Statistical Learning for Data Science

Lecture 19

唐晓颖
电子与电气工程系
南方科技大学

May 24, 2023

Subset Selection

- Hybrid approaches

  Another alternative: hybrid versions of forward and backward stepwise selection, in which variables are added to the model sequentially, in analogy to forward selection. However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.

  Such an approach attempts to more closely mimic best subset selection while retaining the computational advantages of forward and backward stepwise selection.

Subset Selection

- Limitations and potential issues associated

  - Potential omission of important variables: Due to the stepwise addition or elimination strategy, stepwise selection may overlook important variables. Especially when variables are correlated, certain variables may be excluded during the selection process, resulting in model deficiencies.

  - Computational complexity: Stepwise selection can become computationally intensive when dealing with a large number of predictor variables. It requires trying various combinations, which increases the time and computational resources needed for model selection.

  - Multiple comparisons: Stepwise selection involves multiple comparisons at each step when selecting variables based on evaluation criteria. This can lead to overfitting, where variables are chosen based on their performance on a specific dataset but may not generalize well to other datasets.

Choosing the Optimal Model

In best subset selection, forward stepwise selection, and backward stepwise selection, the last step always involves selecting the **best** model.

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.

We need a way to determine being the best!

Choosing the Optimal Model

- The model containing all of the predictors will always have the smallest RSS and the largest R^2, 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. Recall that training error is usually a poor estimate of test error.
- Therefore, RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

Choosing the Optimal Model

In order to select the best model with respect to test error, we need to estimate this test error. There are two common approaches.

1. We can *indirectly* estimate the test error by making an **adjustment** to the training error to account for the bias due to overfitting.

2. We can *directly* estimate the test error, using either a validation set approach or a cross-validation approach, as we previously discussed.

Choosing the Optimal Model

- C_p, AIC, BIC, and Adjusted R^2

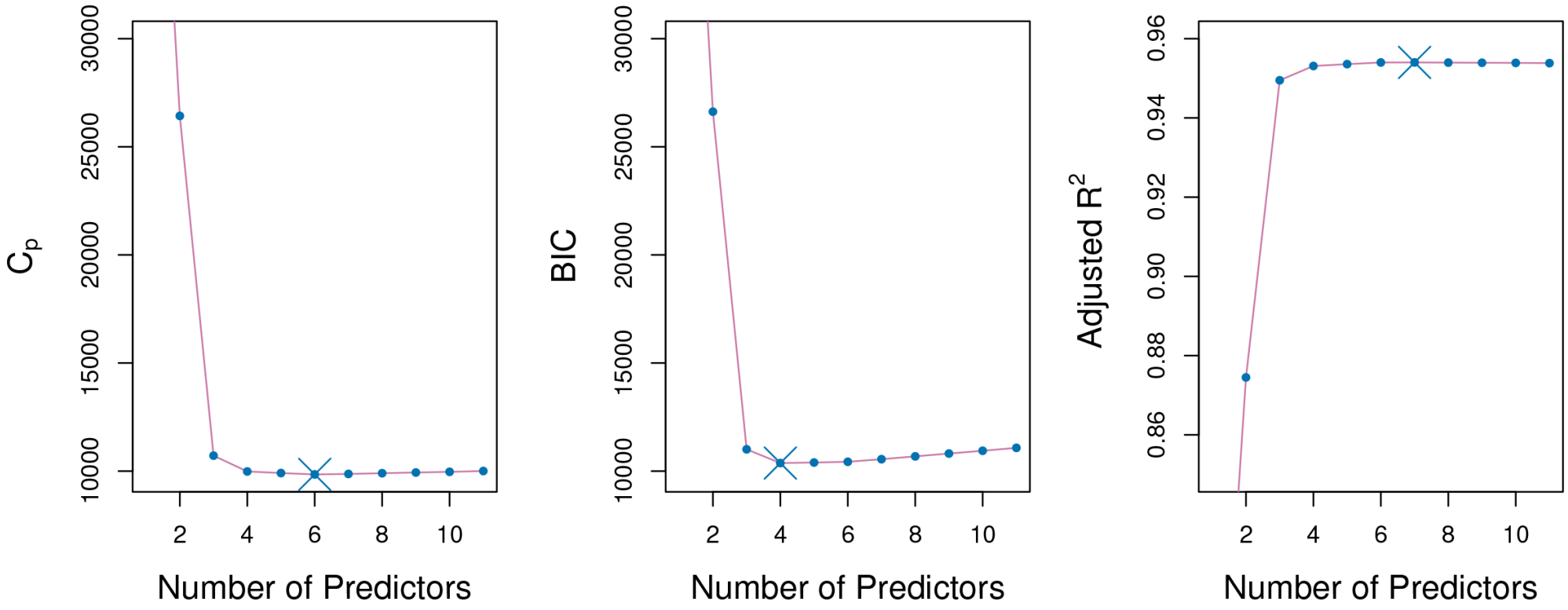
There are a number of techniques for ***adjusting*** the training error for the model size, and can be used to select among a set of models with different numbers of variables.

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

Choosing the Optimal Model

- C_p, AIC, BIC, and Adjusted R^2

Credit data example



Choosing the Optimal Model

- C_p

  For a fitted least squares model containing d predictors, the C_p estimate of test MSE is computed using the equation

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

  d --- the total number of parameters used.

  $\hat{\sigma}^2$ --- an estimate of the variance of the error ε associated with each response measurement.

Choosing the Optimal Model

- C_p

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

  - Essentially, the C_p statistic adds a penalty of $2d\hat{\sigma}^2$ to the training RSS in order to adjust for the fact that the training error tends to under estimate the test error.
  - The penalty $2d\hat{\sigma}^2$ increases as the number of predictors in the model increases; this is intended to adjust for the corresponding decrease in training RSS.

Choosing the Optimal Model

- C_p

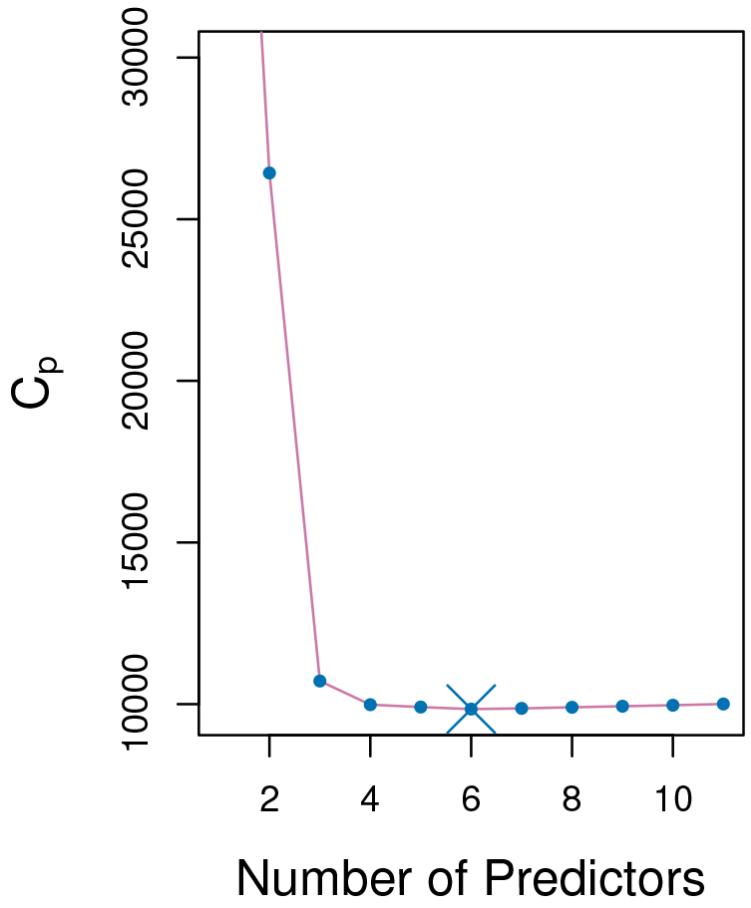
$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

  - If $\hat{\sigma}^2$ is an unbiased estimate of σ^2, then C_p is an unbiased estimate of test MSE (**TA**).

  - Therefore, the C_p statistic tends to take on a small value for models with a low test error, so when determining which of a set of model is best, we choose the model with the lowest C_p value.

Choosing the Optimal Model

- C_p



For the credit data example, C_p select the six-variable model containing the predictors income, limit, rating, cards, age, and student status.

Choosing the Optimal Model

- AIC

The AIC criterion is defined for a large class of models fit by maximum likelihood:

$$\mathrm{AIC} = -2\log L + 2d$$

L --- the maximized value of the likelihood function for the estimated model.

Choosing the Optimal Model

- AIC

  In the case of the linear model with Gaussian errors, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent.

  Prove this. （HW）

Choosing the Optimal Model

- BIC

For a fitted least squares model containing d predictors, the BIC is, up to irrelevant constants, given by

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

d --- the total number of parameters used.

$\hat{\sigma}^2$ --- an estimate of the variance of the error ε associated with each response measurement.

Choosing the Optimal Model

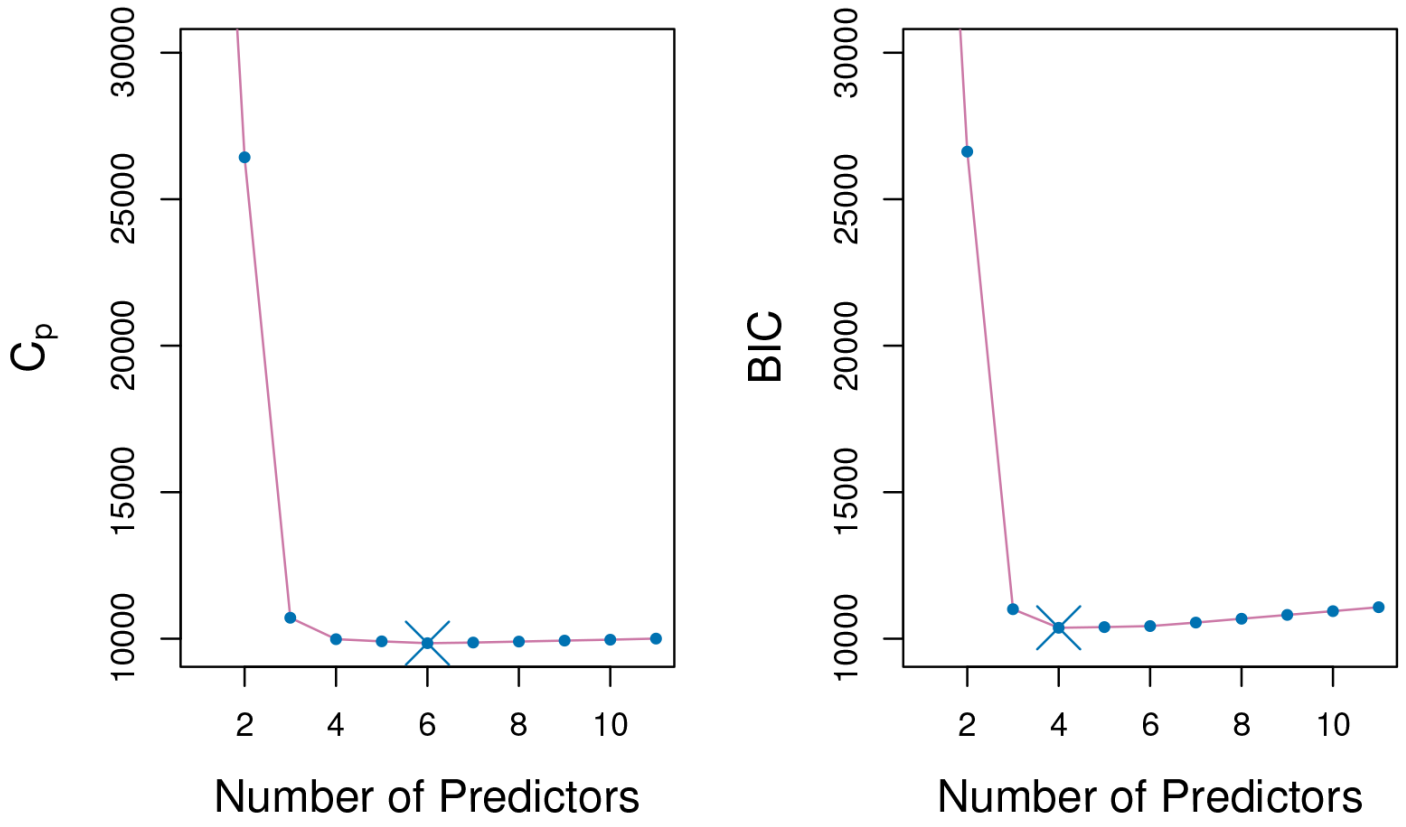
- BIC

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

  - Like C_p , the BIC will tend to take on a small value for a model with a low test error, and so generally we select the 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 observations.
  - Since log n > 2 for any 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 .

Choosing the Optimal Model

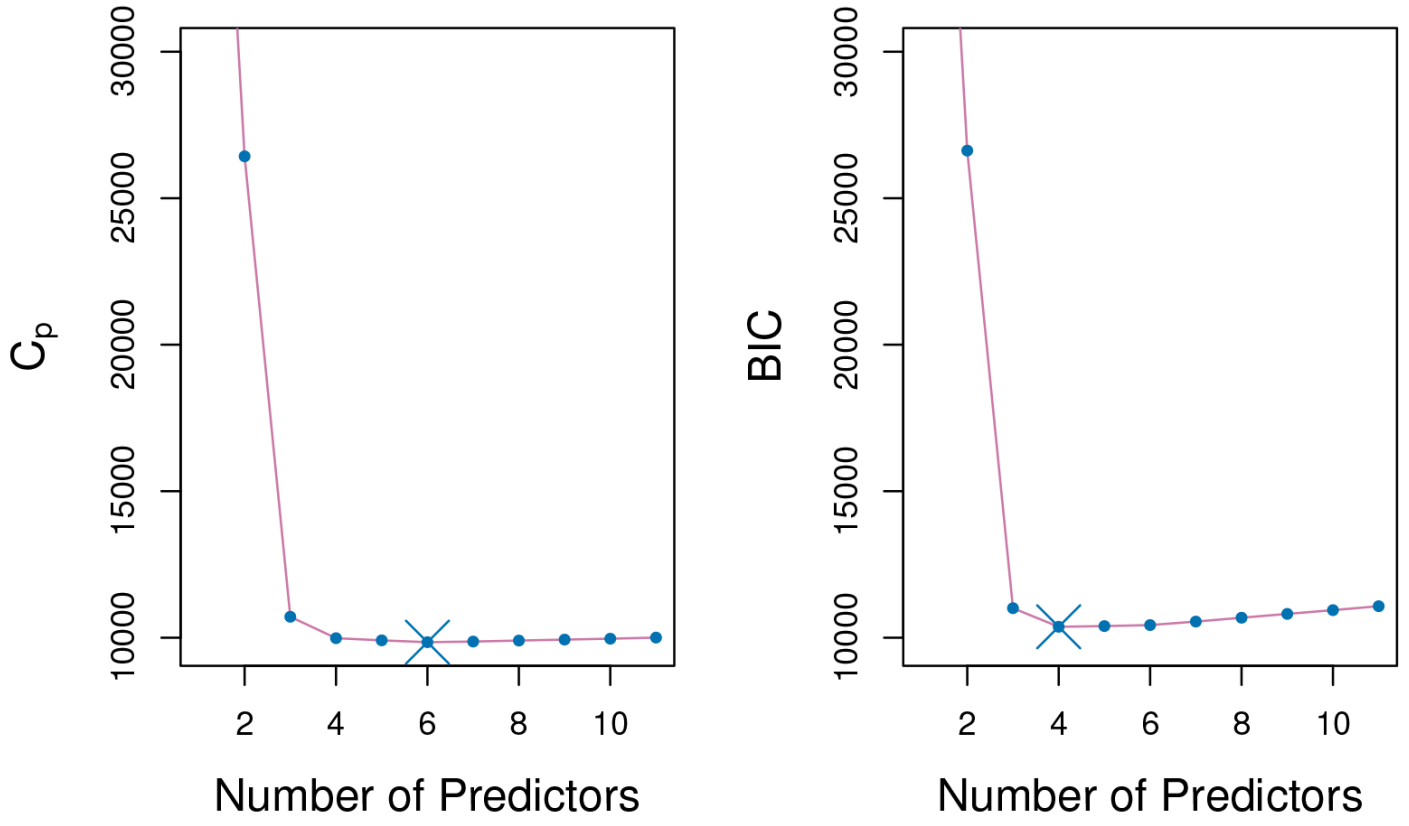
- BIC



- C_p chooses income, limit, rating, cards, age, and student status.
- BIC chooses income, limit, cards, and student status.

Choosing the Optimal Model

- BIC



In this case, the curves are very flat and so there does not appear to be much difference in accuracy between the four-variable and six-variable model.

Choosing the Optimal Model

- Adjusted R^2

The usual R^2 is defined as:

$$R^2 = 1 - \frac{RSS}{TSS},\ TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Since RSS always decreases as more variables are added to the model, the R^2 always increases as more variables are added to the model.

Choosing the Optimal Model

- Adjusted R^2 $\quad R^2 = 1 - \frac{RSS}{TSS},\ TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

  For a fitted least squares model containing d predictors, the $\text{Adjusted } R^2$ is calculated as

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

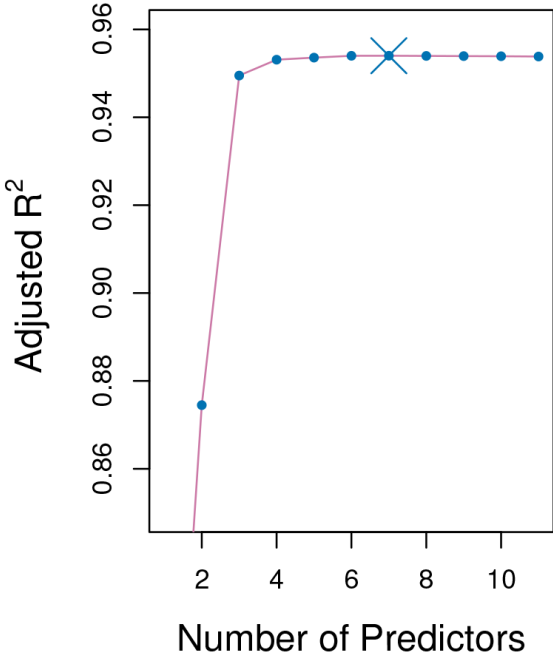
  - Unlike C_p , AIC, and BIC, for which a small value indicates a model with a low test error, a large value of $\text{Adjusted } R^2$ indicates a model with a small test error.
  - Maximizing the $\text{Adjusted } R^2$ is equivalent to minimizing $RSS/(n-d-1)$.

Choosing the Optimal Model

- Adjusted R^2

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

  - Unlike the R^2 statistic, the $\text{Adjusted } R^2$ statistic pays a price for the inclusion of unnecessary variables in the model.



For the credit data example, $\text{Adjusted } R^2$ select the seven-variable model containing the predictors income, limit, rating, cards, age, student status, and gender.

Choosing the Optimal Model

- Validation and Cross-Validation

  - Each of the model selection procedures returns a sequence of models M_k indexed by model size $k = 0, 1, ..., p$. Our job here is to select the optimal $\hat{k}$. Once selected, we will return model $M_{\hat{k}}$.

  - We compute the validation set error or the cross-validation (least-one-out or k-fold) error for each model M_k under consideration, and then select the k for which the resulting estimated test error is smallest.

  - This procedure has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

Choosing the Optimal Model

- Validation and Cross-Validation

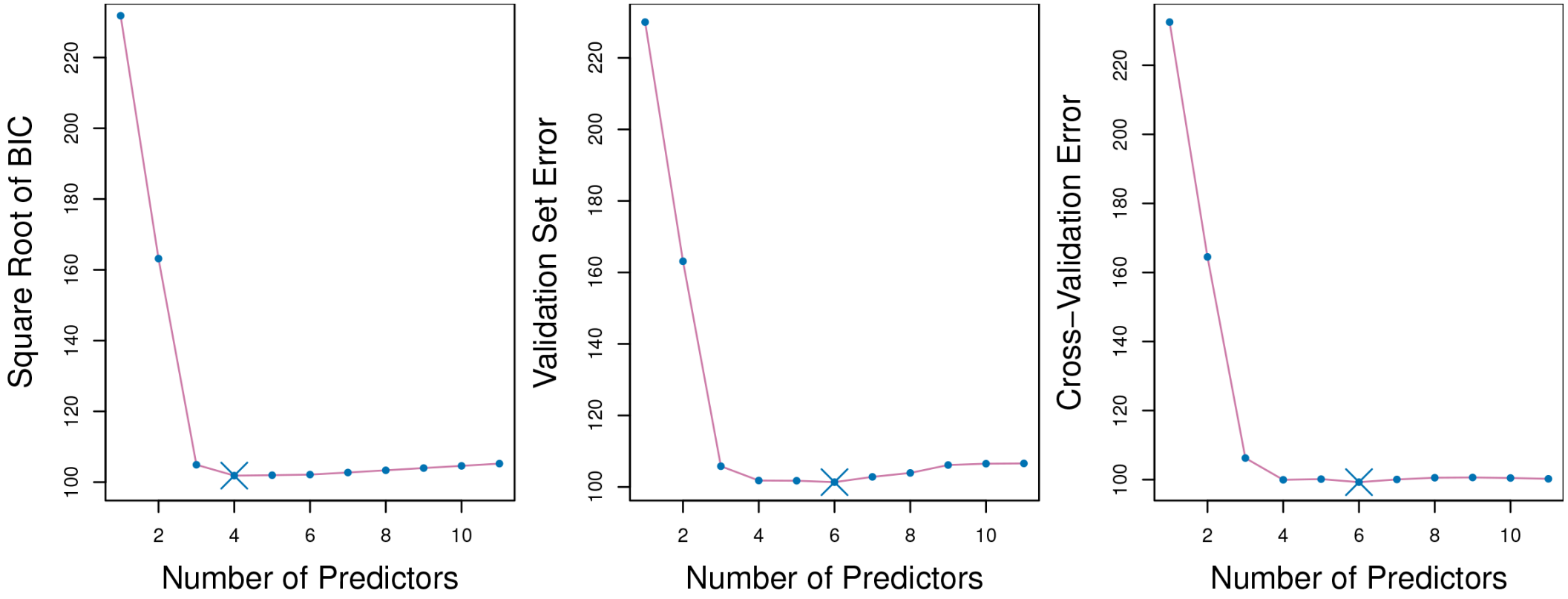
  - It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

  - In the past, performing cross-validation was computationally prohibitive for many problems with large p and/or large n, and so AIC, BIC, C_p and adjusted R^2 were more attractive approaches. However, nowadays with fast computers, the computations required to perform cross-validation are hardly even an issue. Thus, cross-validation is a very attractive approach for selecting from among a number of models under consideration.

Choosing the Optimal Model

- Validation and Cross-Validation

Credit data example



Choosing the Optimal Model

- Validation and Cross-Validation

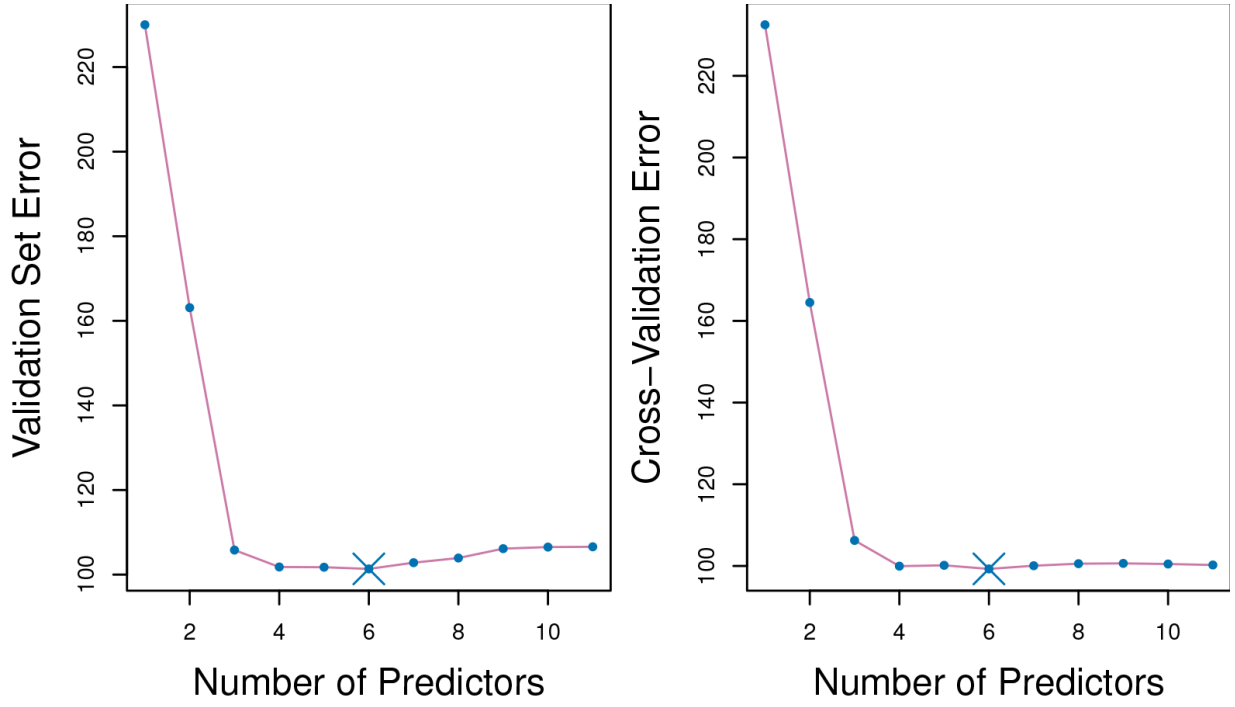
  Credit data example

  - The validation errors were calculated by randomly selecting three-quarters of the observations as the training set, and the remainder as the validation set.
  - The cross-validation errors were computed using k = 10 folds. In this case, the validation and cross-validation methods both result in a six-variable model.
  - However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

Choosing the Optimal Model

- Validation and Cross-Validation

Credit data example



The estimated test error curves are quite flat. From 3 to 11, the estimated test error rates of all models are very similar.

Choosing the Optimal Model

- Validation and Cross-Validation

  Credit data example

  - If we repeat the validation set approach using a different split of the data into a training set and a validation set, or if we repeat the cross-validation using a different set of cross-validation folds, then the precise model with the lowest estimated test error would surely change.

  - In this setting, we can select a model using the *one-standard-error rule*. We first calculate the 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 standard error of the lowest point on the curve.

Shrinkage Methods

- 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.
- Shrinkage methods are more modern techniques in which we don't actually select variables explicitly but rather we 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 relative to the least squares estimates.

Shrinkage Methods

- 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.
- The two best-known techniques for shrinking the regression coefficients towards zero are *ridge regression* and the *lasso*.

Shrinkage Methods

- Ridge Regression

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

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

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

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

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

Shrinkage Methods

- Ridge Regression

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

  - 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 $\beta_1,....,\beta_p$ 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. When $\lambda=0$, the penalty term has no effect, and ridge regression will produce the least squares estimates. However, as $\lambda\to\infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.

Shrinkage Methods

- Ridge Regression

  - Unlike least squares, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates, $\hat{\beta}_{\lambda}^{R}$, for each value of λ .

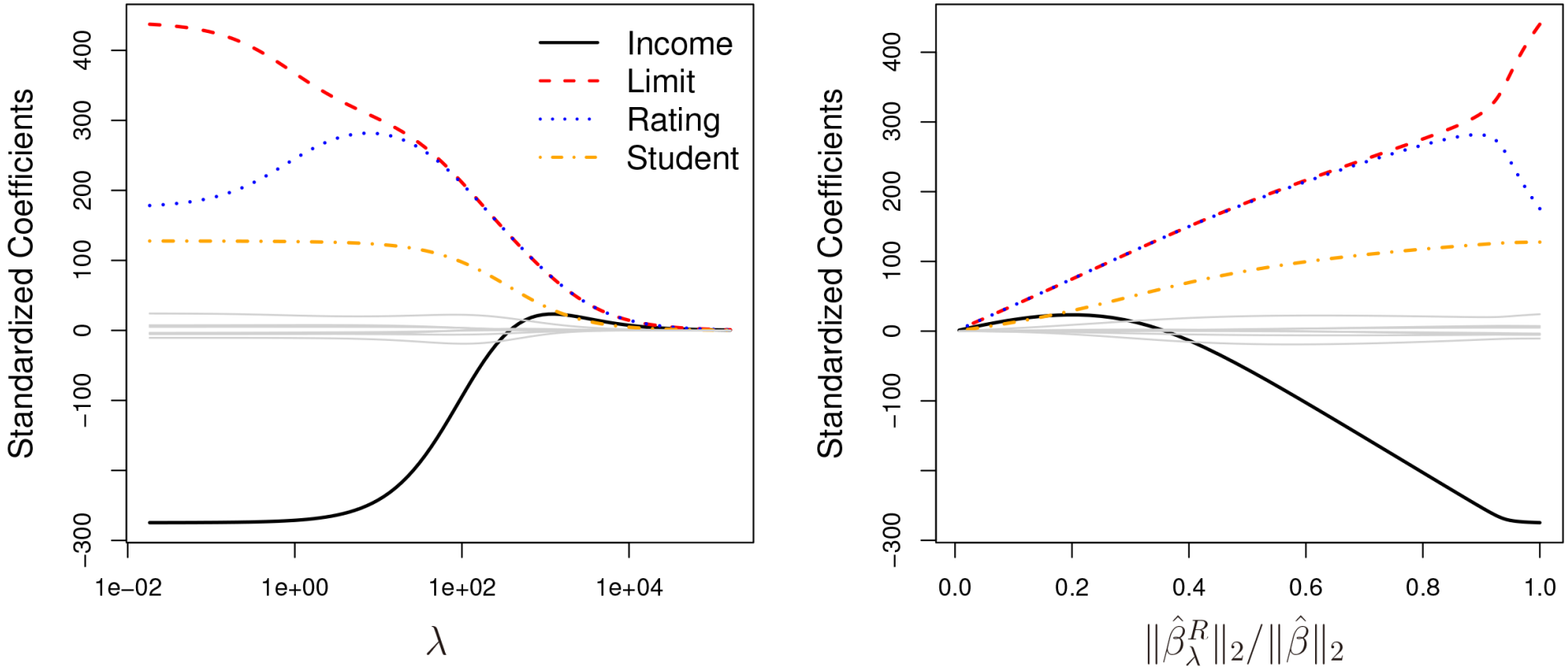
  - Selecting a good value for λ is critical; cross-validation is used for this.

  - The shrinkage penalty is applied to $\beta_1,....,\beta_p$, but not to the intercept β_0 . We want to shrink the estimated association of each variable with the response; however, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when $x_{i1} = x_{i2} = ... = x_{ip} = 0$.

Shrinkage Methods

- Ridge Regression

Credit data example



Shrinkage Methods

- Ridge Regression

  Credit data example

  - In the left-hand panel, each curve corresponds to the standardized ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ .

  - The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but instead of displaying λ on the x-axis, we now display $\left\|\hat{\beta}_\lambda^R\right\|_2 / \left\|\hat{\beta}\right\|_2$ where $\hat{\beta}$ denotes the vector of least squares coefficient estimates.

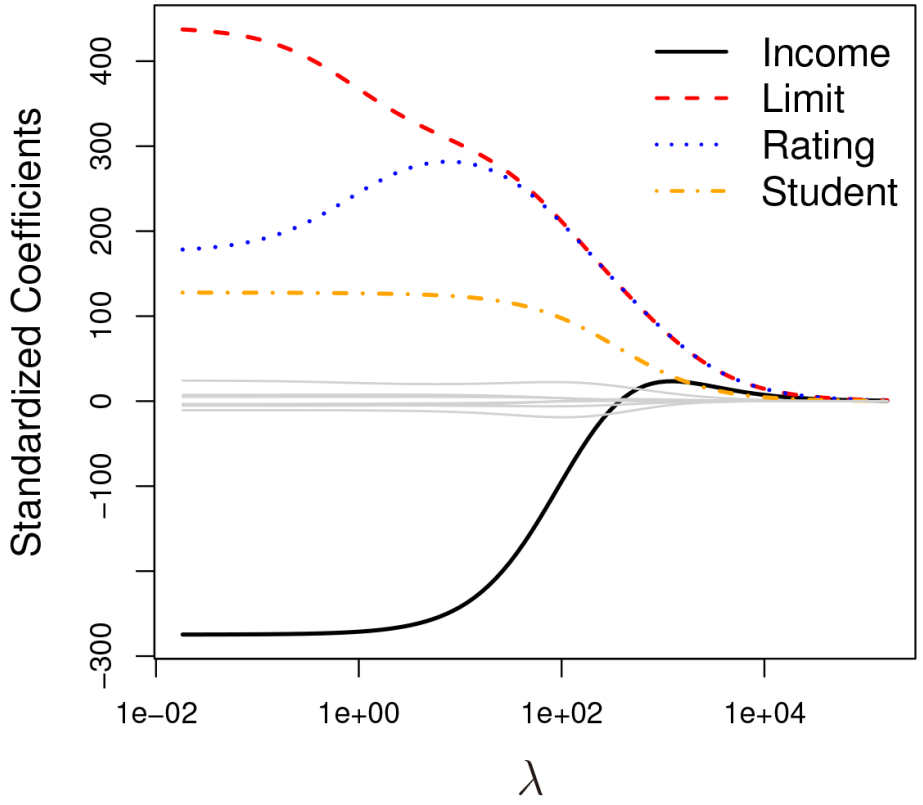
  - The notation $\|\cdot\|_2$ denotes the $l2$ norm of a vector, and is defined as

$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^{\,2}}$$

Shrinkage Methods

- Ridge Regression

Credit data example



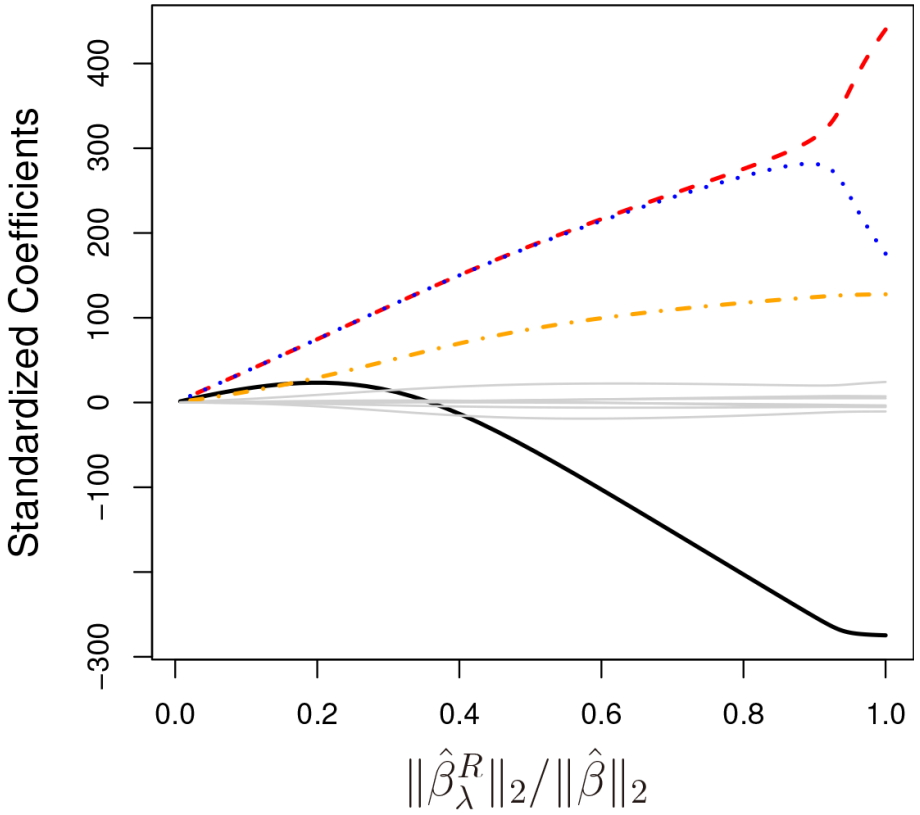
- At the extreme left-hand side of the plot, λ is essentially zero, and so the corresponding ridge coefficient estimates are the same as the usual least squares estimates.

- As λ increases, the ridge coefficient estimates shrink towards zero. When λ is extremely large, then all of the ridge coefficient estimates are basically zero; this corresponds to the null model that contains no predictors.

Shrinkage Methods

- Ridge Regression

Credit data example



- As λ increases, the $l2$ norm of $\hat{\beta}_\lambda^R$ will always decrease, and so will $\left\|\hat{\beta}_\lambda^R\right\|_2 \Big/ \left\|\hat{\beta}\right\|_2$.

  The latter quantity ranges from 1 (when $\lambda = 0$) to 0 (when $\lambda = \infty$).

- We can think of the x-axis as the amount that the ridge regression coefficient estimates have been shrunken towards zero; a small value indicates that they have shrunken very close to zero.

Shrinkage Methods

- Ridge Regression

  Scaling of predictors

  - The standard least squares coefficient estimates are scale equivariant: 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 coefficient 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. In other words, $X_j\hat{\beta}^R_{j,\lambda}$ will depend not only on the value of λ , but also on the scaling of the j-th predictor. In fact, the value of $X_j\hat{\beta}^R_{j,\lambda}$ may even depend on the scaling of the other predictors.

Shrinkage Methods

- Ridge Regression

  Scaling of predictors

  - Therefore, it is best to apply ridge regression after standardizing the predictors, using the formula

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

  so that they are all on the same scale. The denominator is the estimated standard deviation of the jth predictor. Consequently, all of the standardized predictors will have a standard deviation of one. As a result the final fit will not depend on the scale on which the predictors are measured.

Shrinkage Methods

- Ridge Regression

  Why does ridge regression improve over least squares?

  - Rooted in the bias-variance trade-off

    As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

    Experiment: a simulated data set containing p=45 predictors and n=50 observations.

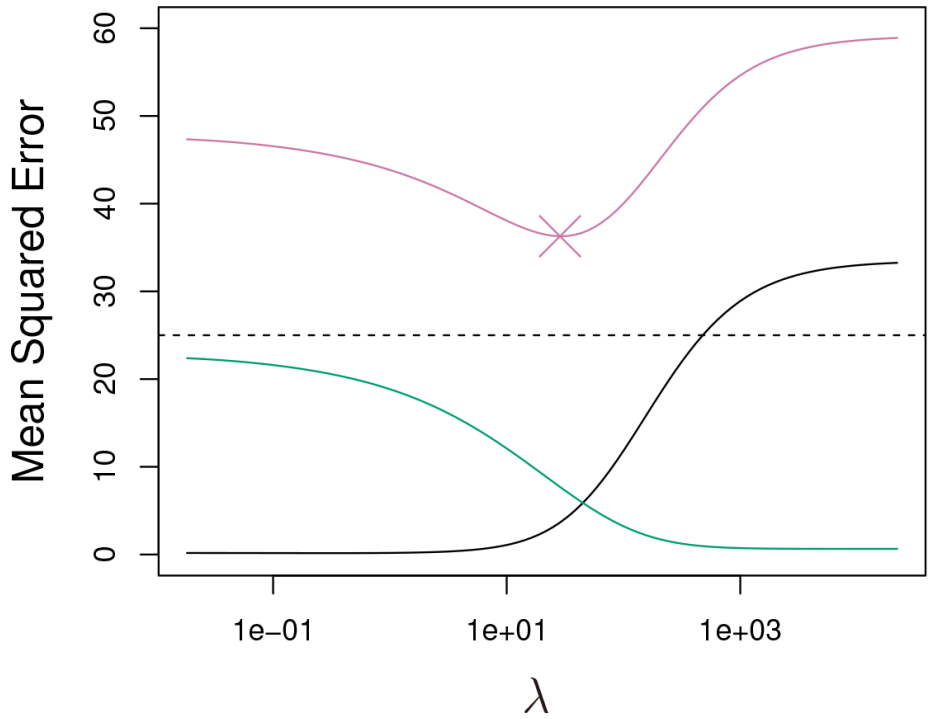
Shrinkage Methods

- Ridge Regression

Why does ridge regression improve over least squares?

➢ Rooted in the bias-variance trade-off

Experiment: a simulated data set containing p=45 predictors and n=50 observations.



--black: squared bias
--green: variance
--purple: test mean squared error

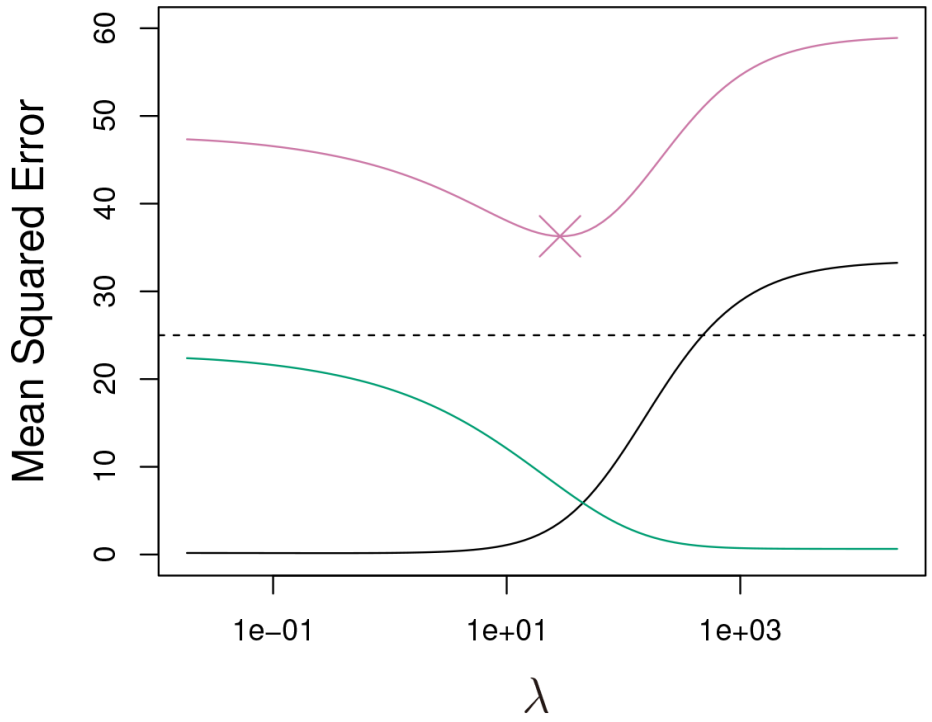
Comment: at the least squares coefficient estimates, which corresponds to ridge regression with $\lambda = 0$, the variance is high but there is no bias.

Shrinkage Methods

- Ridge Regression

Why does ridge regression improve over least squares?

➢ Rooted in the bias-variance trade-off



--black: squared bias
--green: variance
--purple: test mean squared error

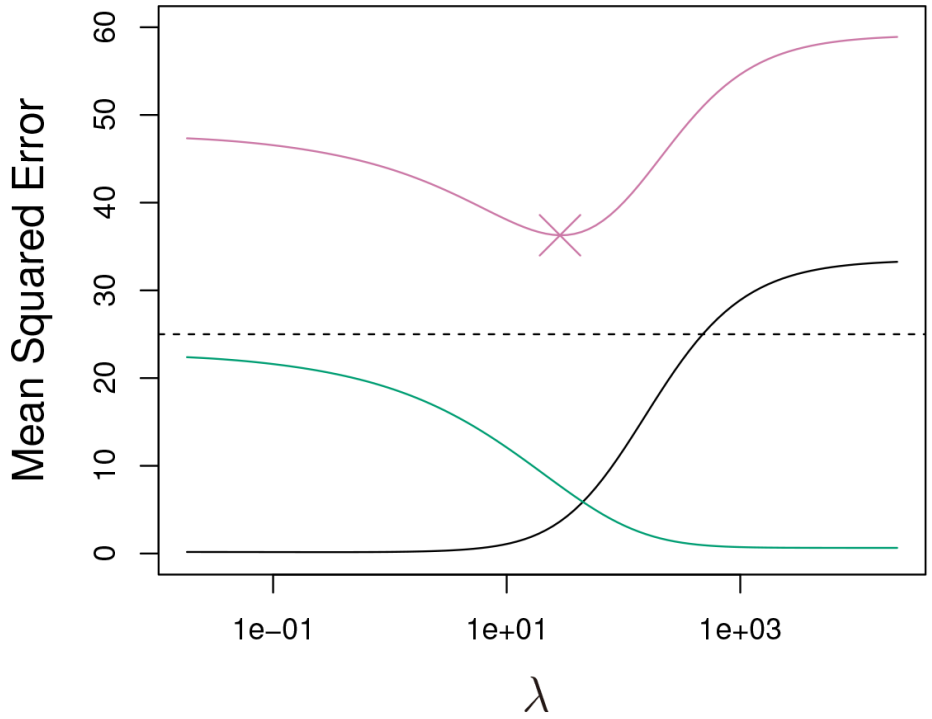
Comment: as λ increases, the shrinkage of the ridge coefficient estimates leads to a substantial reduction in the variance of the predictions, at the expense of a slight increase in bias.

Shrinkage Methods

- Ridge Regression

Why does ridge regression improve over least squares?

➢ Rooted in the bias-variance trade-off



--black: squared bias
--green: variance
--purple: test mean squared error

Comment: the minimum MSE is achieved at approximately $\lambda=30$.

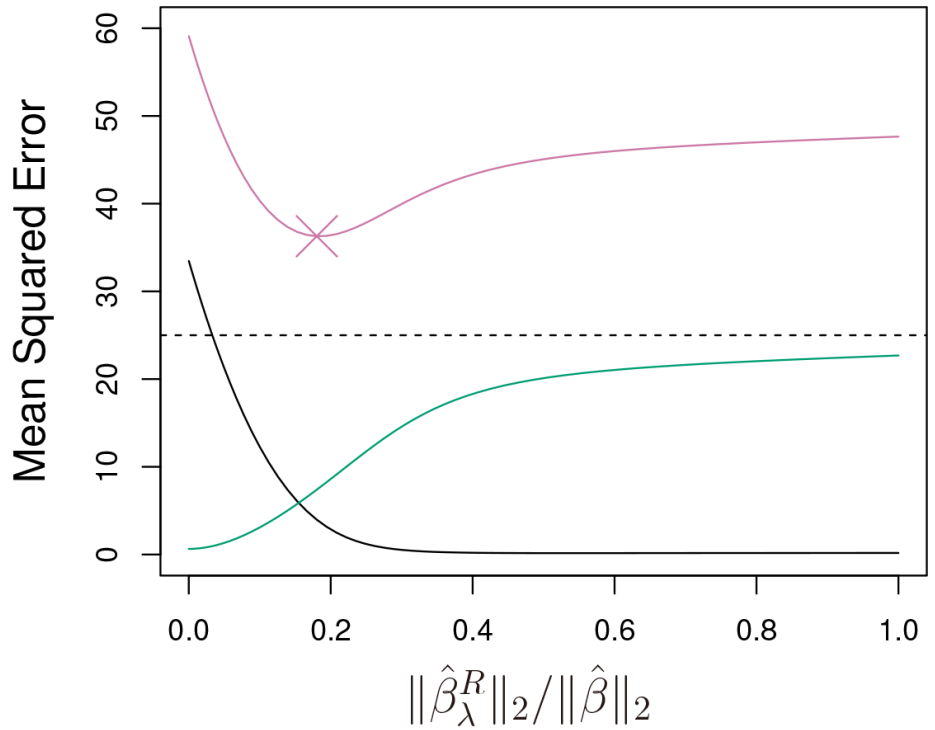
Because of its high variance, the MSE associated with the least squares fit ($\lambda=0$), is almost as high as that of the null model for which all coefficient estimates are zero ($\lambda=\infty$).

Shrinkage Methods

- Ridge Regression

Why does ridge regression improve over least squares?

➢ Rooted in the bias-variance trade-off



--black: squared bias
--green: variance
--purple: test mean squared error

Comment: now as we move from left to right, the fits become more flexible, and so the bias decreases and the variance increases.

Shrinkage Methods

- Ridge Regression

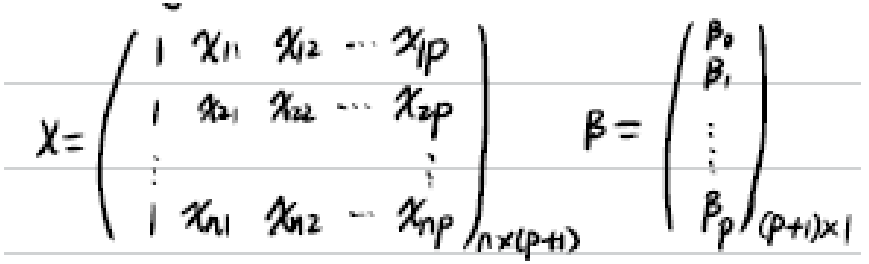
  Why does ridge regression improve over least squares?

  - Rooted in the bias-variance trade-off

    Comment: When the number of variables p is almost as large as the number of observations n, as in the previous example, the least squares estimates will be extremely variable. And if p>n, then the least squares estimates do not even 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.

    Hence, ridge regression works best in situations where the least squares estimates have high variance.

Shrinkage Methods



- Ridge Regression

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

$$E^R = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p}\beta_j{}^2 = (Y - X\beta)^T (Y - X\beta) + \lambda \left\| \beta \right\|_2^2$$

- E^R is convex, and hence has a unique solution.
- Taking derivatives, we have:

$$\frac{\partial E^R}{\partial \beta} = -2X^T(Y - X\beta) + 2\lambda\beta \quad \Rightarrow \quad \hat{\beta}_\lambda^R = (X^T X + \lambda I_{p+1})^{-1} X^T Y$$

Shrinkage Methods

- Ridge Regression

$$E^R = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=0}^{p}\beta_j^2 = (Y - X\beta)^T(Y - X\beta) + \lambda \|\beta\|_2^2$$

$$\hat{\beta}_\lambda^R = (X^T X + \lambda I_{p+1})^{-1} X^T Y$$

Comment: Inclusion of λ makes problem non-singular even if $X^T X$ is not invertible. This was the original motivation for ridge regression (Hoerl and Kennard, 1970)

Shrinkage Methods

- Ridge Regression

  Scaling of predictors

  - The standard least squares coefficient estimates are scale equivariant: 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.

Shrinkage Methods

$$\hat{\beta}_\lambda^R = (X^T X + \lambda I_{p+1})^{-1} X^T Y$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n\times(p+1)} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1}$$

In linear Regression:

我们的优化问题是: $\min_\beta RSS = \min_\beta (Y - X\beta)^T (Y - X\beta) = \min_\beta [Y^T Y + \beta^T (X^T X)\beta - 2 Y^T X \beta]$

$$\frac{\partial RSS}{\partial \beta} = 2(X^T X)\beta - 2(X^T Y) = 0 \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y \Rightarrow \hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y$$

when $X^* = CX$, we have $\hat{\beta}^* = (X^{*T} X^*)^{-1} X^{*T} y$

$$= (CX^T \cdot CX)^{-1} C X^T Y$$

$$= \frac{1}{C^2} \cdot C (X^T X)^{-1} X^T Y$$

$$= \frac{1}{C} (X^T X)^{-1} X^T Y$$

$$\Rightarrow \hat{Y}^* = X^* \hat{\beta}^* = CX \cdot \frac{1}{C} (X^T X)^{-1} X^T Y = X(X^T X)^{-1} X^T Y = \hat{Y}$$

This means, regardless of how the j-th predictor is scaled, $X_j \hat{\beta}_j$ will remain the same

Shrinkage Methods

- Ridge Regression

  Scaling of predictors

  - In contrast, the ridge regression coefficient 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. In other words, $X_j\hat{\beta}^R_{j,\lambda}$ will depend not only on the value of λ , but also on the scaling of the j-th predictor. In fact, the value of $X_j\hat{\beta}^R_{j,\lambda}$ may even depend on the scaling of the other predictors.

Shrinkage Methods

In ridge regression, the optimization problem is:

$$f(x) = \underset{\beta}{argmin}(Y - X\beta)^T(Y - X\beta) + \lambda \|\beta\|_2^2$$

$$\frac{\partial f(x))}{\partial \beta} = 2(X^TX)\beta - 2X^TY + 2\lambda\beta = 0 \Rightarrow \hat{\beta}_\lambda^R = (X^TX + \lambda I_{p+1})^{-1}X^TY$$
$$\Rightarrow \hat{Y}^R = X(X^TX + \lambda I_{p+1})^{-1}X^TY$$

$When\ X^* = CX, we\ have:$

$$\begin{aligned}\hat{\beta}_\lambda^{*R} &= (X^{*T}X^* + \lambda I_{p+1})^{-1}X^{*T}Y \\ &= (C^2X^TX + \lambda I_{p+1})^{-1}CX^TY \\ &= (CX^TX + \frac{1}{C}\lambda I_{p+1})^{-1}X^TY \\ \hat{Y}^{*R} &= X^*\hat{\beta}_\lambda^{*R} = C^2X(C^2X^TX + \lambda I_{p+1})^{-1}X^TY \neq \hat{Y}^R\end{aligned}$$

So, in ridge regression, $X_j\hat{\beta}_{j,\lambda}^R$ will depend not only on the value of λ, but also on the scaling of the j-th predictor.

Shrinkage Methods

- The Lasso

  - Ridge regression does have one obvious disadvantage: unlike subset selection (best, forward stepwise, or backward stepwise), which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.

    The penalty $\lambda\sum_{j=1}^{p}\beta_j^2$ will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda=\infty$).

    This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of variables p is quite large.

Shrinkage Methods

- The Lasso
  - The Lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$,minimize the quantity

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

  - In statistical parlance, the lasso uses an $l1$ penalty instead of an $l2$ penalty. The $l1$ norm of a coefficient vector β is given by $\|\beta\|_1=\sum\left|\beta_j\right|$.

  - Unlike ridge regression, the lasso solution does not have a closed form.

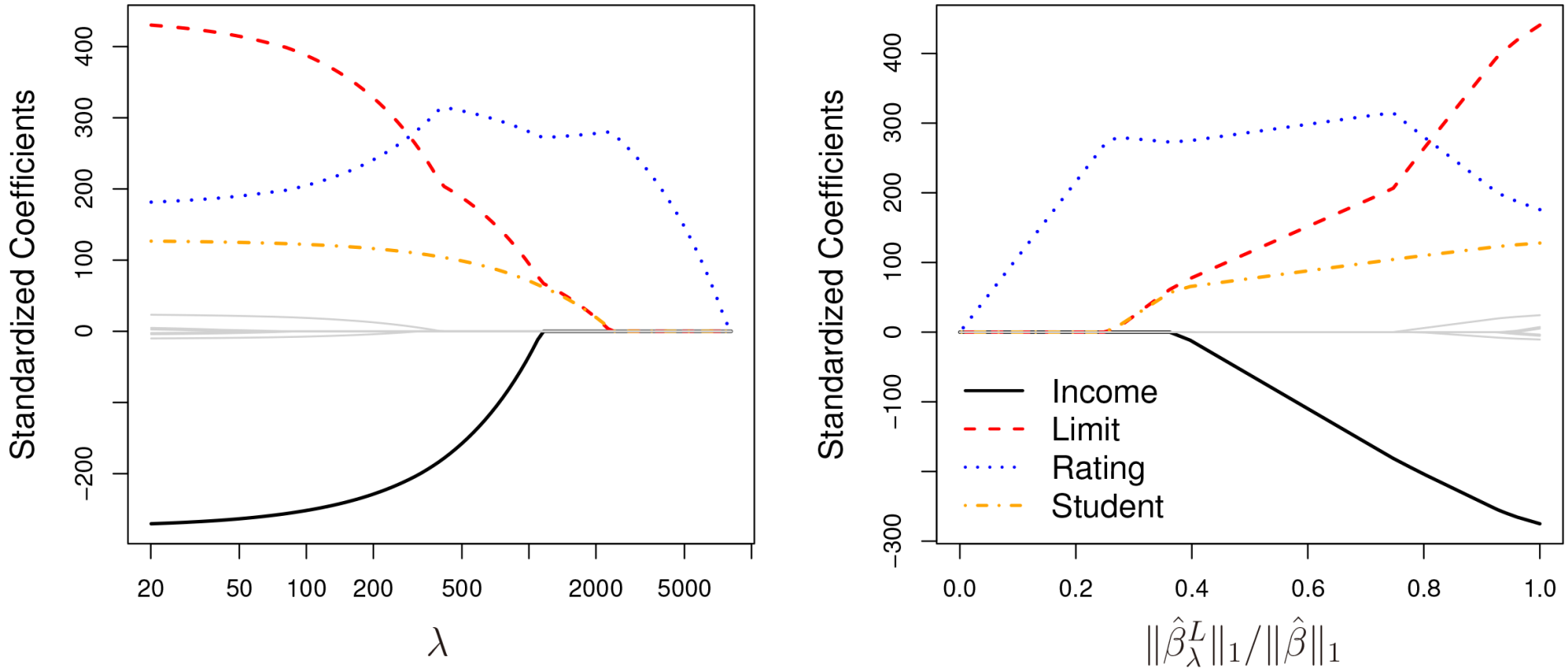
Shrinkage Methods

- The Lasso $$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|=RSS+\lambda\sum_{j=1}^{p}\left|\beta_j\right|$$
  - As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
  - However, in the case of the lasso, the $l1$ 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 *variable selection*.
  - We say that the lasso yields *sparse* models -- that is, models that involve only a subset of the variables.
  - As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice.

Shrinkage Methods

- The Lasso

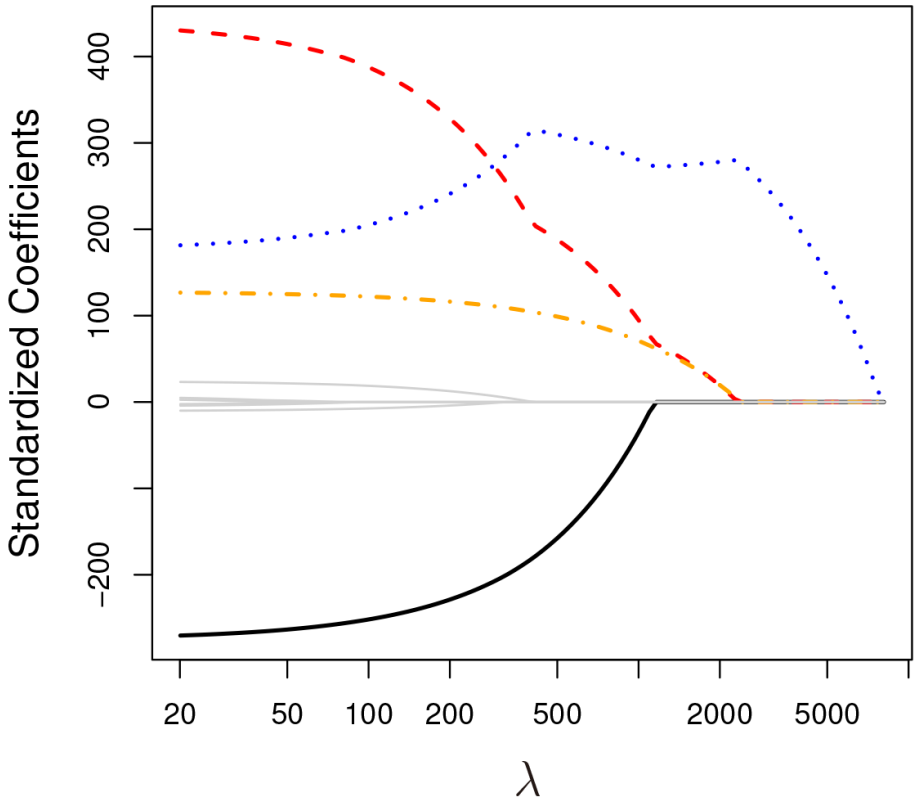
Credit data example

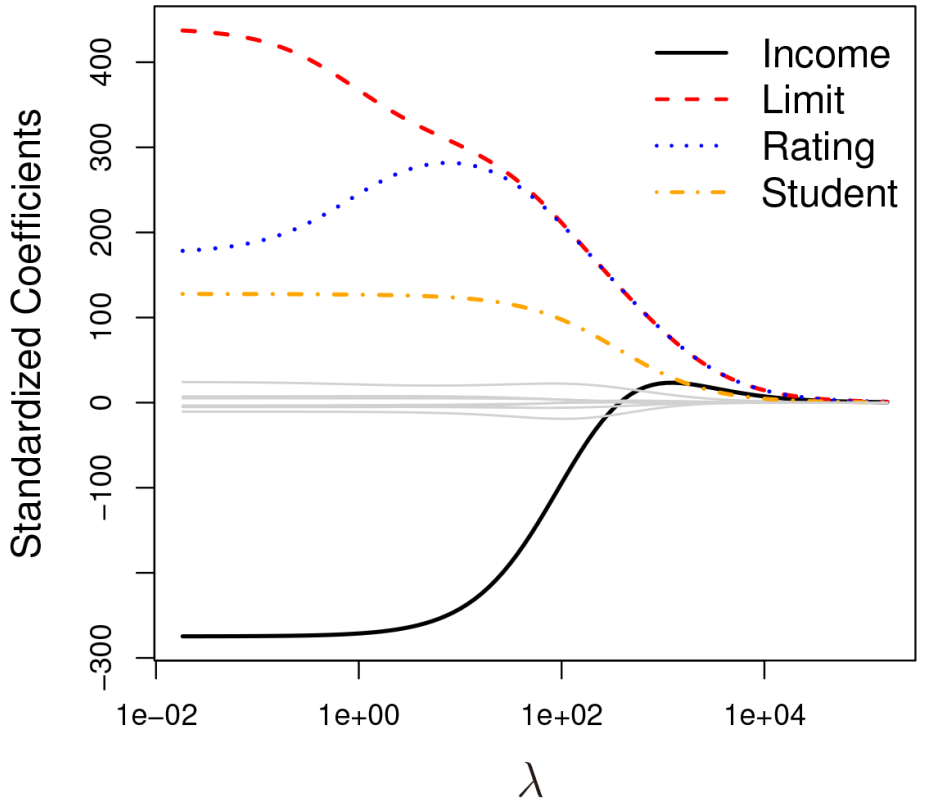


Shrinkage Methods

- The Lasso

Credit data example

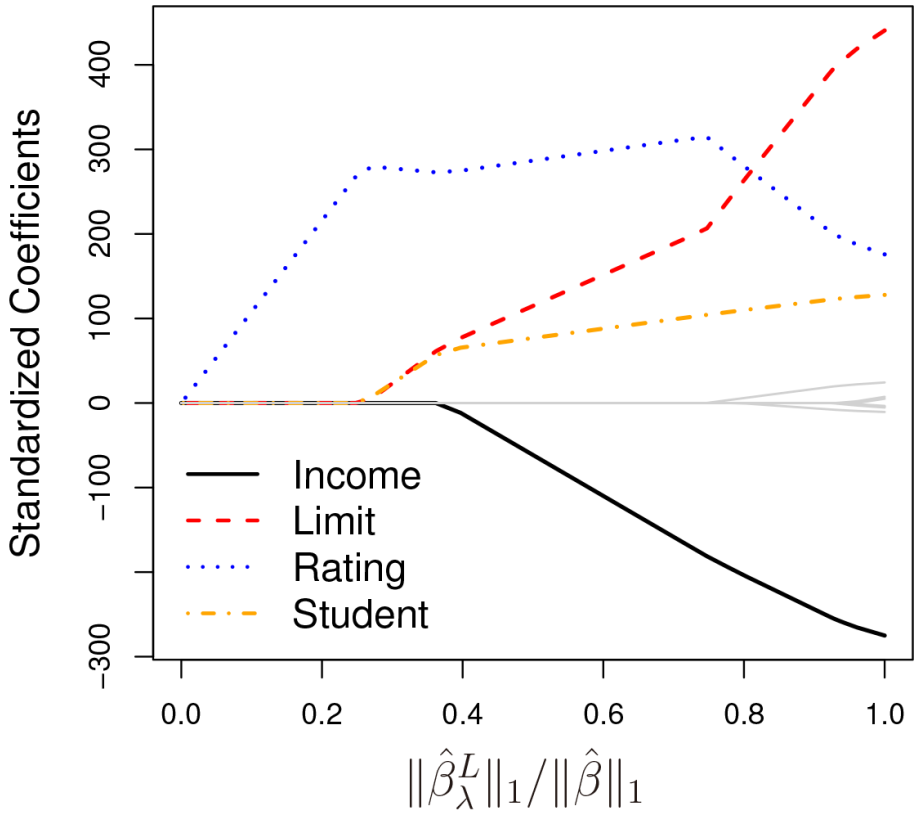




Shrinkage Methods

- The Lasso

Credit data example



Comment: moving from left to right, we observe that at first the lasso results in a model that contains only the rating predictor. Then student and limit enter the model almost simultaneously, shortly followed by income. Eventually, the remaining variables enter the model.

Shrinkage Methods

- The Lasso

Comment: Depending on the value of λ , the lasso can produce a model involving any number of variables. In contrast, ridge regression will always include all of the variables in the model, although the magnitude of the coefficient estimates will depend on λ .

Shrinkage Methods

- Another formulation for ridge regression and the lasso

The lasso

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

Ridge regression

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

Shrinkage Methods

- Another formulation for ridge regression and the lasso

The lasso

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

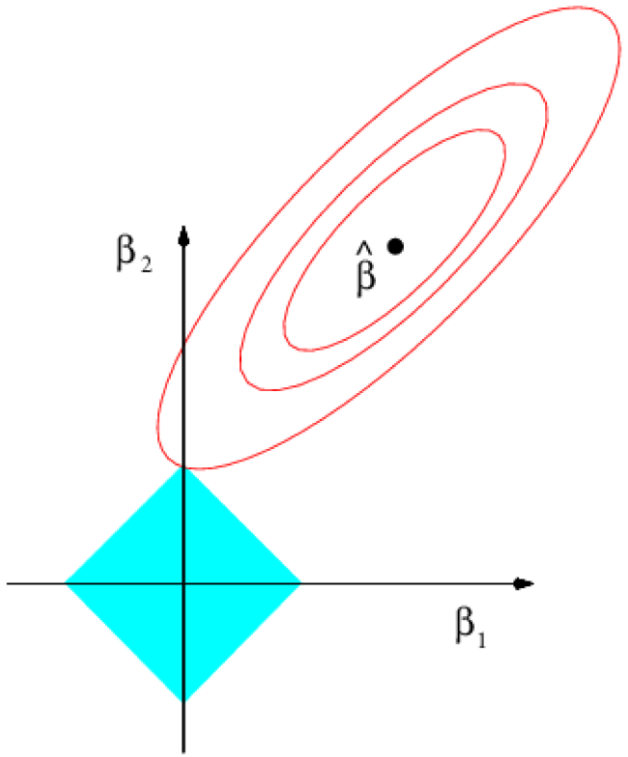
When p=2, the above indicates that the lasso coefficient estimates have the smallest RSS out of all points that lie within the diamond defined by $|\beta_1| + |\beta_2| \leq s$.

Shrinkage Methods

- Another formulation for ridge regression and the lasso

The lasso

When p=2, the above indicates that the lasso coefficient estimates have the smallest RSS out of all points that lie within the diamond defined by $|\beta_1| + |\beta_2| \leq s$.



--solid blue area: the constrains regions
--red ellipses: the contours of the RSS

Comment: If s is large enough, the least squares solution falls within the constraint region, then it will simply yield the least squares solution.

Shrinkage Methods

- Another formulation for ridge regression and the lasso

Ridge regression

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j{}^2 \leq s$$

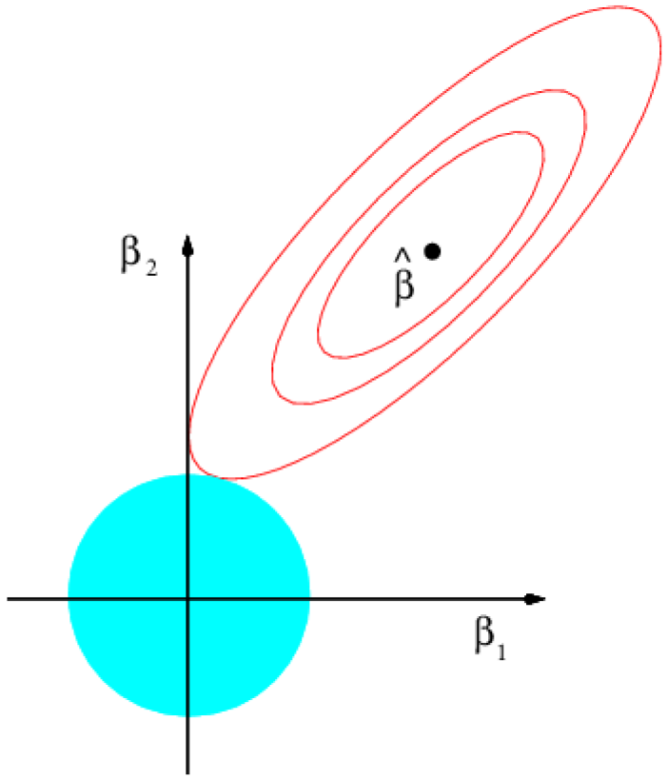
When p=2, the above indicates that the ridge coefficient estimates have the smallest RSS out of all points that lie within the circle defined by $\beta_1{}^2 + \beta_2{}^2 \leq s$.

Shrinkage Methods

- Another formulation for ridge regression and the lasso

Ridge regression

When p=2, the above indicates that the ridge coefficient estimates have the smallest RSS out of all points that lie within the circle defined by $\beta_1^2 + \beta_2^2 \leq s$.



--solid blue area: the constrains regions
--red ellipses: the contours of the RSS

Shrinkage Methods

- Similar formulation for best subset selection

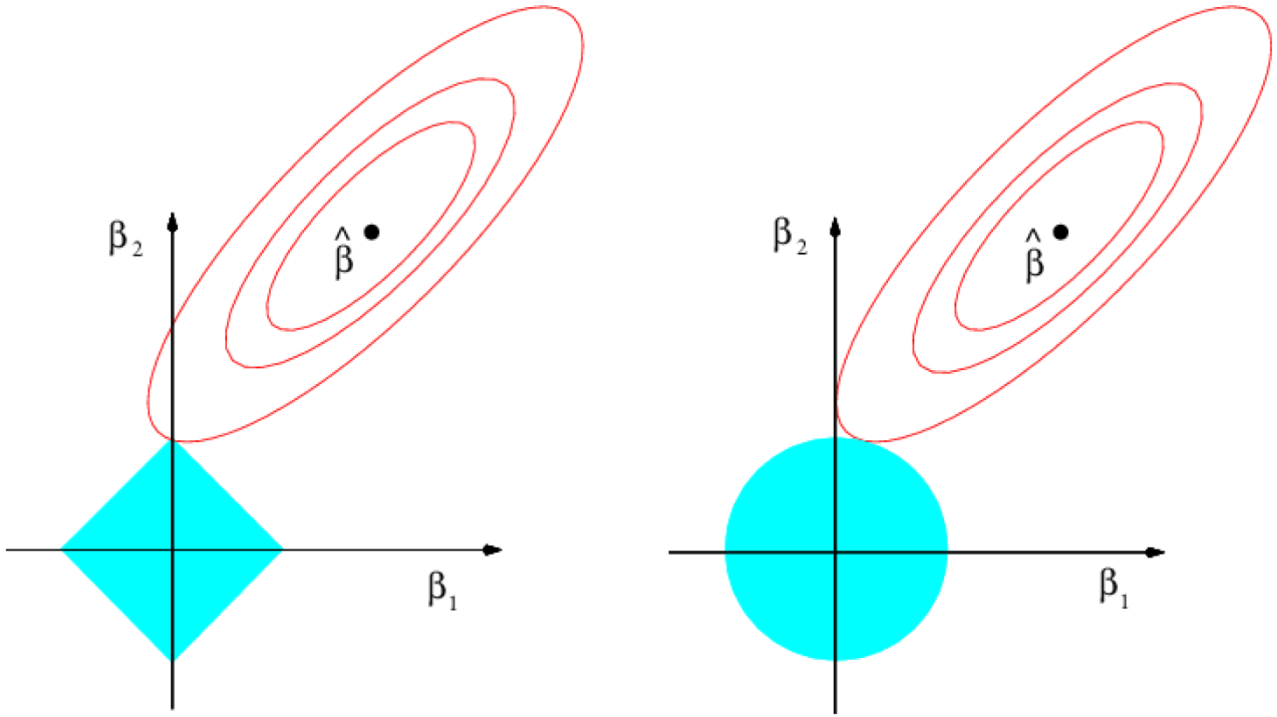
$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s$$

The above formulation amounts to finding 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. This is equivalent to *best subset selection*.

Shrinkage Methods

- The 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?



Since ridge regression has a circular constraint with no sharp points, the intersection will not generally occur on an axis.

The lasso constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis.

Shrinkage Methods

- The Variable Selection Property of the Lasso

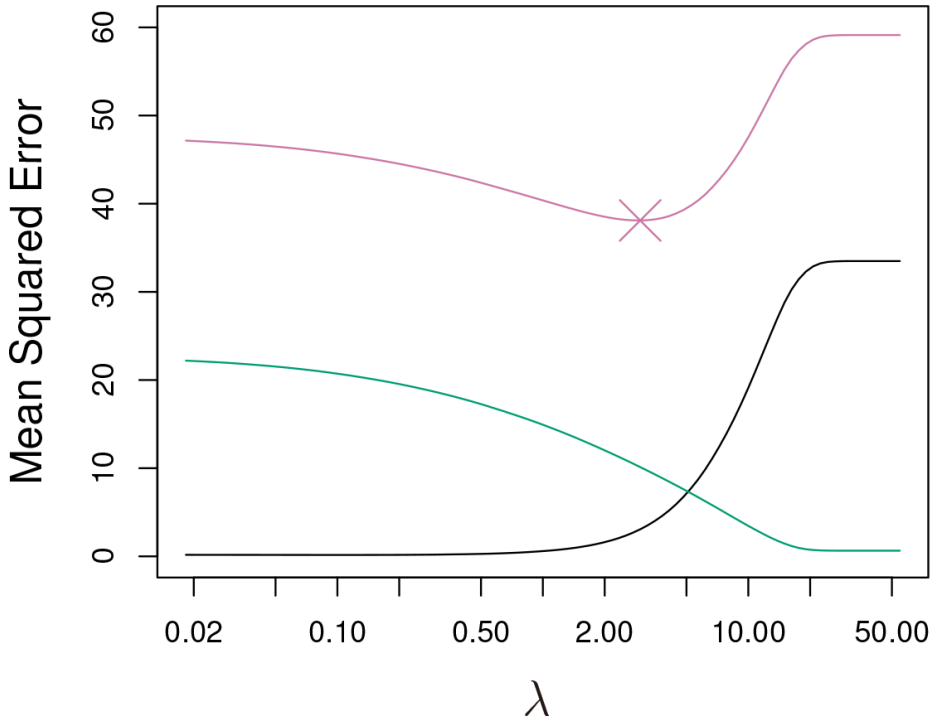
Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?

Comment: When p=3, then the constraint region for ridge regression becomes a sphere, and the constraint region for the lasso becomes a polyhedron. When p>3, the constraint for ridge regression becomes a hypersphere, and the constraint for the lasso becomes a polytope.

The lasso leads to feature selection when p>2 due to the sharp corners of the polyhedron or polytope.

Shrinkage Methods

- Comparing the lasso and ridge regression

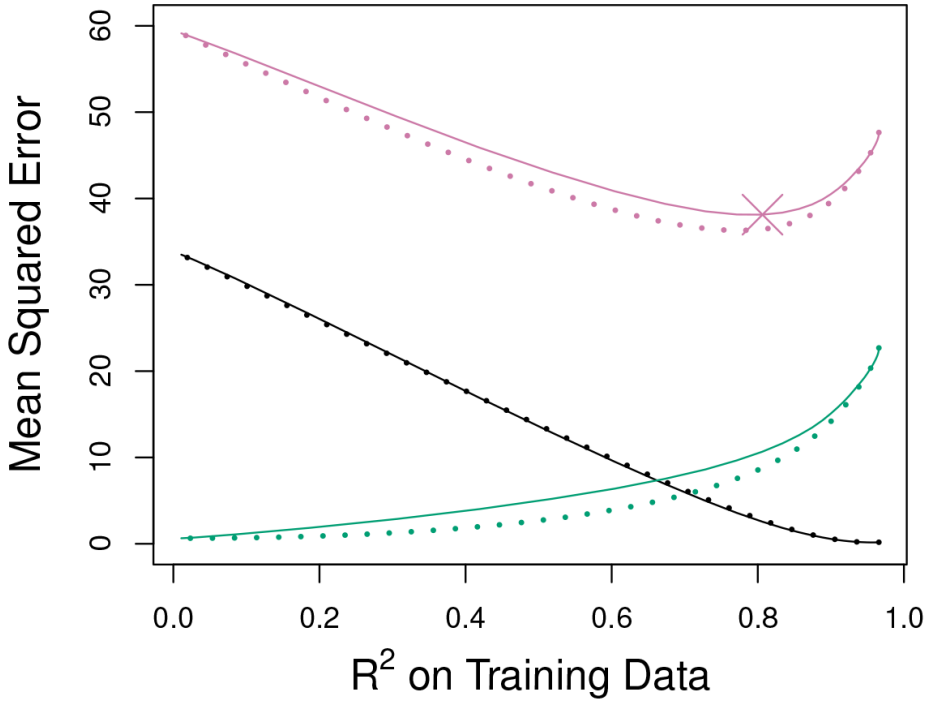


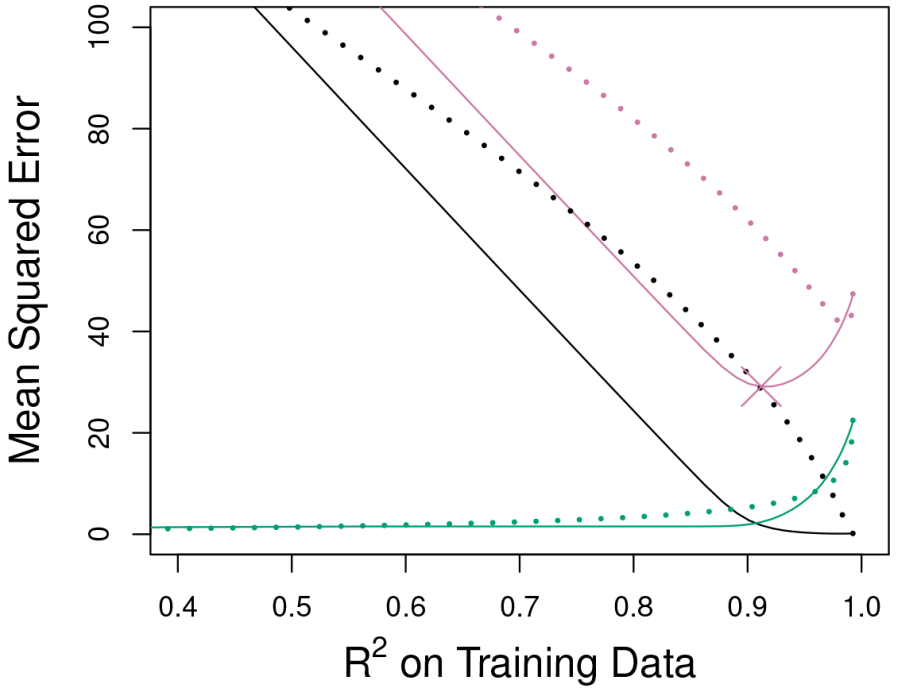
--black: squared bias
--green: variance
--purple: test mean squared error

Comment: the lasso leads to qualitatively similar behavior to ridge regression, in that as λ increases, the variance decreases and the bias increases.

Shrinkage Methods

- Comparing the lasso and ridge regression

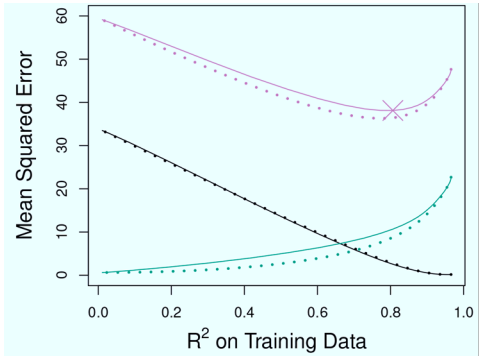


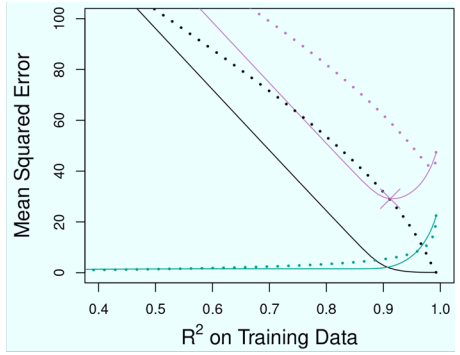


--black: squared bias
--green: variance
--purple: test mean squared error

--dotted: ridge regression
--solid: the lasso

Shrinkage Methods





- Comparing the lasso and ridge regression

  - These two examples illustrate that neither ridge regression nor the lasso will universally dominate the other.

  - In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors. Ridge regression will perform better when the response is a function of many predictors.

  - However, the number of predictors that is related to the response is never known a priori for real data sets.

  - A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.