High-Dimensional Methods and Inference on Structural and Treatment Effects [1]

Jasmine. Hao[1]

[1]University of Hong Kong

ECON 6083: Machine Learning

[1]This section is based on [James et al., 2013], Chapter 5,6.

Outline

Linear Model

Recall the linear model

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

Linear model has distinct advantages in terms of its interpretability and often shows good predictive performance.

Hence the linear model can be improved by replacing ordinary least squares fitting with some alternative fitting procedures.

Why Consider Alternatives to Least Squares?

- Prediction Accuracy
  Especially when $p > n$, there is no longer a unique least squares coefficient estimate: the variance is **infinite** so the least squares estimates cannot be used at all. We can **constrain** or **shrink** the estimated coefficients to reduce the variance at the cost of a negligible increase in bias.
- Model Interpretability by removing **irrelevant** features,
  - setting the corresponding coefficient estimates to zero,
  - automatically performing **feature selection** or **variable selection**.

3 Classes of Methods

- **Subset Selection** Select a subset of the p predictors believed to be related to the response and fit a model using least squares.
- Shrinkage Fit a model with all p predictors, shrinking coefficients towards zero relative to least squares estimates. This reduces variance and can select variables.
- Dimension Reduction Project p predictors into an M-dimensional subspace ($M < p$) using M linear combinations. Use these projections to fit a linear regression model by least squares.

Subset Selection

We consider two methods for selecting subsets of predictors

- Best Subset Selection
- Stepwise Selection

Outline

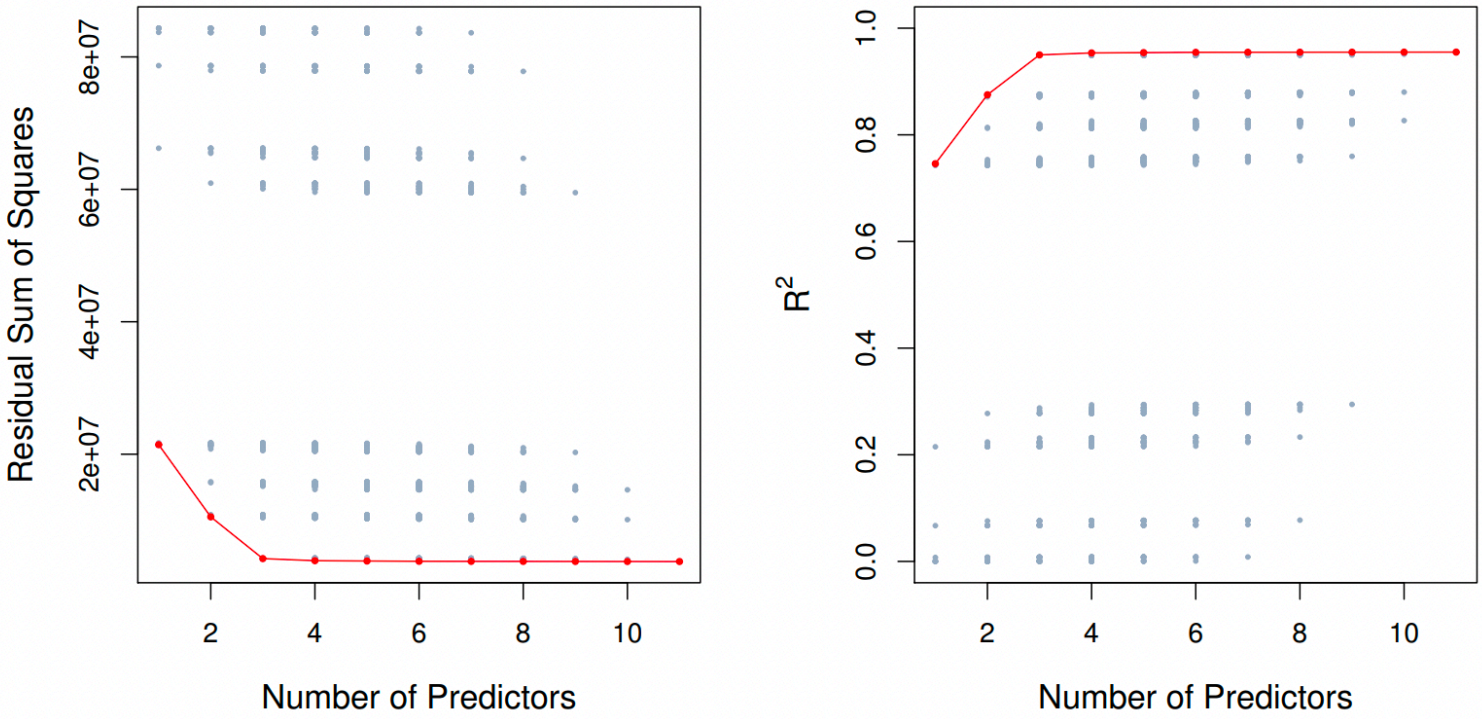
Best Subset Selection

We perform least squares regression for every combination of p predictors, fitting models from one predictor up to p, and evaluate all 2^p possible models to find the optimal one.
The procedure is outlined as follows:

1. Define M_0 as the null model, predicting the sample mean.
2. For each $k = 1, 2, \cdots, p$:
   1. Fit all $\binom{p}{k}$ models with k predictors.
   2. Select the model M_k with the lowest RSS or highest R^2.
3. Choose the best model from M_0 to M_p based on cross-validation, C_p, AIC, BIC, or adjusted R^2.

Example: the Credit Data Set

Figure: Best Subset Selection (James et al. (2021), 228.)



The red frontier tracks the best model for a given number of predictors, according to RSS and R^2.

Outline

Stepwise Selection

Stepwise selection includes:

- Forward Stepwise Selection
- Backward Stepwise Selection

Forward Stepwise Selection I

- 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 II

The steps are as follows:

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \cdots, p-1$:
   1. Consider all $p-k$ models that augment the predictors in M_k with one additional predictor.
   2. Choose the best among these $p-k$ models, and call it M_{k+1}. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $M_0, \cdots, M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Example: the Credit Data Set

Figure: Best Subset vs. Forward Stepwise (James et al. (2021), 231.)

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income` `student`, `limit`	`rating`, `income`, `student`, `limit`

The first three models are identical but the fourth models differ.

Problems: Not Guarantee to Find Best

Although forward stepwise tends to do well in practice, it is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

For instance, suppose that in a given data set with $p = 3$ predictors:

- the best possible one-variable model contains X_1
- the best possible two-variable model instead contains X_2 and X_3

Then forward stepwise selection will fail to select the best possible two-variable model, because M_1 will contain X_1, so M_2 must also contain X_1 together with one additional variable.

Backward Stepwise Selection I

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

Backward Stepwise Selection II

The steps are as follows:

1. Let M_p denote the full model, which contains all p predictors.
2. For $k = p, p-1, \cdots, 1$:
   1. Consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictors.
   2. Choose the best among these k models, and call it M_{k-1}. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $M_0, \cdots, M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward Stepwise vs. Backward Stepwise

Similarities with forward stepwise selection:

- Backward selection explores only $1 + p(p+1)/2$ models, suitable when p is too large for best subset selection.
- It does not ensure the best model with a subset of p predictors.

Differences from forward stepwise selection:

- Backward selection requires $n > p$ to fit the full model.
- Forward stepwise works even if $n < p$, making it feasible for very large p.

Hybrid Approaches

Hybrid approaches combine elements of best subset, forward stepwise, and backward stepwise selection, producing models that are similar yet distinct.

- Variables are added sequentially as in forward selection, but any that do not enhance model fit are subsequently removed.
- The goal is to mimic the best subset selection's effectiveness while preserving the computational efficiency of stepwise methods.

Outline

Choosing the Optimal Model

Models with all predictors have the smallest RSS and largest R^2 due to their relation to training error. However, our goal is a model with low **test error**, not just low training error, because:

- Training error often misrepresents test error.
- Consequently, RSS and R^2 are inadequate for choosing the best model from a set with varying numbers of predictors.

Choosing the Optimal Model

Two methods for estimating test error:

- Indirect estimation: Adjust training error for overfitting bias.
- Direct estimation: Use validation set or cross-validation.

Indirect Approach

Four approaches adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables:

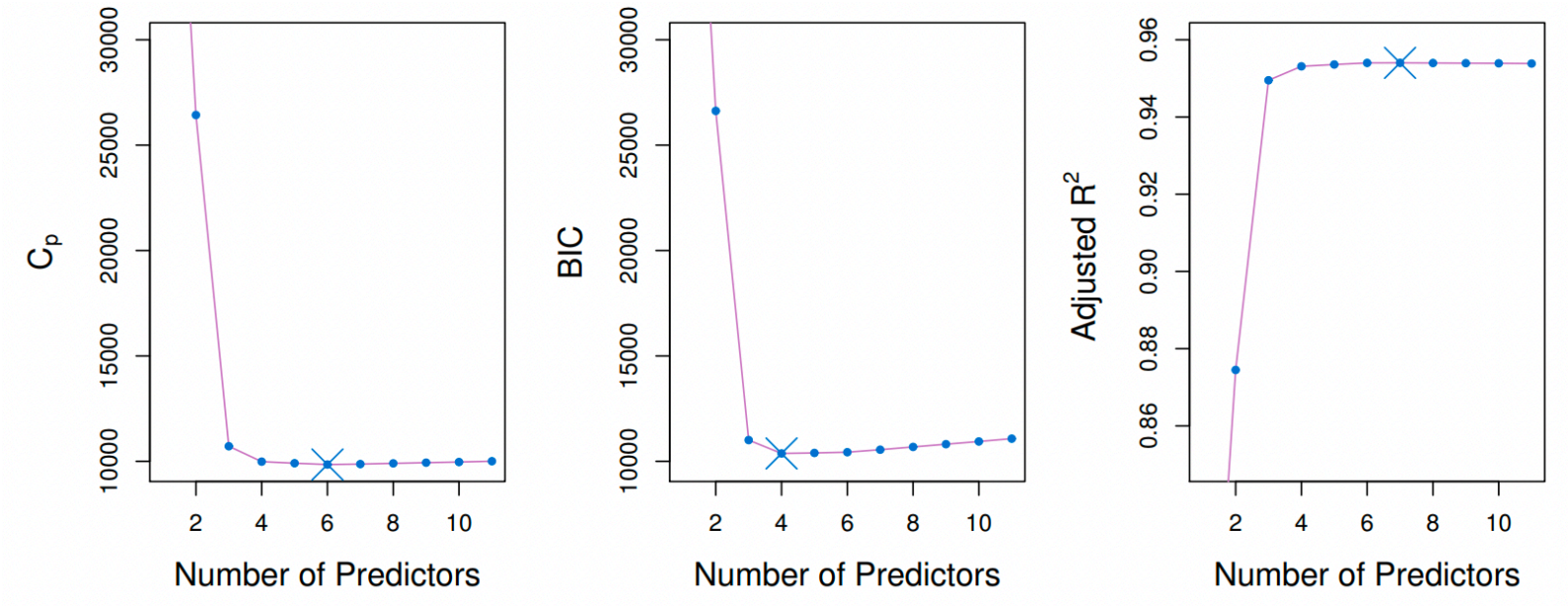
- C_p
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- adjusted R^2

C_p , AIC and BIC

- $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$,
  - where d is the number of parameters, and $\hat{\sigma}^2$ is the error variance estimate per response.
- $\text{AIC} = -2\log L + 2d$,
  - L is the likelihood function's maximized value.
  - In linear models with Gaussian errors, maximum likelihood equals least squares, making C_p and AIC equivalent.
- $\text{BIC} = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$,
  - BIC adds a larger penalty for models with more variables, as $\log n > 2$ for $n > 7$.
- Adjusted $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$,
  - Maximizing adjusted R^2 minimizes $\frac{RSS}{n-d-1}$.

Example: the Credit Data Set

Figure: C_p, BIC, and Adjusted R^2 (James et al. (2021), 233.)



- For C_p, AIC, and BIC, a small value indicates a low test error.
- For adjusted R^2, a large value indicates a low test error.

Direct Approach

Two approaches:

- Validation set approach
- Cross-validation approach

Validation Set Approach

- Samples are randomly split into a **training set** and a **validation set**.
- The model is trained on the training set and predictions are made for the validation set.
- Validation set error estimates the test error, typically using MSE for quantitative responses and misclassification rate for qualitative responses.

Drawbacks of validation set approach:

- Validation estimate of test error varies significantly based on the composition of training and validation sets.
- Only a subset of observations, those in the training set, are used to fit the model, limiting data usage.
- The validation set error likely **overestimates** the test error for the model trained on the full data set.

Cross-Validation Approach I

To deal with the drawbacks of validation set approach, we introduce the cross-validation approach.

The idea of K-fold cross-validation is:

1. Randomly divide the data into K equal-sized parts
2. Leave out part k, fit the model to the other $K-1$ parts (combined), and then obtain predictions for the left-out kth part.
3. Do the second step for each part $k = 1, 2, \cdots, K$, and then the results are combined.

Cross-Validation Approach II

The steps are as follows:

1. Let the K parts be $C_1, C_2, \cdots, C_K$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k: if N is a multiple of K, then $n_k = n/K$
2. Compute

$$\text{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k,$$

   where $\text{MSE}_k = \sum_{i \in C_k} \frac{(y_i - \hat{y}_i)^2}{n_k}$, and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.
3. Setting $K = n$ yields n-fold or leave-one out cross-validation (LOOCV).

Cross-Validation Approach III

In the past, performing cross-validation was computationally prohibitive for many problems with large p and/or large n, and so indirect approaches were more attractive approaches for choosing among a set of models.

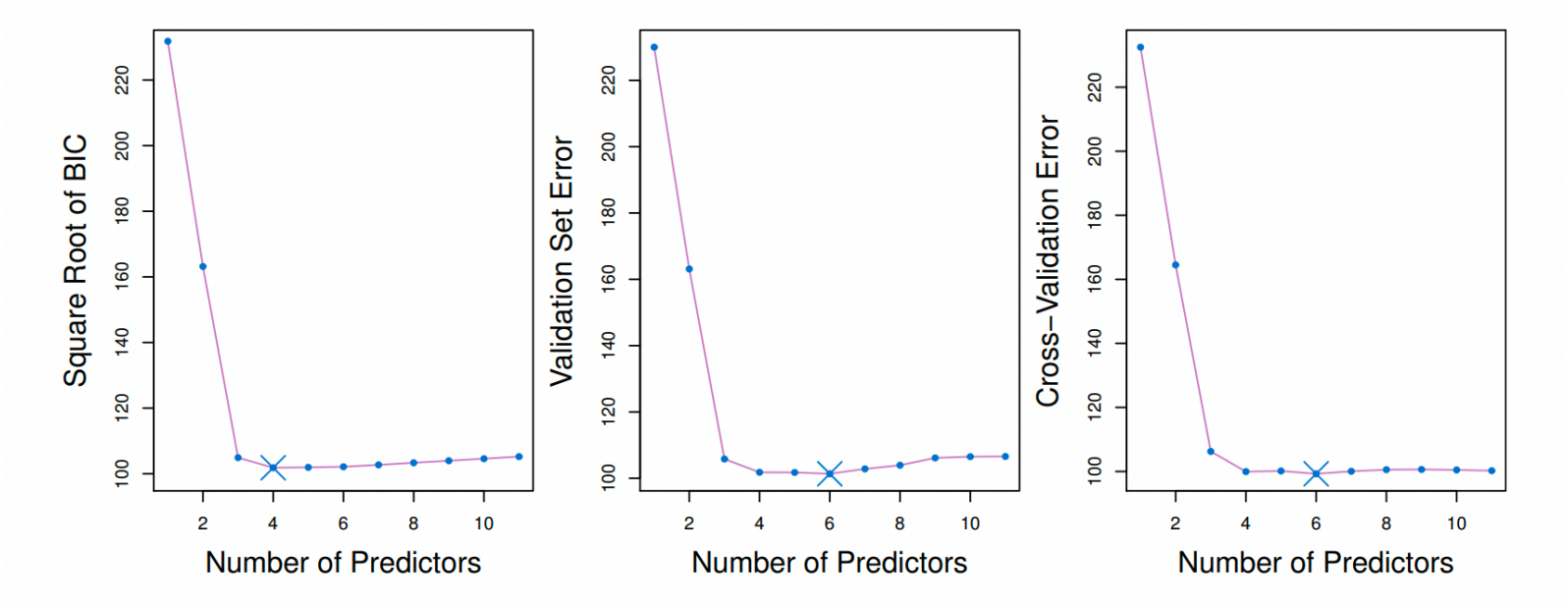
However, nowadays with fast computers, the computations required to perform cross-validation are hardly ever an issue.

The cross-validation method has advantages because it:

- provides a direct estimate of the test error.
- makes fewer assumptions about the true underlying model, such as do not require an estimate of the error variance σ^2.
- can also be used in a wider range of model selection tasks, even if 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.

Example: the Credit Data Set

Figure: BIC, Validation Set, Cross-Validation (James et al. (2021), 236.)



Although three approaches do not suggest the same number of predictors, they suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

Shrinkage

- 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, or equivalently, that **shrinks the coefficient estimates towards zero**.
- Two regressions: **ridge** and **lasso**.

Outline

Ridge I

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

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

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

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

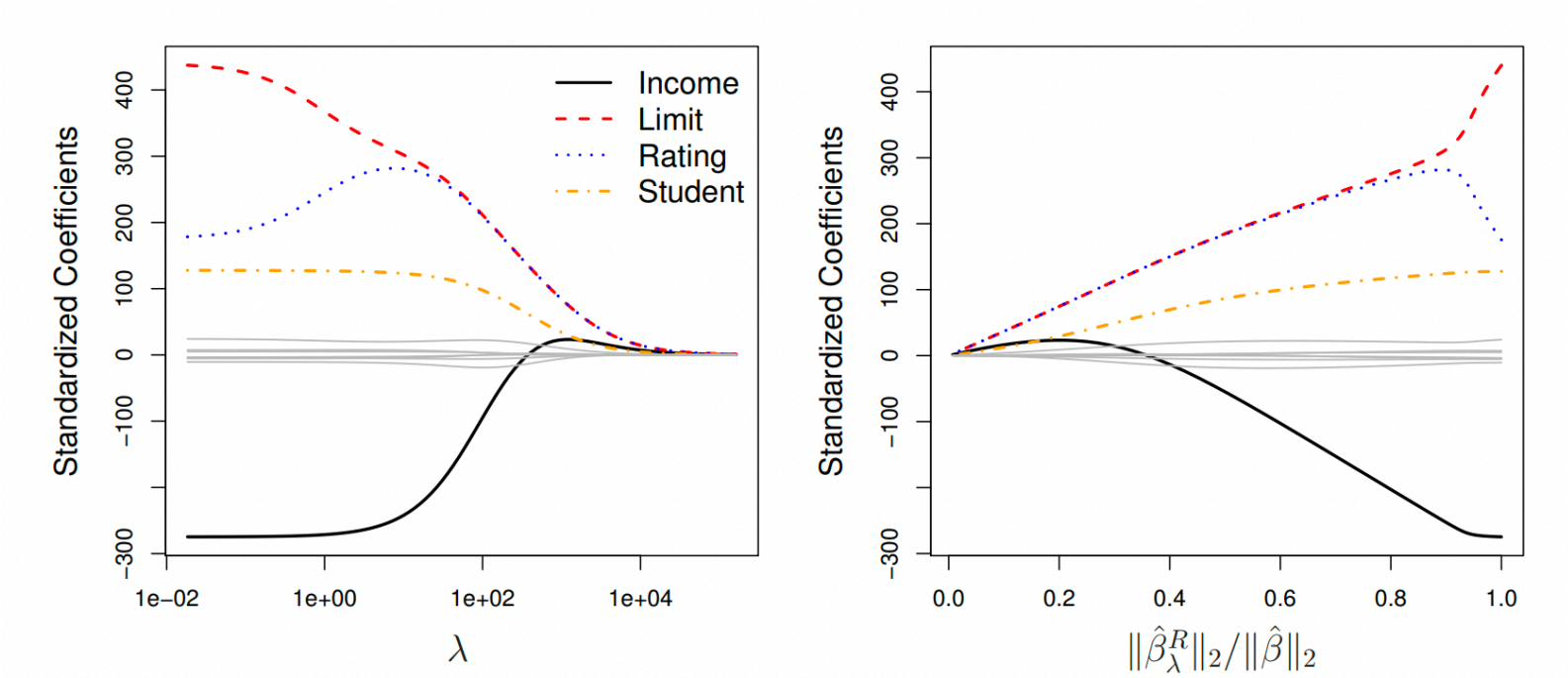
where $\lambda \geq 0$ is a **tuning parameter**, to be determined separately.

Ridge II

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
- However, the second term, $\lambda \sum_{j=1}^{p} \beta_j^2$, called a **shrinkage penalty**, is small when $\beta_0, \beta_1, \cdots, \beta_p$ 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 is used for this.

Example: the Credit Data Set

Figure: λ and coefficients (James et al. (2021), 238.)



Here, $\hat{\beta}$ denotes the vector of least squares coefficient estimates, and $||\beta||_2$ denotes the l_2 norm, and is defined as $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$.

As λ increases, $||\hat{\beta}^R_\lambda||_2$ will always decrease, and so will $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$.

Standardization

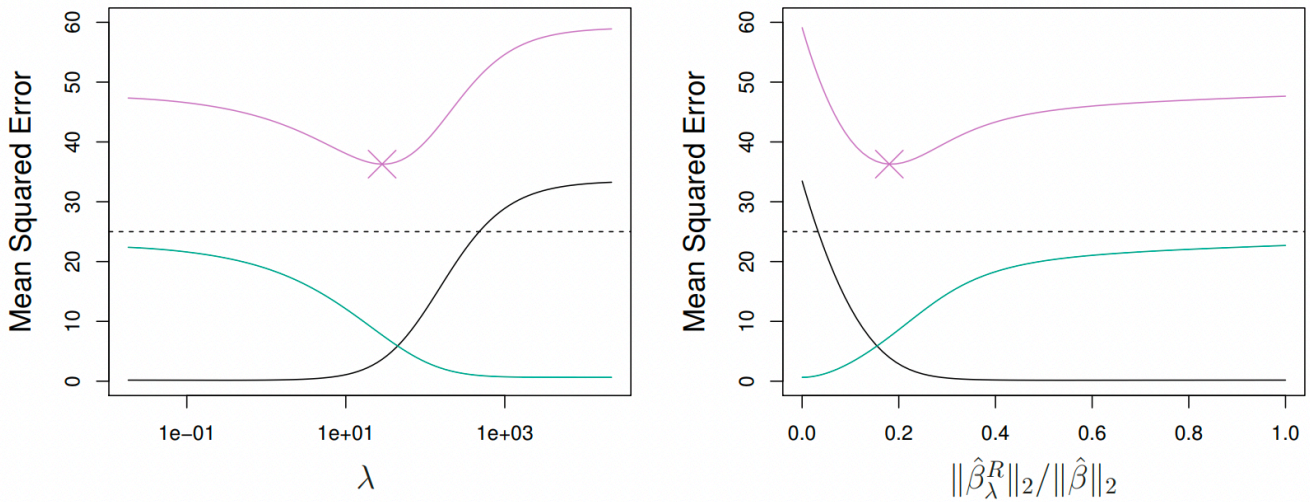
- Standard least squares coefficients are scale equivariant: multiplying X_j by c scales the coefficients by $1/c$. Thus, $X_j\hat{\beta}_j$ remains constant.
- Ridge regression coefficients, however, are sensitive to scaling due to the penalty on squared coefficients in its objective function.
- It's advisable to standardize predictors before applying ridge regression, using:

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

Ridge Improves Over Least Squares I

Ridge regression's advantage is rooted in the **bias-variance trade-off**. As λ increases, variance decreases but bias increases.

Figure: λ and coefficients (James et al. (2021), 240.)



Black: squared bias; green: variance; purple: test MSE.
Horizontal dashed line: the minimum possible MSE.
Purple cross: ridge regression models for which the MSE is smallest.

Ridge Improves Over Least Squares II

- When the relationship between response and predictors is nearly linear, least squares estimates have low bias but high variance.
- This means that small changes in training data can cause large changes in the coefficients.
- When p is nearly as large as n, least squares estimates become extremely variable, and when $p > n$, they don't even have a unique solution.
- Ridge regression can perform well in such cases by trading off a slight increase in bias for a significant decrease in variance.
- Thus, ridge regression is most effective when least squares estimates have high variance.

Outline

Lasso Regression

- Ridge regression includes all p predictors, unlike subset selection which selects only a subset.
- Lasso, a newer alternative to ridge regression, addresses this by using coefficients, $\hat{\beta}_\lambda^L$, that minimize a specific quantity.
- Unlike ridge regression that minimizes

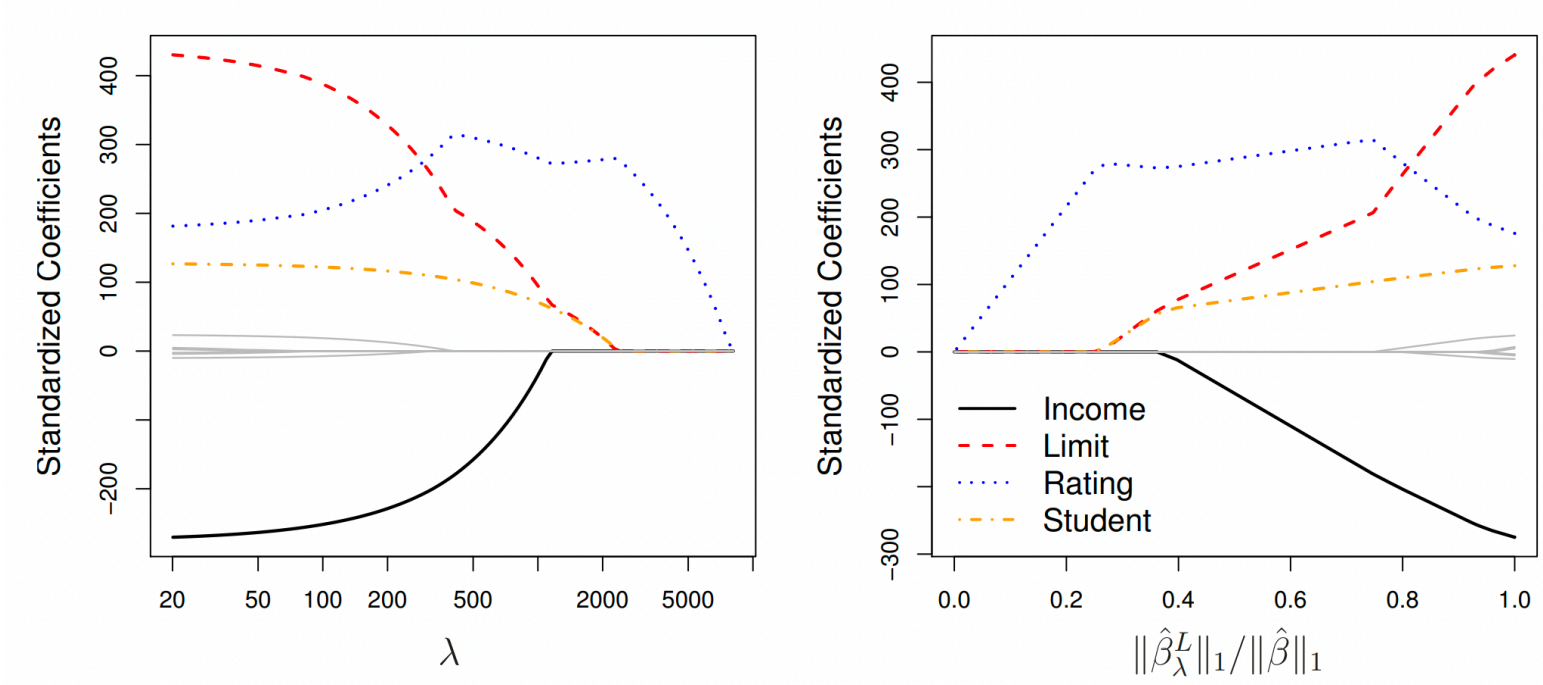
$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|,$$

lasso uses an l_1 penalty, $||\beta||_1 = \sum |\beta_j|$, promoting sparsity.

- Like ridge regression, lasso also shrinks coefficients towards zero.
- The l_1 penalty in lasso forces some coefficients to zero if λ is large, enabling variable selection.
- Lasso produces **sparse models**, involving only a subset of variables.
- Choosing an appropriate λ for lasso is crucial, with cross-validation preferred for this purpose.

Example: the Credit Data Set

Figure: λ and coefficients (James et al. (2021), 242.)



Outline

Another Formulation I

We noticed that:
Ridge regression cannot result in coefficient estimates exactly equal to zero, while lasso can. Why?

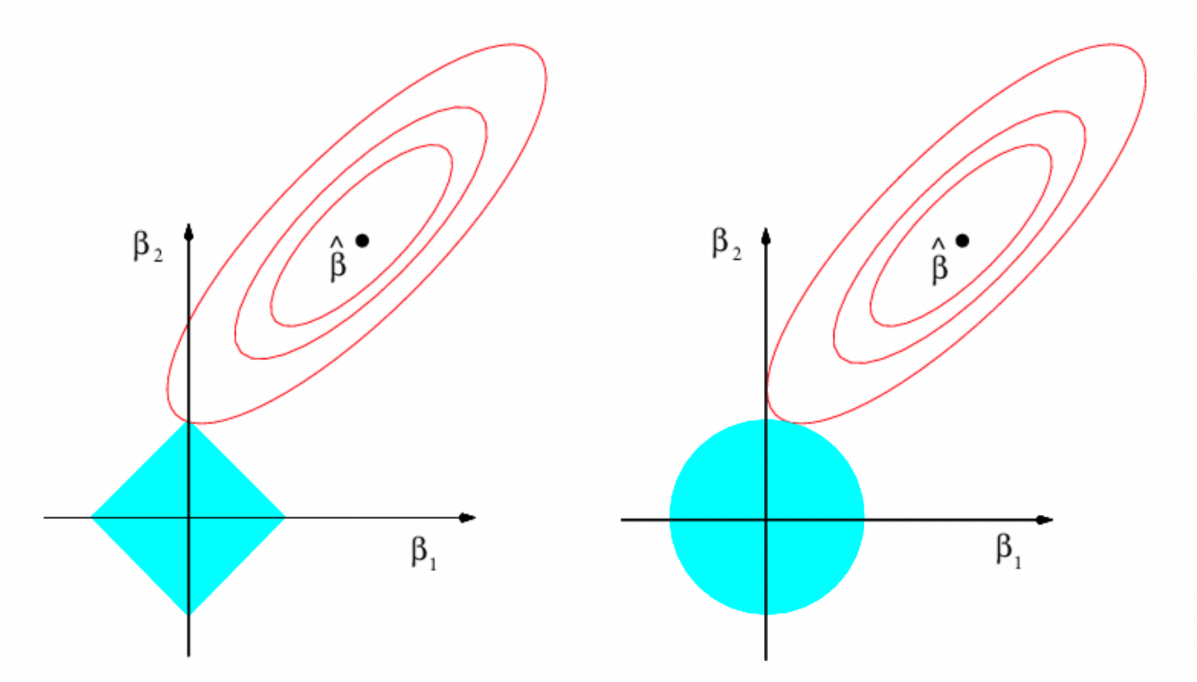
- The lasso regression coefficient estimates solve the problem

$$\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2, \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s.$$

- The ridge regression coefficient estimates solve the problem

$$\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2, \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s.$$

Figure: Lasso (left) vs. Ridge (right) (James et al. (2021), 244.)



- All of the points on a particular ellipse have the same RSS value.
- As the ellipses expand away from the least squares coefficient estimates, the RSS increases.

Lasso coefficient estimates can be zero.

- This is because asso constraint has corners at each of the axes, making it more likely for the ellipse to intersect the constraint region at an axis.
- In higher dimensions, many of the coefficient estimates may equal zero simultaneously.

Ridge coefficient estimates will be exclusively non-zero.

- This is because ridge regression has a circular constraint that lacks sharp points.

Therefore, lasso has a major advantage over ridge regression:

- Lasso produces simpler and more interpretable models that involve only a subset of the predictors.

Which Method to Choose?

Figure: Example: ridge outperforms lasso (James et al. (2021), 245.)

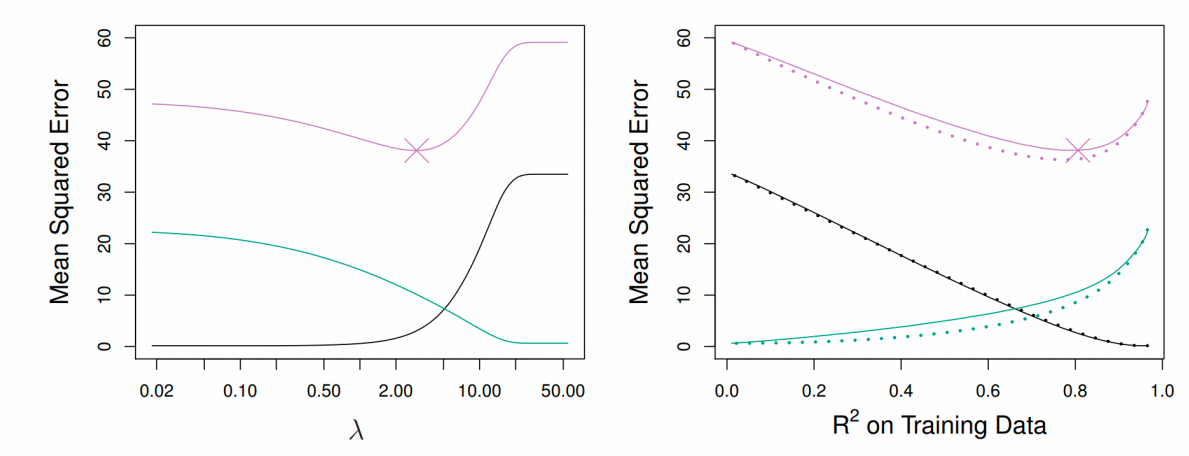
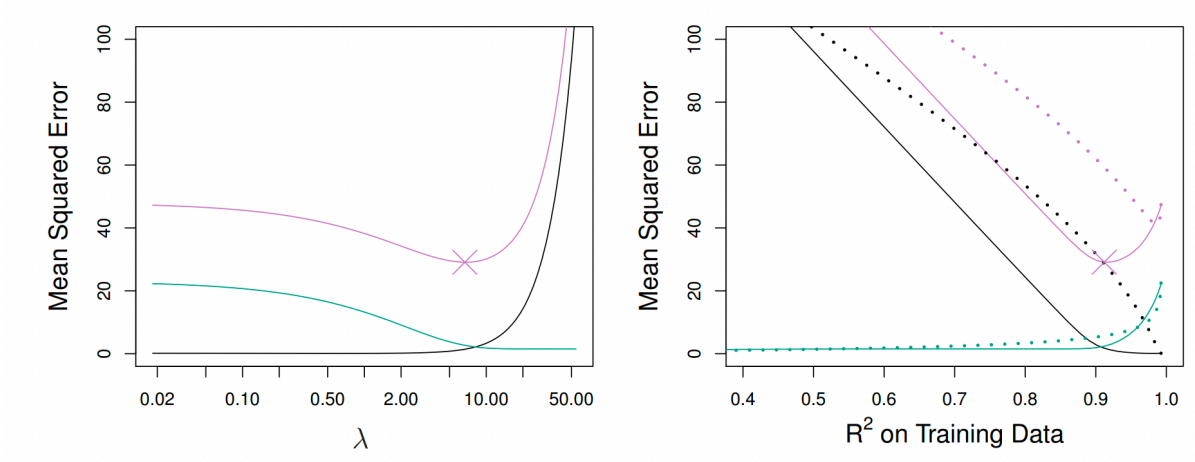


Figure: *

Black: squared bias; green: variance; purple: test MSE; lasso: solid; ridge: dotted.
Purple cross: lasso models for which the MSE is smallest.

- All 45 predictors influenced the response, with none of the coefficients $\beta_1, \cdots, \beta_{45}$ being zero.
- Lasso assumes some coefficients are zero, leading to ridge regression's superior prediction accuracy in this scenario.

Figure: Lasso (left) vs. Ridge (right) (James et al. (2021), 246.)



- Here, the response is a function of only 2 out of 45 predictors.
- Consequently, lasso tends to outperform ridge regression in terms of bias, variance, and MSE.

- Neither ridge regression nor the lasso consistently outperforms the other.
- The lasso tends to perform better when few predictors influence the response.
- The exact number of relevant predictors is typically unknown beforehand in real data sets.
- Cross-validation can help determine the more effective approach for a specific data set.

Outline

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 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. The steps are as follows:

1. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ.
2. We then select the tuning parameter value for which the cross-validation error is smallest.
3. Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

Examples I

Figure: Small λ (James et al. (2021), 250.)

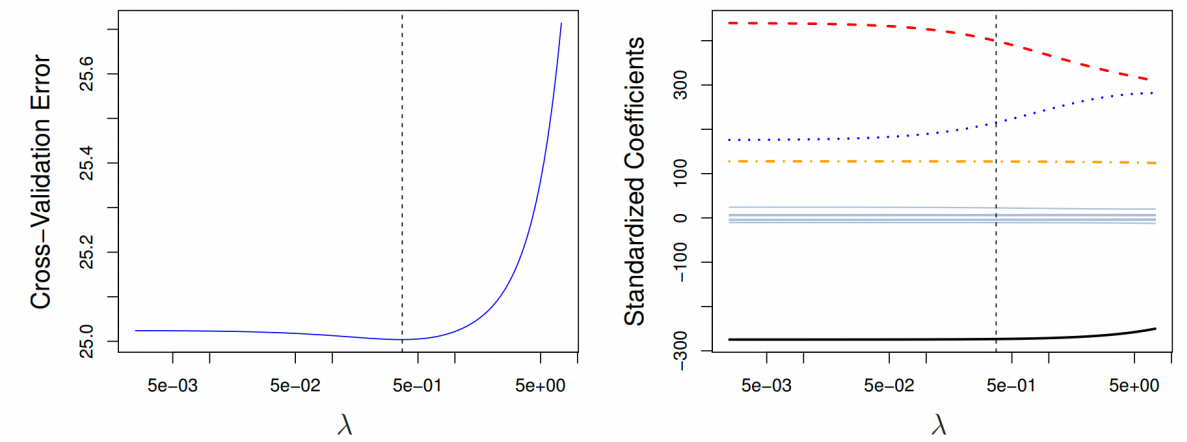


Figure: *

Left: Cross-validation errors result from applying ridge to Credit data set.
Right: The coefficient estimates as a function of λ.

Vertical dashed lines: selected λ.

- The selected λ is relatively small, indicating that the optimal fit only involves a small amount of shrinkage relative to the least squares solution.
- So, we might simply use the least squares solution.

Examples II

Figure: Only 2 predictors related (James et al. (2021), 251.)

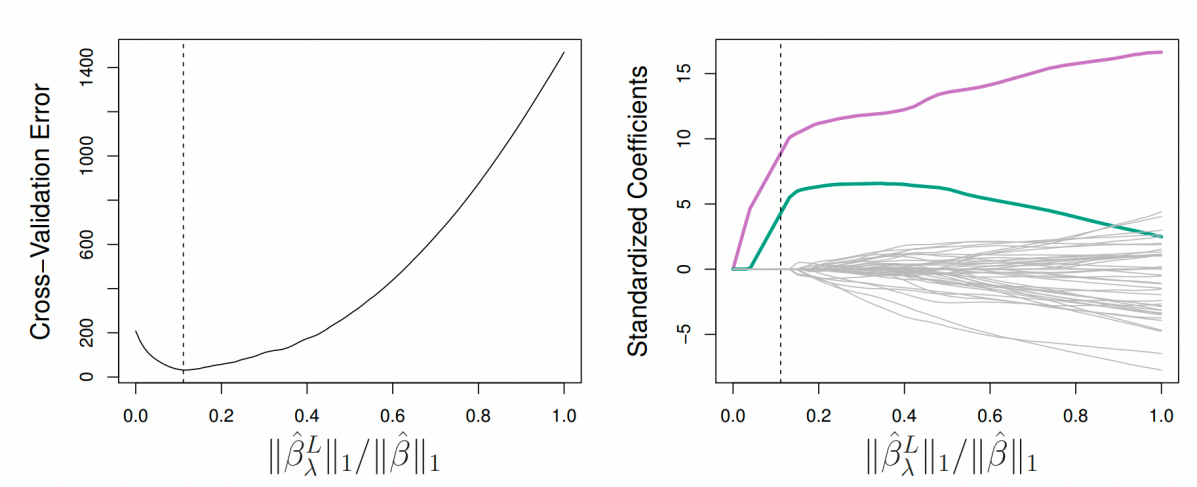


Figure: *

Ten-fold cross-validation for lasso, applied to the sparse simulated data set. Setting: $n = 50$ obs, 2 predictors (signal variables) related to the response (color), 43 unrelated predictors (noise variables) to the response (grey).

Dimension Reduction

- The subset selection and the shrinkage methods use the original predictors, $X_1, X_2, \cdots, X_p$.
- The **dimension reduction** methods **transform** the predictors and then fit a least squares model using the transformed variables.
- Let $Z_1, Z_2, \cdots, Z_M$ represent $M < p$ linear combinations of our original p predictors. That is,

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j \tag{1}$$

for some constants $\phi_{m1}, \cdots, \phi_{mp}$.

- Fit the linear regression model using ordinary least squares:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, i = 1, \cdots, n, \tag{2}$$

- Reduce the problem dimension from $p+1$ to $M+1$.
- Dimension reduction can outperform OLS if constants $\phi_{m1}, \cdots, \phi_{mp}$ are chosen wisely.
- Note that $\sum_{m=1}^{M} \theta_m z_{im} = \sum_{j=1}^{p} \beta_j x_{ij}$, where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj}. \tag{3}$$

  Model (2) is a special case of the original model, constraining β_j to form (3).
- This constraint may bias estimates but can be advantageous in the bias-variance tradeoff.

- Dimension reduction methods work in two steps:
  1. the transformed predictors $Z_1, Z_2, \cdots, Z_M$ are obtained.
  2. the model is fit using these M predictors.
- Two approaches for dimension reduction:
  - **principal components**
  - **partial least squares**

References

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.