IST 5535: Machine Learning Algorithms and Applications

Langtao Chen, Spring 2021

Linear Model Selection and Regularization

Reading

- Book Chapter 6 (6.1, 6.2, 6.5, 6.6)

Outline

- (I) Need of Linear Model Selection and Regularization
- (II) Subset Selection
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model
- (III) Shrinkage Methods
  - Ridge Regression
  - The Lasso

AGENDA

- Need of Linear Model Selection and Regularization
- Subset Selection
- Shrinkage/Regularization Methods

Linear Models

- Recap: linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p + \varepsilon$$

- Despite its simplicity, linear model has distinct advantages:
  - Model interpretation is easy;
  - Predictive performance is often good on real-world problems.

- We usually use OLS to estimate linear model. However, other methods may yield better prediction accuracy and model interpretability.

OLS Failure: Include All Predictors in a Model

- A dataset with complete information is not available or expensive to collect
- May have a serious missing data issue with more predictors
- May not be able to accurately measure some predictors
- Using predictors that are unrelated with the response will increase the variance of the prediction

A parsimonious model helps to unveil the underlying relationships with stable estimates of coefficients (especially for an explanatory model).

When OLS May Not Work?

- 1. Prediction Accuracy
  - Given that the true relationship between the response and predictors is approximately linear, the OLS estimates have low bias.
  - According to the tradeoff between bias and variance, an optimal model should also have low variance.
    - If $n >> p$, OLS estimates tend to have low variability.
    - However, if n is not much larger than p, OLS estimates can have high variability and lead to over fitting and poor prediction.
    - If $n < p$, OLS will fail (variance is infinite).

When OLS May Not Work? (cont.)

- 2. Model Interpretability
  - When we have a large number of predictors, some or many variables will be in fact not associated with the response.
  - Including such *irrelevant* variables leads to unnecessarily complexity in the model, thus making the model hard to interpret.
  - The model could be easier to interpret if we remove those irrelevant variables (or set their coefficients as zeros).

Three Classes of Methods Alternative to OLS

- Subset Selection (a.k.a. Feature Selection or Variable Selection)
  - Identify a subset of the p predictors that we believe to be related to the response.
  - Then fit a model using least squares on the reduced set of variables.

- Shrinkage (a.k.a. Regularization)
  - Fit a model involving all p predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates.
  - This shrinkage has the effect of reducing variance.
  - Some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform variable selection.

- Dimension Reduction
  - Project the p predictors into a M-dimensional subspace, where $M < p$.
  - Then these M projections are used as predictors to fit a linear regression model by least squares.

AGENDA

- Need of Model Selection and Regularization
- Subset Selection
- Shrinkage/Regularization Methods

Subset Selection

- Best Subset Selection

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Quiz

- Given a dataset with potentially 3 predictors, how many models will be evaluated during the best subset selection process?
  - (A) 3
  - (B) 4
  - (C) 7
  - (D) 8

Quiz

- Given a dataset with potentially 3 predictors, how many models will be evaluated during the best subset selection process?
  - (A) 3
  - (B) 4
  - (C) 7
  - (D) 8

Answer

Null model $M_{0:}$ 1

Models with one predictor $M_{1:}$ $\binom{1}{3} = 3$

Models with two predictors $M_{2:}$ $\binom{2}{3} = 3$

Models with three predictors $M_{3:}$ $\binom{3}{3} = 1$

In total: 1 + 3 + 3 + 1 = 8 models

Subset Selection

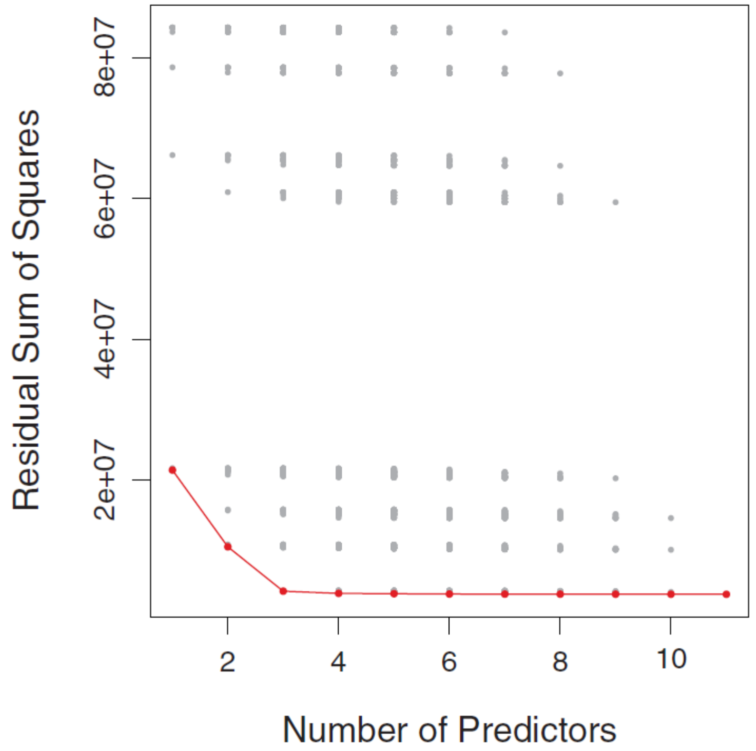
- Best Subset Selection is computationally intensive especially when p is large.

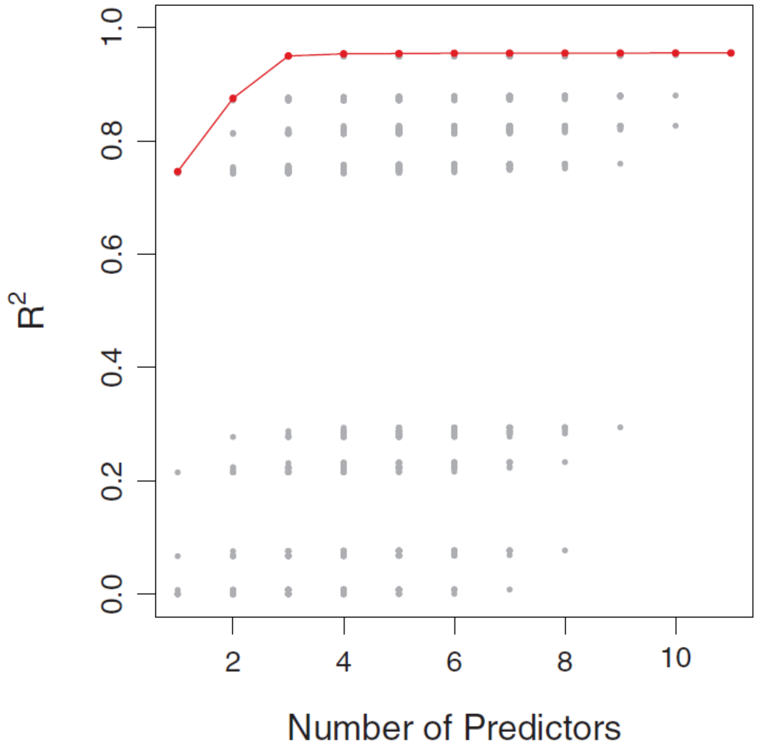
- More attractive alternative: stepwise selection
  - Forward Stepwise Selection
    - Start with no predictors
    - Add them one by one (add the one with largest contribution)
    - Stop when the addition is not statistically significant

  - Backward Stepwise Selection
    - Start with all predictors
    - Successively eliminate least useful predictors one by one
    - Stop when all remaining predictors have statistically significant contribution

Choose the Optimal Model

- As the number of predictors increases, RSS always decreases and R^2 always increases.
- Thus, RSS and R^2 are not suitable for selecting the best model among models with different number of predictors.





Other Measures to Consider for Model Comparison

- The following measures add a heavier penalty on models with many variables:
  - C_p statistic

    $$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) \text{ where } d \text{ is the number of predictors}$$
  - AIC (Akaike information criterion)

    $$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$
  - BIC (Bayesian information criterion)

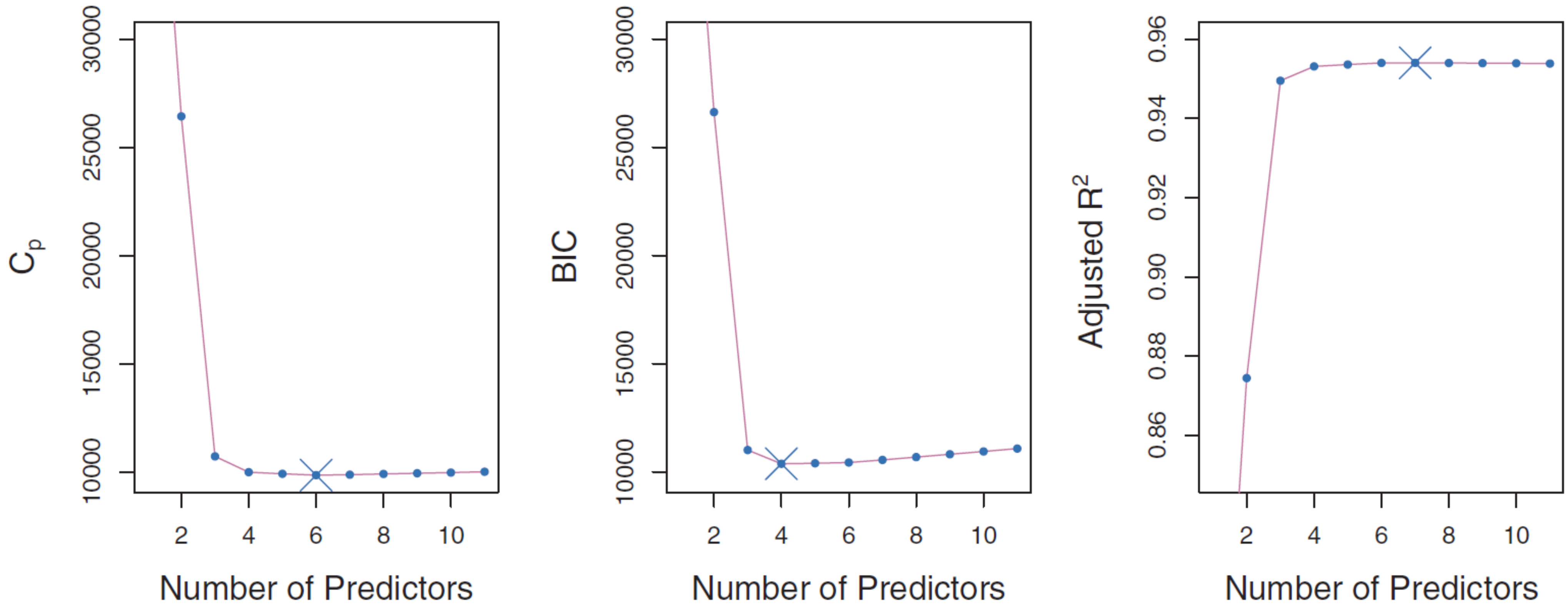
    $$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$
  - Adjusted R^2

    $$R^2_{adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Smaller C_p, AIC, and BIC are better; Larger adjusted R^2 is better.

An Example of Model Selection

- Smaller C_p, AIC, and BIC are better;
- Larger adjusted R^2 is better.



AGENDA

- Need of Model Selection and Regularization
- Subset Selection
- Shrinkage/Regularization Methods

Shrinkage/Regularization Methods

- The above subset selection methods involve using OLS 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, or equivalently, that *shrinks* the coefficient estimates towards zero.

- Shrinking the coefficient estimates can significantly reduce their variance.

- Regularization reduces parameters and shrinks the model, thus avoiding over-fit.

- Two best known shrinkage methods: *ridge regression* and the *lasso*

Ridge Regression

- The OLS fitting procedure minimizes the RSS

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

- The ridge regression minimizes

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

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

shrinkage penalty

Shrinkage Penalty

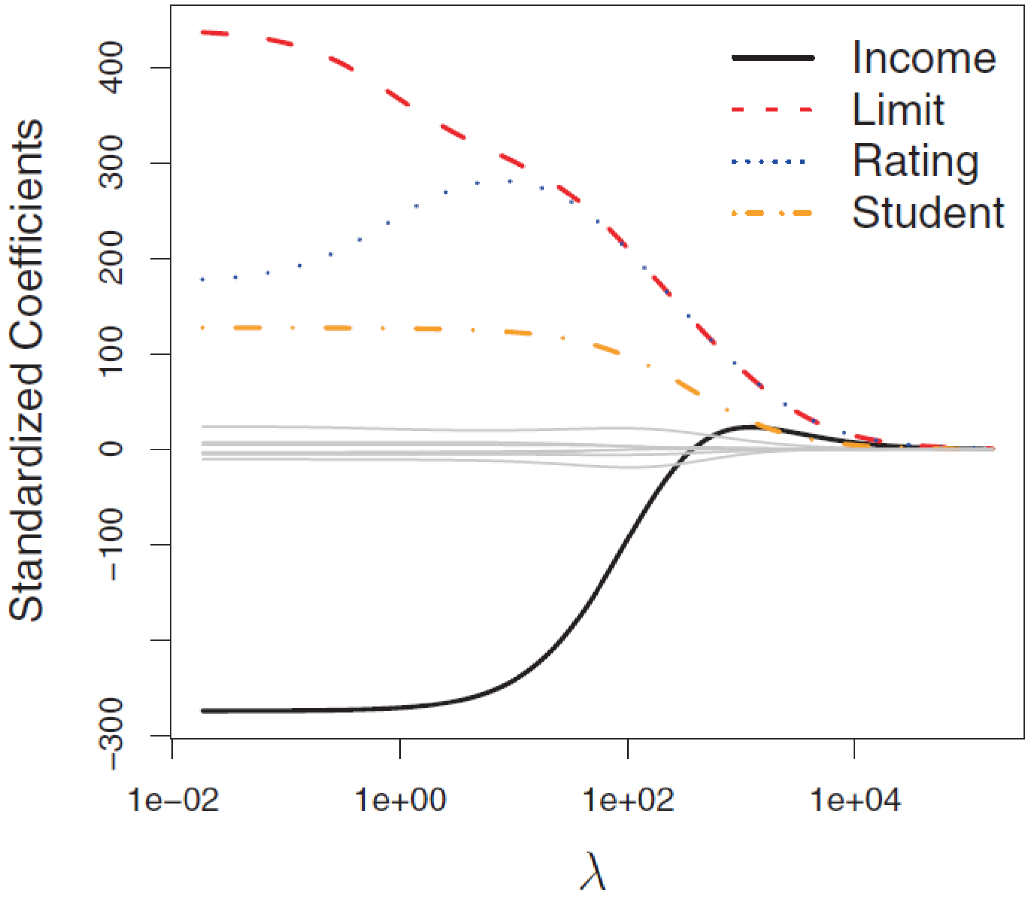
- The ℓ_2 penalty term $||\beta_j||_2 = \sum_{j=1}^{p} \beta_j^2$ has the effect of shrinking coefficient estimates β_j towards zero.

- The tuning parameter $\lambda \geq 0$ controls the relative importance of the penalty term in the overall optimization of the objective function.
  - When $\lambda = 0$, the penalty term does not have effect. Ridge regression results in OLS estimates;
  - When λ is large, the impact of the penalty term grows. $\beta_j (j = 1,2, \dots, p)$ has to be close to zero.

- It's critical to select an appropriate value for λ. In practice, cross-validation is used to tune this parameter.

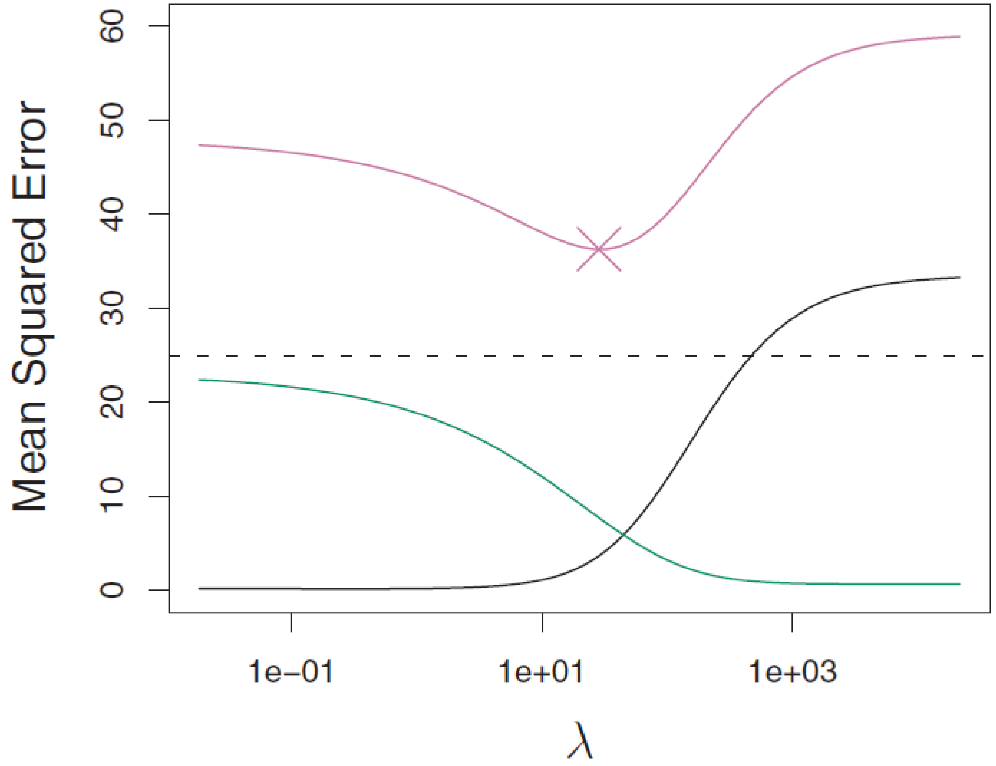
Example: Ridge Regression on Credit Data

- As λ increases, the ridge coefficient estimates shrink towards zero.



Why Does Ridge Regression Improve Over OLS?

- OLS estimates have low bias. However, if the condition $n >> p$ does not hold, OLS estimates may have large variance.
- By adding the shrinkage penalty, ridge regression leads to more biased but less variable estimates.
- Ridge regression can make a better trade-off between bias and variance, thus improving over OLS.
- Ridge regression works best in situations where OLS estimates have high variance.



Black: Squared Bias
Green: Variance
Purple: Test MSE

The Lasso

- One problem for ridge regression:
  - It shrinks all coefficients towards zero, but it will not set any of them exactly to zero;
  - Thus, ridge regression cannot conduct variable selection.
  - As all *p* variables will be included in the final model, there could be a challenge in model interpretation.

- Lasso is a more recent alternative to ridge regression that overcomes this disadvantage. Lasso coefficients minimize:

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

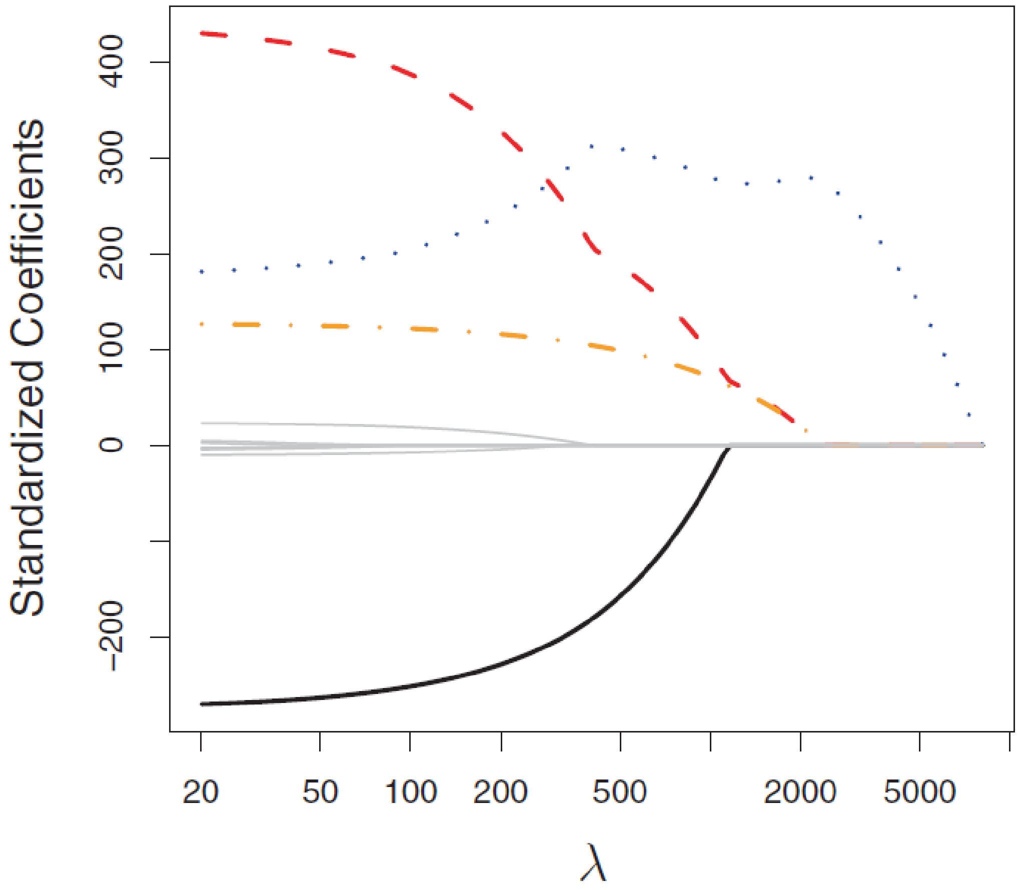
Lasso works in similar way as ridge regression, except using an ℓ_1 penalty.

Lasso Penalty Term

- The lasso ℓ_1 penalty $||\beta_j||_1 = \sum_{j=1}^{p} |\beta_j|$ can force some coefficient estimates to be exactly equal to zero, when the tuning parameter λ is large enough.
- Thus, lasso performs variable selection.
- Models generated from lasso are generally much easier to interpret than those produced by ridge regression.

Example: Lasso on Credit Data

- Depending on the value of λ, lasso can produce a model with any number of variables.

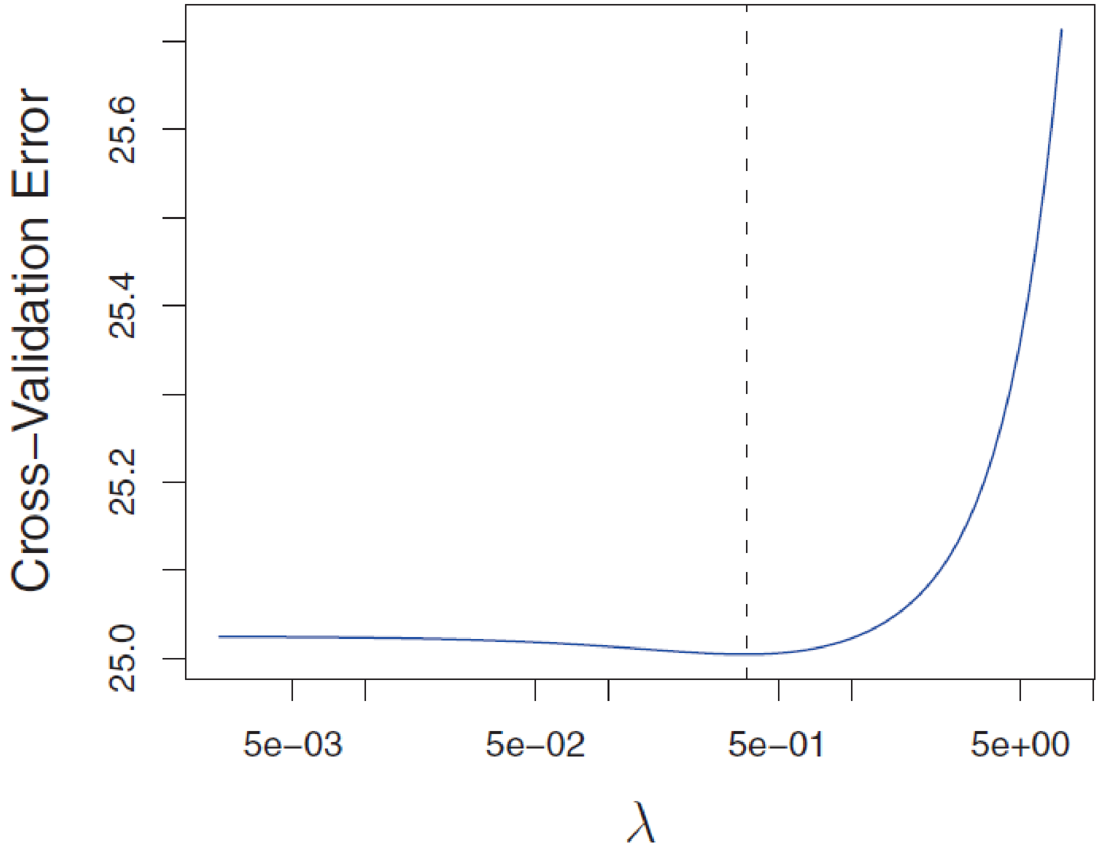


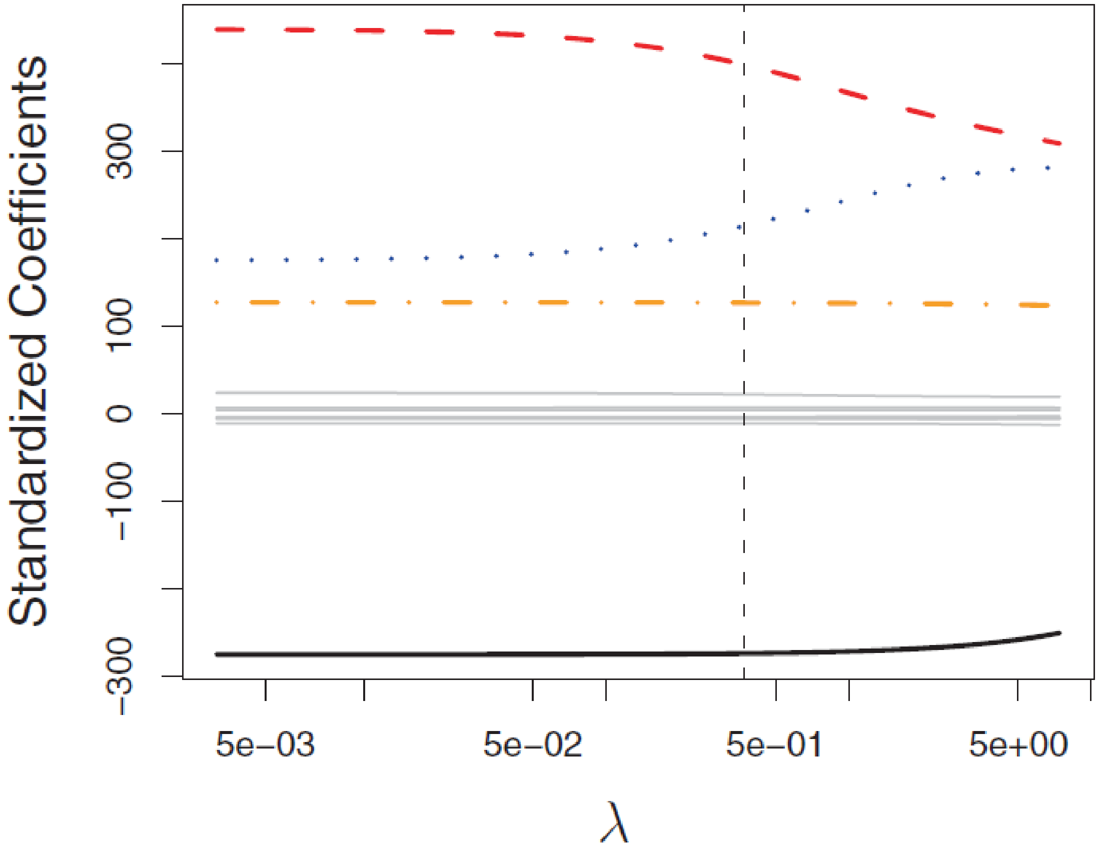
Selecting the Tuning Parameter

- Implementing ridge regression and the lasso requires a method of selecting an optimal value for the tuning parameter λ.
- Cross-validation provides a simple way to tune parameters.

```
Define a grid of parameter values
for each parameter value do
    for each cross-validation iteration do
        Hold-out specification samples
        [Optional] Pre-process the data
        Fit the model on the remainder
        Predict the hold-out samples
    end
    Calculate the average performance across all iterations
end
Determine the optimal parameter value
Fit the final model to all training data using the optimal parameter value
```

Example: Tuning λ for Ridge Regression





RECAP: OUTLINE

- (I) Need of Linear Model Selection and Regularization
- (II) Subset Selection
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model
- (III) Shrinkage Methods
  - Ridge Regression
  - The Lasso

Q & A