Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 6: Linear Model Selection and Regularization (ISL ch. 6)

Instructor: Ma, Jun

Renmin University of China

October 28, 2021

Linear model and regression

- The linear model:

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

- Despite its simplicity, the linear model has advantages:
  - good interpretability;
  - often shows good predictive performance.
- Training data: $(Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n)$, where $X_i = \left(1, X_{1,i}, X_{2,i}, \cdots, X_{p,i}\right)^\top$.
- Linear regression coefficients:

$$\widehat{\beta} = \underset{b_0, b_1, ..., b_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1,i} - b_2 X_{2,i} - \cdots - b_p X_{p,i}\right)^2.$$

Estimation of test error

- $RSS = \sum_{i=1}^{n} \left(Y_i - X_i^\top \widehat{\beta}\right)^2$.
- The training error (regression MSE $\text{MSE}_{\text{Tr}} = RSS/n$) underestimates the test error (test MSE $\text{MSE}_{\text{Te}} = \text{E}\left[\left(Y_0 - X_0^\top \widehat{\beta}\right)^2\right]$, where (Y_0, X_0) is a future data point).
- The direct approach (cross-validation) estimates the test error by holding out a subset of the training observations from estimation, and then applying the method to those held out observations.
- Another approach uses a mathematical adjustment to the training error in order to estimate the test error.
  - Mallows' C_p.
  - Akaike information criterion (AIC) (for linear regression, AIC is proportional to C_p).
  - Bayesian information criterion (BIC).

Extensions of linear regression

- We focus on the regression model in this chapter and study two extensions of the linear regression method.
  - Model selection: best subset selection, stepwise selection.
  - Regularization/shrinkage: ridge regression, LASSO.
- Why these extensions?
  - Control the variance for prediction accuracy (especially when $p > n$), by shrinking coefficient estimates at the cost of increase in bias or using a smaller model with less regressors.
  - Model Interpretability: by setting some of coefficient estimates to zero (removing irrelevant variables), we can obtain a model that is more easily interpreted.
  - LASSO achieves both simultaneously.

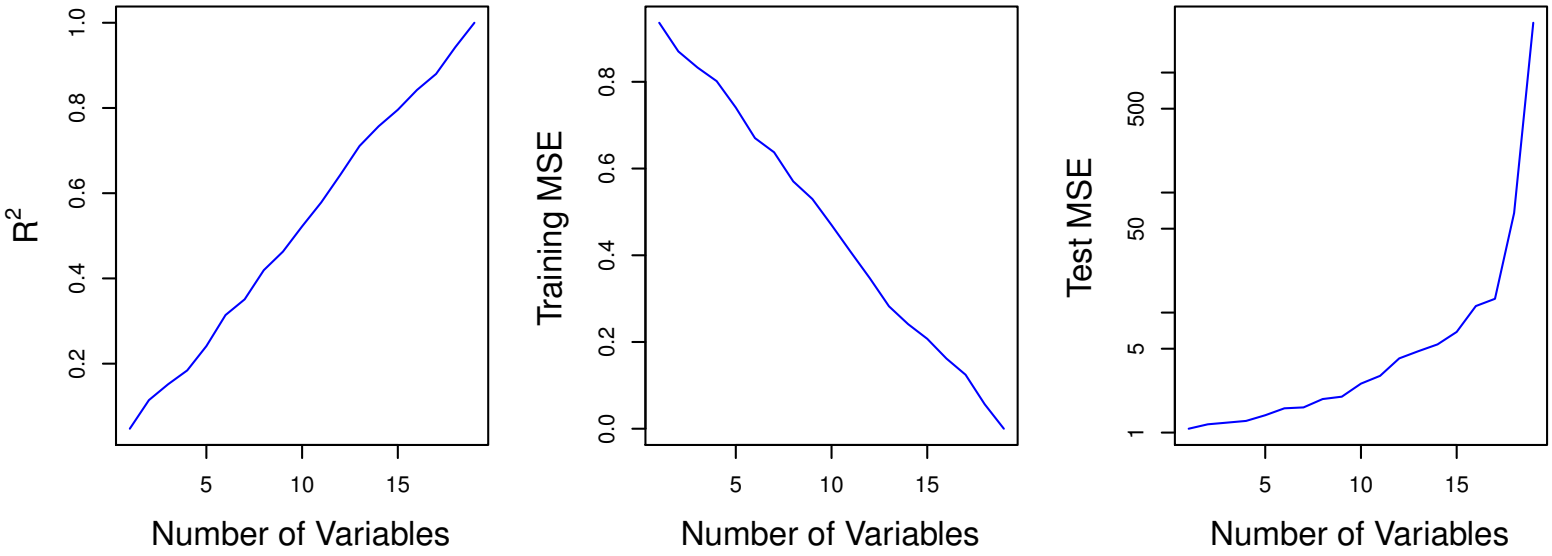
Model selection and shrinkage

- We have p predictors that we believe to be related to the response. When p is large, regression using all the predictors may have large variance.
- Model selection: identify a subset of the and then do linear regression using the reduced set of variables.
- Regularization/shrinkage: fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to OLS. This shrinkage/regularization has the effect of reducing variance and can also perform variable selection.
- Both model selection and shrinkage aim at finding a smaller model that has smaller variance.

High-dimensional data

- Most traditional statistical techniques are intended for the low-dimensional setting $n > p$.
  - Predict a patient's blood pressure using a data set with 200 patients and three predictors: age, gender and body mass index (BMI).
- Forward stepwise selection, ridge regression and LASSO work in high-dimensional settings $n < p$.
  - One might collect measurements for half a million "single nucleotide polymorphisms" (genetic characteristics) for inclusion in the predictive model for blood pressure. Then $n \approx 200$ and $p \approx 500000$.
  - A marketing analyst interested in understanding people's online shopping patterns could treat as features all of the search terms entered by users of a search engine. For a given user, each of the search terms is scored present (0) or absent (1), creating a large binary feature vector.

OLS cannot be performed in high-dimensional settings



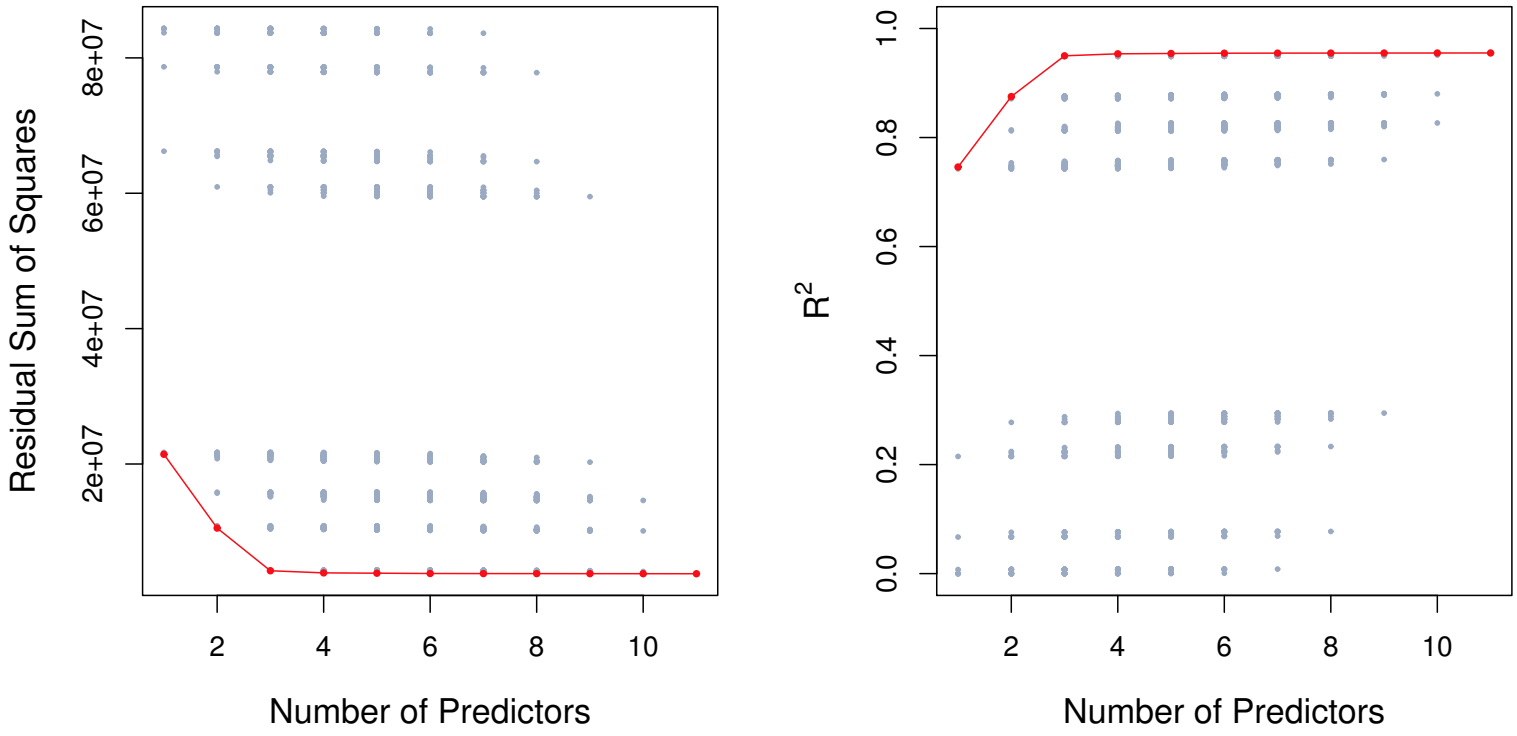
ISL Figure 6.23

- OLS will yield perfect fit to the data, such that the residuals are zero. Perfect fit certainly leads to overfitting.
- Simulated data with $n = 20$. OLS with 1 to 20 predictors, each of which was completely unrelated to the response.
- R^2 increases to 1 as the number of included predictors p increases. Test MSE becomes extremely large as p increases, because including the additional predictors leads to a vast increase in the variance of the coefficient estimates.

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 $k = 1, 2, ..., p$, find the model $\mathcal{M}_k$ with the largest R^2 in the collection of all $\left(\begin{array}{c} p \\ k \end{array} \right)$ models that contain exactly k predictors.
3. Select a single best model among $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ using cross-validation, C_p (AIC), BIC or adjusted R^2.
   - Note that RSS or R^2 cannot be used to select from among a set of models with different numbers of variables.

Example: the `Credit` data



ISL Figure 6.1

- Each model contains a subset of the ten predictors in the `Credit` data set.
- One categorical variable is represented by two dummy variables.
- Red frontier tracks the best model for a given number of predictors, according to RSS and R^2.

Stepwise selection

- In best subset selection, we fit all $\sum_{k=0}^{p} \left(\begin{array}{c} p \\ k \end{array} \right) = 2^p$ possible models. 2^p grows rapidly as p increases and is computationally infeasible if $p > 40$.
- Forward stepwise selection:
  - begins with a model containing no predictors, and then adds predictors to the model;
  - iteratively add one variable that gives the greatest additional improvement to the fit;
  - requires fitting just a total of $1 + p\,(p+1)\,/2$ models.
- Backward stepwise selection:
  - begins with the full model containing all p predictors;
  - iteratively removes the least useful predictor;
  - requires fitting just a total of $1 + p\,(p+1)\,/2$ models.

Forward stepwise selection

1. Let $\mathcal{M}_0$ denote the null model.
2. For $k = 0, 1, ..., p-1$, given the model $\mathcal{M}_k$ that has k predictors, let $\mathcal{M}_{k+1}$ denote the model with the largest R^2 in the collection of $p-k$ models with one more predictor.
3. Select a single best model among $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ using cross-validation, C_p (AIC), BIC or adjusted R^2.

- Forward stepwise selection is not guaranteed to select the best model in the collection with the same number of variables.
- Forward stepwise selection can be applied even in the high-dimensional setting where $n < p$.
- However, in the case of $n < p$, it is possible to construct sub-models $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_{n-1}$ only, since no unique solution to the least squares problem if $p \geq n$.

Example: the `Credit` data

# 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 four selected models for best subset selection and forward stepwise selection on the `Credit` data set.
- The first three models are identical but the fourth models differ.

Backward stepwise selection

1. Let $\mathcal{M}_p$ denote the full model, which contains all p predictors.
2. For $k = p, p-1, ..., 1$, given the model $\mathcal{M}_k$ that has k predictors, consider all k models that contain all but one of the predictors in $\mathcal{M}_k$. Let $\mathcal{M}_{k-1}$ denote the model with the highest R^2 among these k models.
3. Select a single best model among $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ using cross-validation, C_p (AIC), BIC or adjusted R^2.

C_p, AIC, BIC and adjusted R^2

- These techniques adjust the training error for the model size to obtain an estimate of the test error.
- Mallow's C_p:

$$C_p = \frac{1}{n}\left(RSS + 2d\widehat{\sigma}^2\right),$$

  where is the total number of parameters and $\widehat{\sigma}^2$ is an estimate of the variance of the error ϵ.
- $\widehat{\sigma}^2$ is estimated using the full model containing all predictors.
- C_p adds a penalty of $2d\widehat{\sigma}^2$ to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error.
- C_p statistic is an estimate of the test error and tends to take on a small value for models with a low test error.

- The Akaike information criterion (AIC) is defined for a large class of models fit by maximum likelihood:

$$\text{AIC} = -2\log L + 2 \cdot d,$$

where L is the maximized value of the likelihood function for the estimated model.

- For the regression model with a normally distributed ϵ,

$$\text{AIC} = \frac{1}{\widehat{\sigma}^2 n}\left(RSS + 2d\widehat{\sigma}^2\right),$$

which is proportional to C_p.

- Bayes information criterion:

$$\text{BIC} = -2\log L + \log(n) \cdot d,$$

- For the regression model with a normally distributed ϵ,

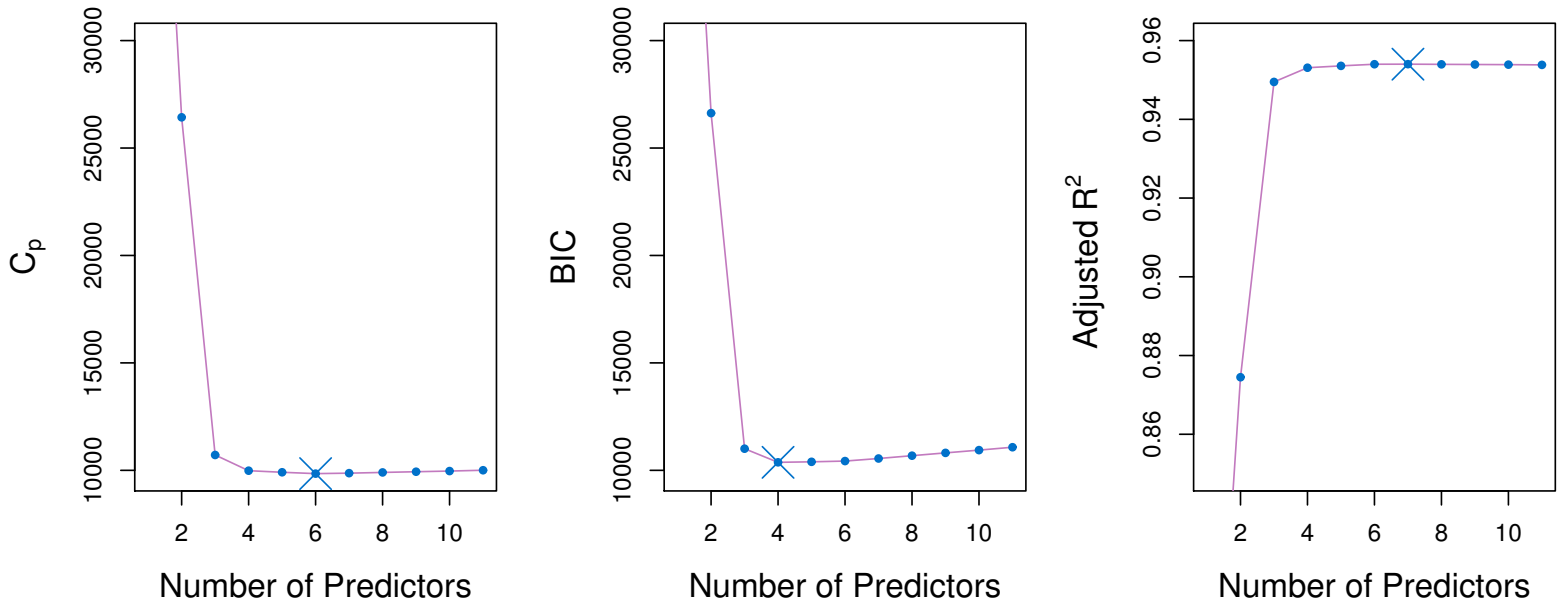
$$\text{BIC} = \frac{1}{\widehat{\sigma}^2 n}\left(RSS + \log(n)\, d\widehat{\sigma}^2\right).$$

- Since $\log(n) > 2$ for any $n > 7$, the BIC generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p or AIC.

- For a least squares model with d variables, the adjusted R^2 is calculated as

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)},$$

  where TSS is the total sum of squares.
- While RSS always decreases as the number of variables in the model increases, $RSS/(n-d-1)$ may increase or decrease, due to the presence of d.
- C_p, AIC and BIC all have rigorous theoretical justification. Adjusted R^2 is not as well motivated in statistical theory.

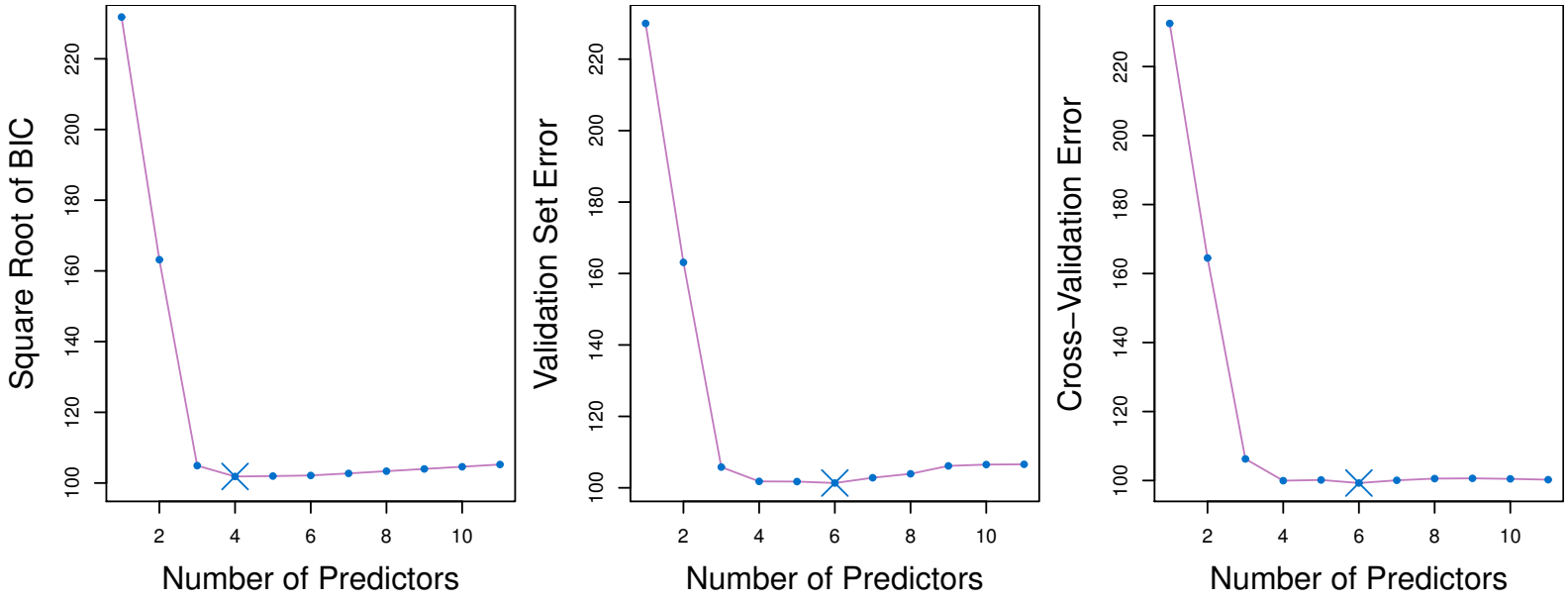


ISL Figure 6.2

- C_p, BIC and adjusted R^2 are shown for the best models of each size for the `Credit` data set.

Cross-Validation

- C_p, AIC and BIC are not appropriate in high-dimensional settings, because $\widehat{\sigma}^2 = 0$.
- As a alternative to C_p, BIC and AIC, we can directly estimate the test error using the validation set and cross-validation.
- Cross-validation can be applied in high-dimensional settings.
- Cross-validation is more computationally burdensome.
- We can compute the validation set error or the cross-validation error for each of $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$, and then select the model for which the resulting estimated test error is smallest.
- Cross-validation provides a direct estimate of the test error and its theoretical justification requires fewer assumptions about the true underlying model.



ISL Figure 6.3

- BIC, validation set errors, and cross-validation errors on the `Credit` data, for the best d-variable model.
- The validation and cross-validation methods both result in a six-variable model.
- The four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

One-standard-error rule

- For a given number of predictors, the test MSE is an unknown parameter.
- We first calculate the one standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one error standard error of the lowest point on the curve.
- If a set of models appear to be more or less equally good, then we should choose the simplest model.

Ridge regression

- Linear regression coefficients:

$$\begin{aligned}\widehat{\beta} &= \underset{b_0,b_1,...,b_p}{\operatorname{argmin}} \sum_{i=1}^{n}\left(Y_i - b_0 - \sum_{k=1}^{p} b_k X_{k,i}\right)^2 \\ &= \underset{b_0,b_1,...,b_p}{\operatorname{argmin}}\ RSS\left(b_0, b_1, ..., b_p\right).\end{aligned}$$

- Ridge regression:

$$\widehat{\beta}^R_\lambda = \underset{b_0,b_1,...,b_p}{\operatorname{argmin}}\ RSS\left(b_0, b_1, ..., b_p\right) + \lambda \sum_{k=1}^{p} b_k^2,$$

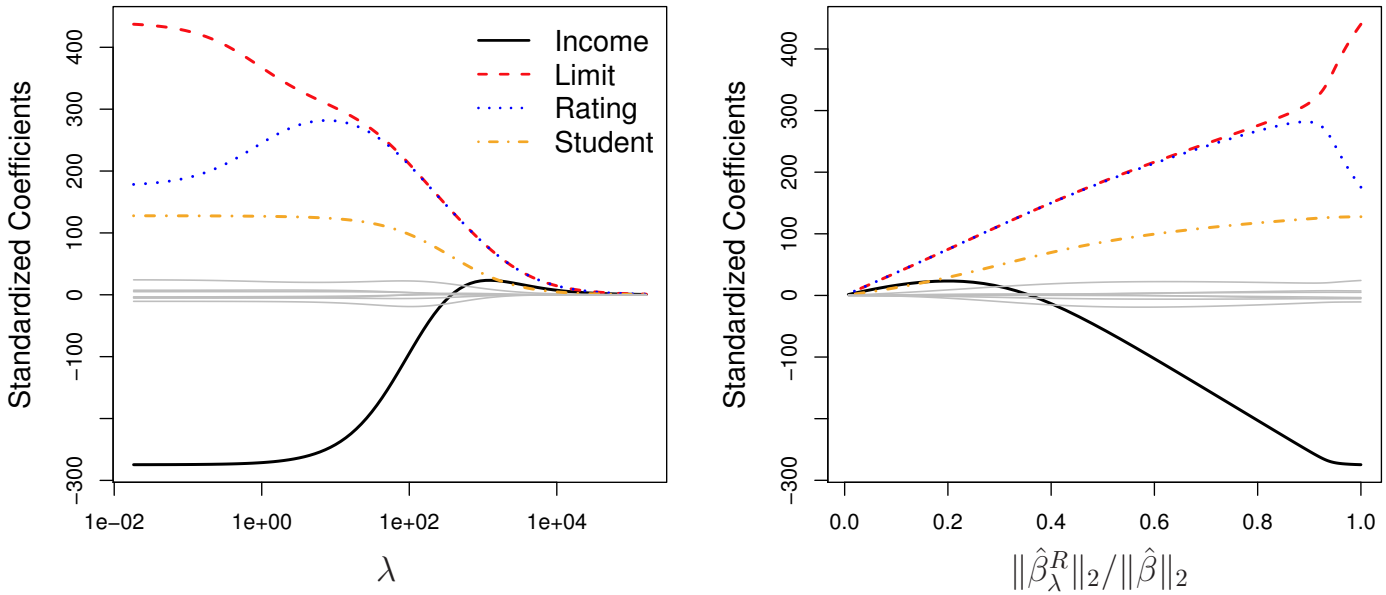
  where $\lambda \geq 0$ is a tuning parameter selected by the user.
- The ridge regression is a constrained least squares: for every λ, there exists $s > 0$ such that $\widehat{\beta}^R_\lambda$ solves

$$\min_{b_0,b_1,...,b_p} RSS\left(b_0, b_1, ..., b_p\right) \text{ subject to } \sum_{j=1}^{p} b_j^2 \leq s.$$

  And λ and s are inversely related.

- Ridge regression seeks coefficient estimates that fit the data well, by making $RSS\left(b_0, b_1, ..., b_p\right)$ small.
- The shrinkage penalty $\lambda \sum_{k=1}^{p} b_k^2$ is small when $b_1, b_2, ..., b_k$ are close to zero. The tuning parameter λ controls the relative impact of these two terms.
- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce OLS.
- $\widehat{\beta}_{\lambda}^{R} = \left(\widehat{\beta}_{\lambda,0}^{R}, \widehat{\beta}_{\lambda,1}^{R}, ..., \widehat{\beta}_{\lambda,p}^{R}\right)$ depends on λ. Optimal λ (test MSE minimizing) depends on the underlying data generating mechanism. Use cross-validation to select λ.
- Note that we do not want to shrink the intercept.

Example: the `Credit` data



ISL Figure 6.4

- Left: ridge regression coefficient estimates for each of the ten variables, plotted as a function of λ.
- Right: ridge regression coefficient estimates for each of the ten variables, plotted as a function of $\left\|\widehat{\beta}^R_\lambda\right\|_2 / \left\|\widehat{\beta}\right\|_2$

$(\left\|(x_1, x_2, ..., x_k)^\top\right\|_2 = \sqrt{\sum_{j=1}^k x_j^2})$.

Scaling of predictors

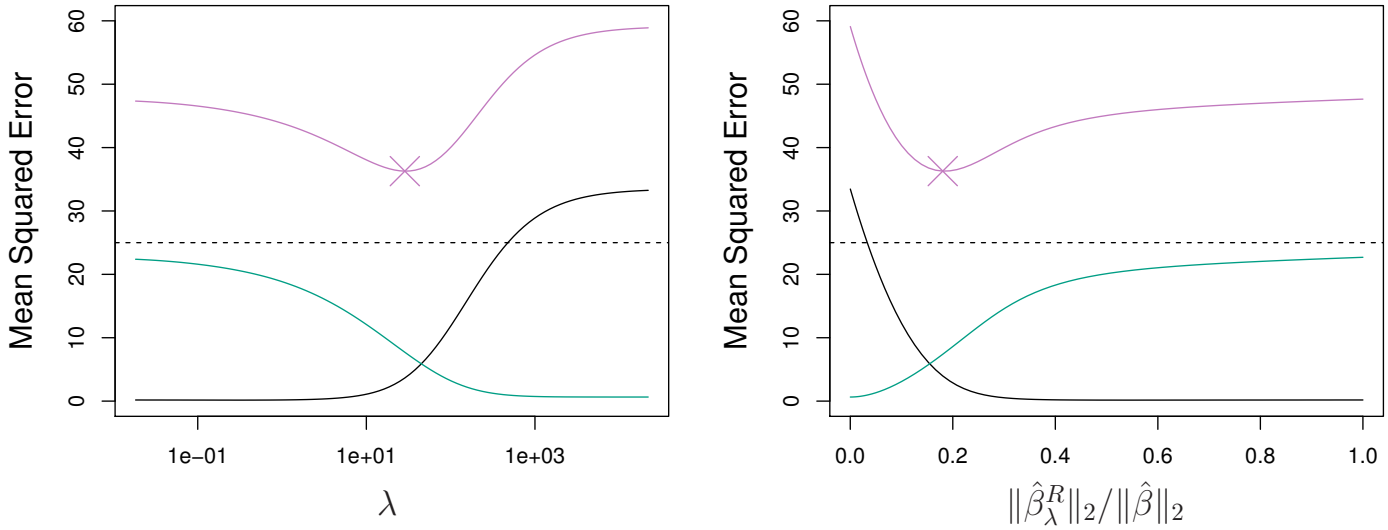
- The standard OLS estimates are scale equivariant: multiply $X_{j,i}$ by a constant c leads to a scaling of the OLS coefficient $\widehat{\beta}_j$ by a factor of $1/c$. regardless of how the j-th predictor is scaled, $X_{j,i}\widehat{\beta}_j$ will remain the same. In contrast, for ridge regression, $X_{j,i}\widehat{\beta}^R_{\lambda,j}$ will not remain the same and depend on λ.
- For instance, consider the income variable, which is measured in dollars. One could reasonably have measured income in thousands of dollars.
- It is best to apply ridge regression after standardizing the predictors:

$$\tilde{X}_{j,i} = \frac{X_{j,i}}{\sqrt{n^{-1}\sum_{i=1}^{n}\left(X_{j,i} - \overline{X}_j\right)^2}}.$$

- All of the standardized predictors will have a standard deviation of one.

Why does ridge regression improve over OLS?

- As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.
- Sacrificing variance a little bit may lead to substantial improvement in bias.
- In general, in situations where the relationship between the response and the predictors is close to linear, OLS will have low bias but may have high variance. In particular, when the number of variables p is almost as large as the number of observations n, the OLS will be extremely variable.
- If $p > n$, OLS does not have a unique solution, whereas ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance.
- Ridge regression also has substantial computational advantages over best subset selection. In contrast, for any fixed value of λ, ridge regression only fits a single model.



ISL Figure 6.5

- Simulated data with $n = 50$ observations, $p = 45$ predictors, all having nonzero coefficients.
- Black: square bias; green: variance; test MSE: purple.
- $\lambda = 0$: OLS.

LASSO

- Unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.
- This problem can create a challenge in model interpretation when p is quite large.
- In the `Credit` data, it appears that the most important variables are `income`, `limit`, `rating`, and `student`. So we might wish to build a model including just these predictors. However, ridge regression will always generate a model involving all ten predictors.
- The LASSO is a relatively recent alternative to ridge regression: LASSO will always generate a sparse model that involves only a subset of the variables..
- For some tuning parameter $\lambda \geq 0$,

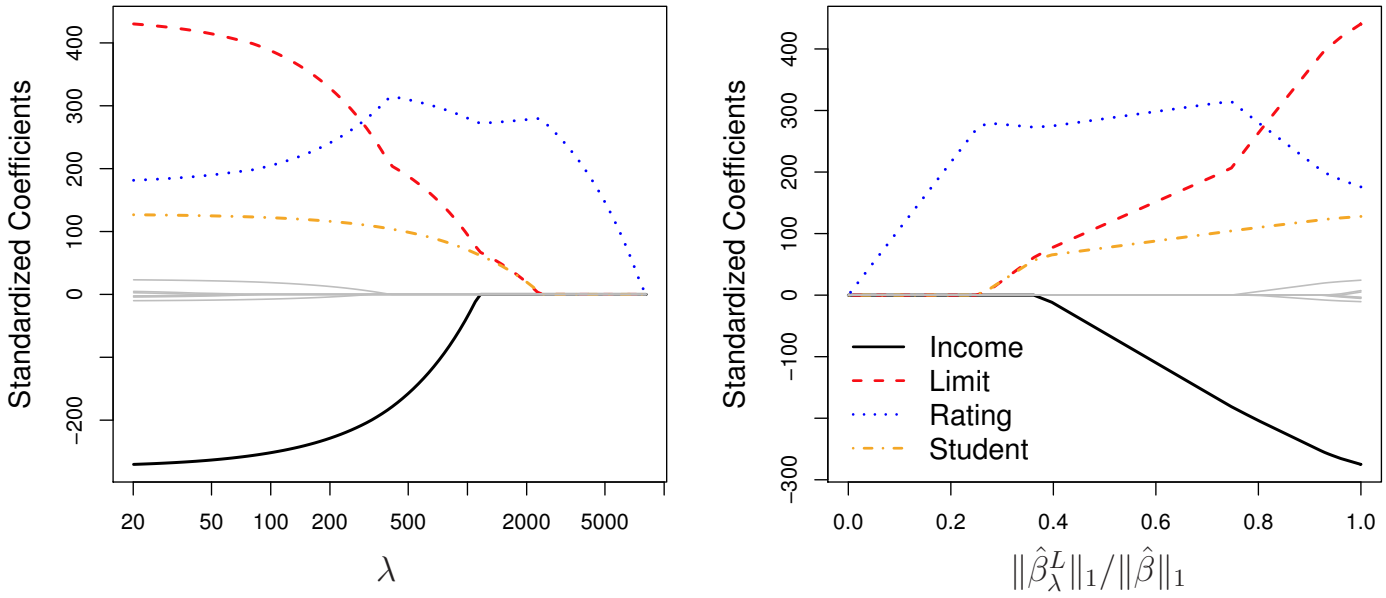
$$\widehat{\beta}_{\lambda}^{L} = \underset{b_0, b_1, ..., b_p}{\operatorname{argmin}} \; RSS\left(b_0, b_1, ..., b_p\right) + \lambda \sum_{k=1}^{p} |b_k| \, .$$

- LASSO is a constrained least squares: for every λ, there exists $s > 0$ such that $\widehat{\beta}_{\lambda}^{L}$ solves

$$\min_{b_0, b_1, ..., b_p} RSS\left(b_0, b_1, ..., b_p\right) \text{ subject to } \sum_{j=1}^{p} \left|b_j\right| \leq s.$$

- In comparison to ridge regression, LASSO uses an $\mathscr{L}^1$ penalty with $\left\|(x_1, x_2, ..., x_k)^\top\right\|_1 = \sum_{j=1}^{k} \left|x_j\right|$, while ridge uses an $\mathscr{L}^2$ penalty with $\left\|(x_1, x_2, ..., x_k)^\top\right\|_2 = \sqrt{\sum_{j=1}^{k} x_j^2}$.
- As with ridge regression, LASSO shrinks the coefficient estimates towards zero.
- However, in the case of the LASSO, the $\mathscr{L}^1$ penalty has the effect of forcing some of the coefficient estimates to be exactly 0 when λ is large.
- LASSO performs variable/model selection and estimation in one step.
- We do cross-validation to select λ.

Example: the `Credit` data



ISL Figure 6.5

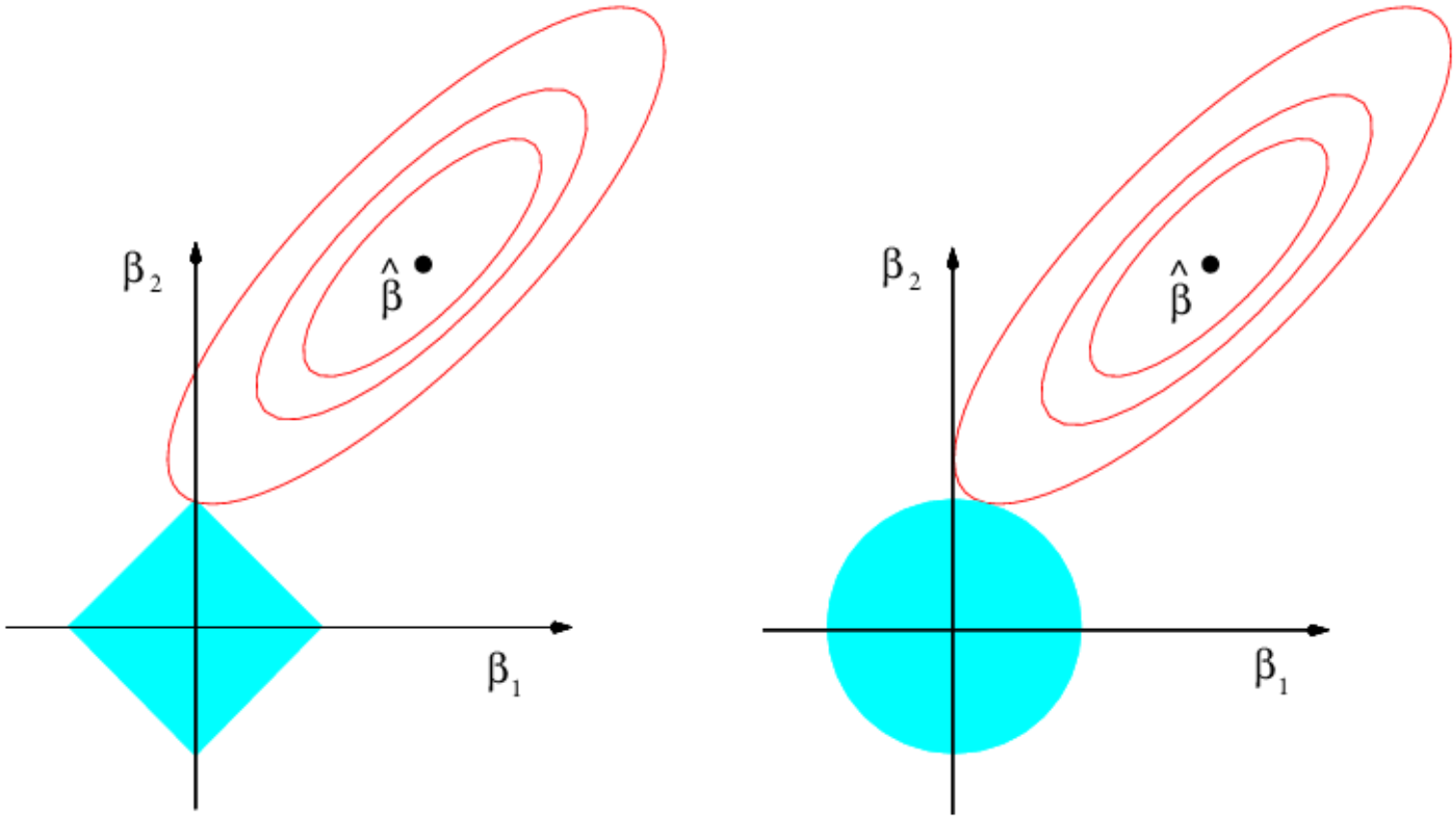
- Depending on the value of λ, LASSO can produce a model involving any number of variables.
- In contrast, ridge regression will always include all of the variables.

The variable selection property of LASSO

- Why is it that the LASSO, unlike ridge regression, results in coefficient estimates that are exactly 0?
- LASSO and ridge regression coefficient estimates solve the problems

$$\text{Ridge:} \min_{b_0, b_1, ..., b_p} RSS\left(b_0, b_1, ..., b_p\right) \text{ subject to } \sum_{j=1}^{p} b_j^2 \leq s$$

$$\text{LASSO:} \min_{b_0, b_1, ..., b_p} RSS\left(b_0, b_1, ..., b_p\right) \text{ subject to } \sum_{j=1}^{p} \left|b_j\right| \leq s.$$



ISL: Figure 6.7

- $\hat{\beta}$: the OLS estimator.
- Red contours: constant residual sum of squares with respect to (b_1, b_2).
- Left shaded rectangle: $\{(b_1, b_2) : |b_1| + |b_2| \leq s\}$; right shaded rectangle: $\left\{(b_1, b_2) : b_1^2 + b_2^2 \leq s\right\}$.

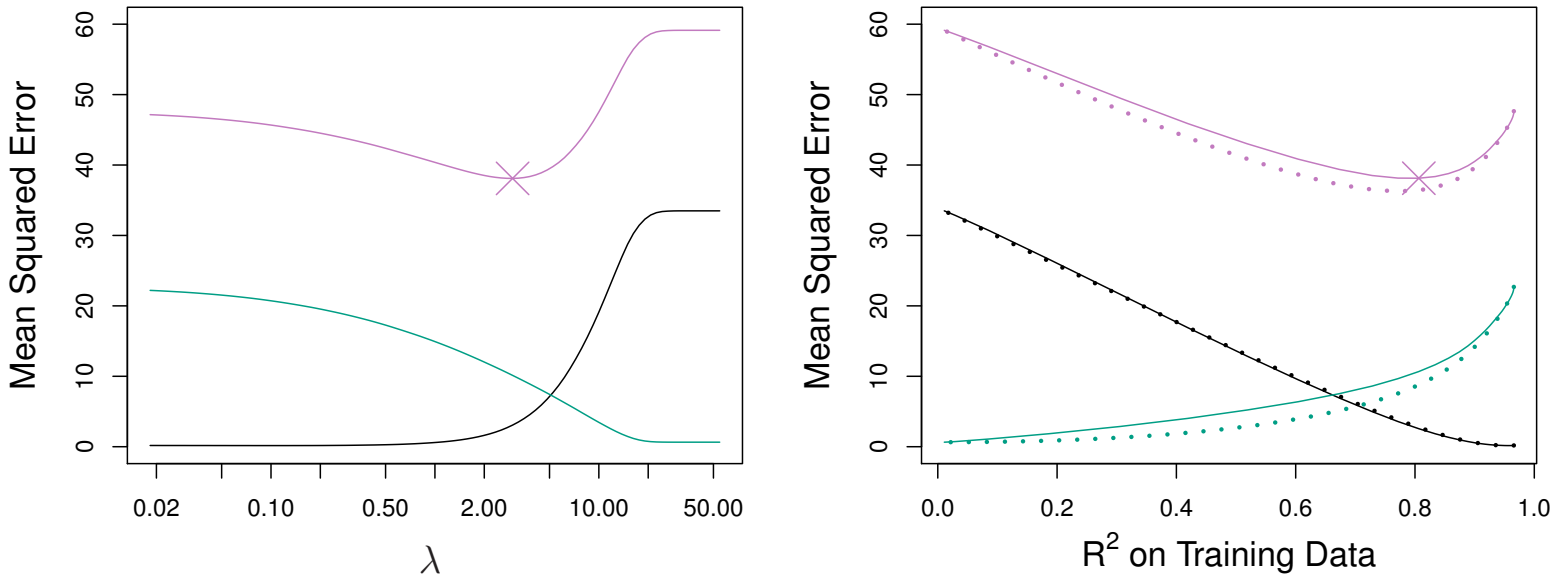
Alternative formulation of best subset selection

- The best subset selection amounts to solving

$$\min_{b_0, b_1, ..., b_p} RSS\left(b_0, b_1, ..., b_p\right) \text{ subject to } \sum_{j=1}^{p} 1\left(b_j \neq 0\right) \leq s$$

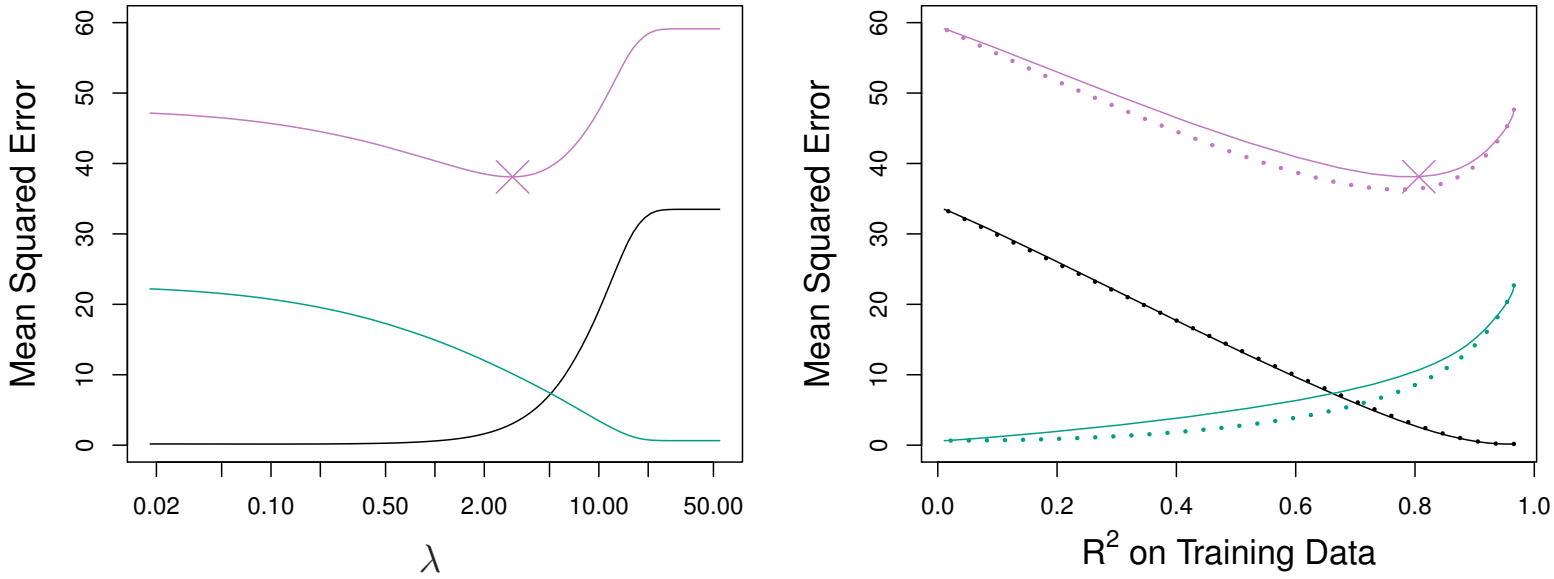
  where s is an integer tuning parameter. $1\left(b_j \neq 0\right)$ takes on value of 1 if $b_j \neq 0$ and 0 otherwise.
- We find a set of coefficient estimates such that RSS is as small as possible, subject to the constraint that no more than s coefficients can be nonzero with $0 \leq s \leq p$.
- This is computationally infeasible when p is large, since it requires considering all $\begin{pmatrix} p \\ s \end{pmatrix}$ models containing s predictors.
- Ridge regression and LASSO are computationally feasible alternatives to best subset selection that replace the intractable constraint with forms that are much easier to solve.

Comparing the LASSO and ridge regression



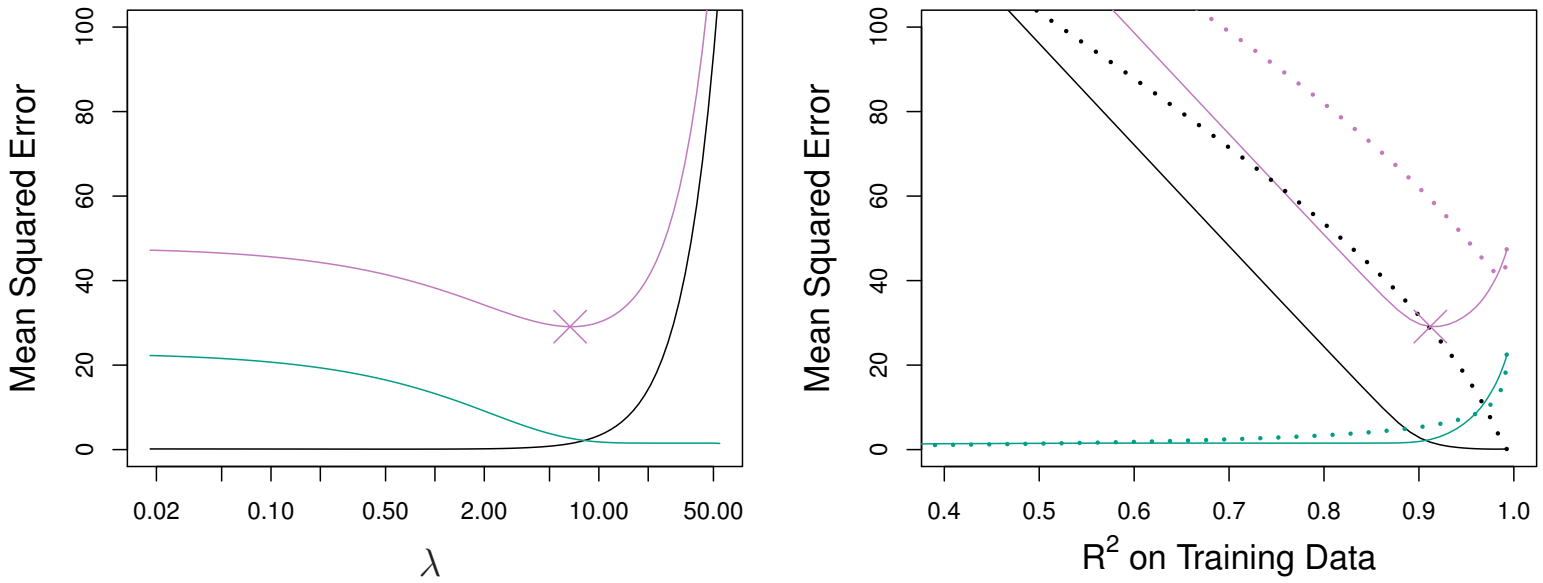
ISL Figure 6.8

- Compared with ridge regression, LASSO produces simpler and more interpretable models that involve only a subset of the predictors. However, which method leads to better prediction accuracy?
- Left: square bias (black), variance (green) and test MSE (purple) for LASSO.



ISL Figure 6.8

- Right: Comparison of squared bias, variance, and test MSE between LASSO (solid) and ridge (dotted).
- Both are plotted against their R^2 on the training data. This is another useful way to index models, and can be used to compare models with different types of regularization.
- The minimum MSE of ridge regression is slightly smaller than that of the LASSO.



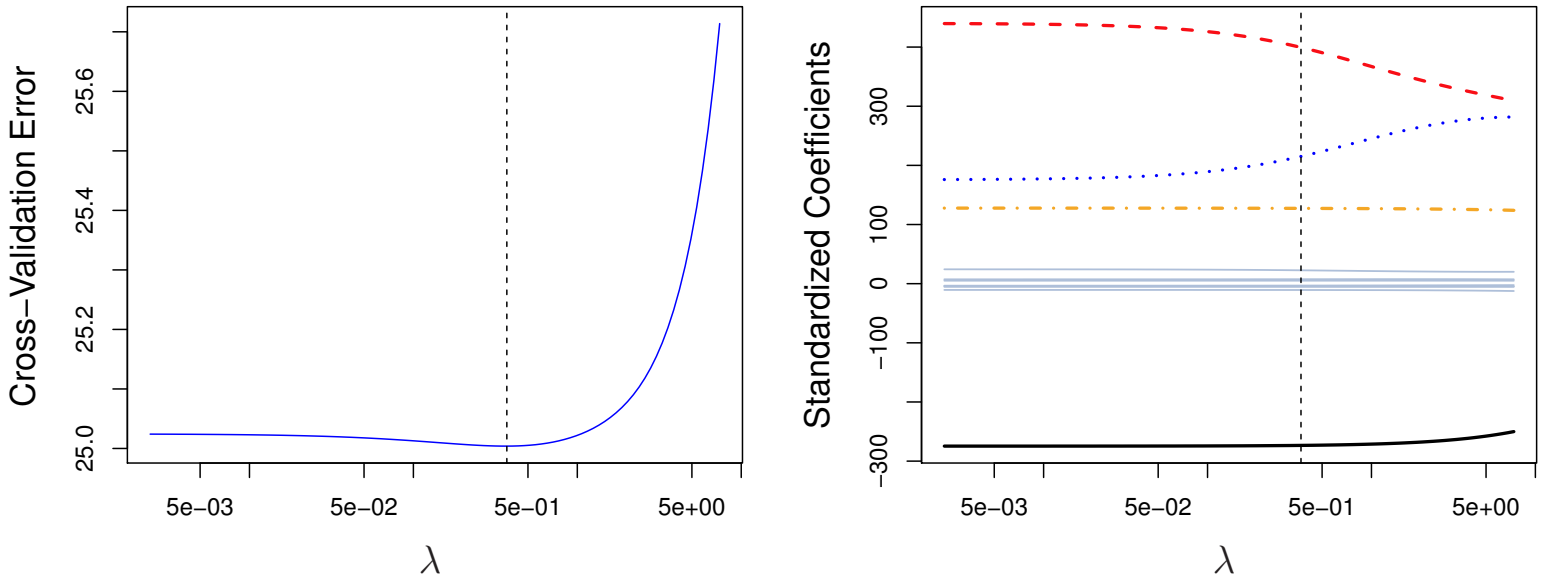
ISL Figure 6.9

- Right: Comparison of squared bias, variance, and test MSE between LASSO (solid) and ridge (dotted).
- In the simulated data, only two predictors are related to the response.

- LASSO implicitly assumes that a number of the coefficients truly equal zero.
- Ridge regression outperforms the lasso in terms of prediction error when all predictors are related to the response.
- If the true model that generates the data is sparse (a relatively small number of nonzero coefficients), LASSO tends to outperform ridge regression in terms of bias, variance, and MSE.
- The number of predictors that is related to the response is never known for real data sets. Cross-validation can be used in order to determine which approach is better on a particular data set.

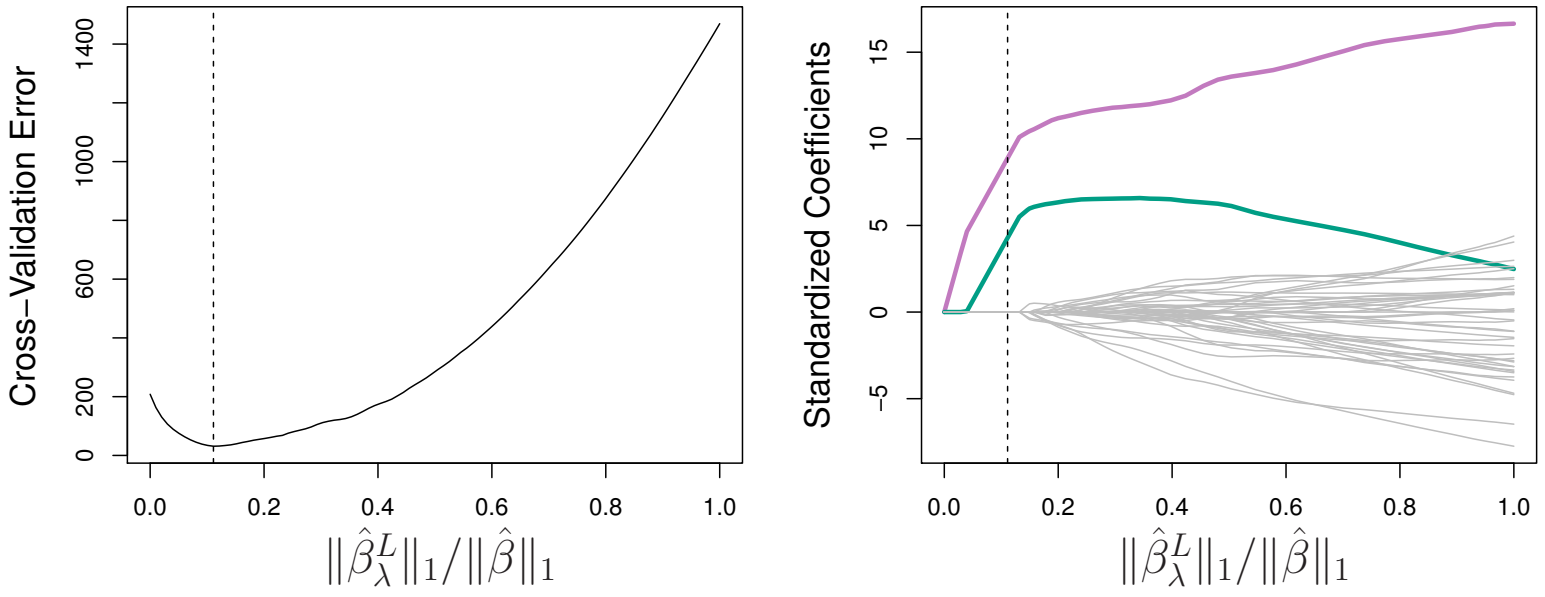
Selecting the tuning parameter

- Implementing ridge regression and the lasso requires a method for selecting a value for the tuning parameter λ.
- We choose a grid of λ values, and compute the cross-validation error 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.



ISL Figure 6.12

- Left: leave-one-out cross-validation errors that result from applying ridge regression to the `Credit` data set with various values of λ.
- Right: The coefficient estimates as a function of λ.
- Vertical dashed lines: λ selected by cross-validation.



ISL Figure 6.13

- 10-fold cross-validation applied to the LASSO fits on the sparse simulated data with only two nonzero coefficients.
- Cross-validation together with LASSO has correctly identified the two variables with nonzero coefficients.
- OLS displayed on the far right of the right-hand panel assigns a large coefficient estimate to only one of the two variables.