\beta}} ||Y - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda\Big((1-\alpha)/2||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1\Big)$$

- Can consider it a weighted combination (mixture) of ℓ_1 and ℓ_2 penalties

- α controls how much penalty is applied from the Ridge and LASSO method.
- λ is still here to apply the overall mixture of the penalty of ℓ_1 and ℓ_2.

THE UNIVERSITY OF
SYDNEY