Topic 5. Linear Model Selection and Regularization

Review: Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

β_0 ---- intercept (i.e., the average value of Y if all inputs are zero)

β_j ---- slope for the *j*th predictor (the average increase in Y when X_j is increased by 1 and all other predictors are held constant)

Review: Estimating Coefficients

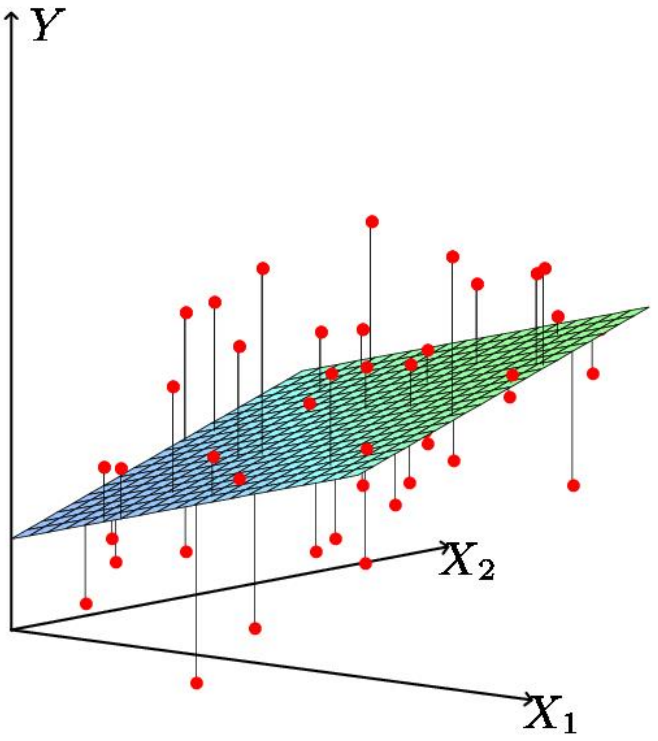
- Given a training data set

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$$
$$\boldsymbol{x}_i = (x_1, x_2, \ldots, x_p)$$

- Use Least Squares (LS) method to find coefficient estimates

$$\text{minimize } RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p\right)^2$$



Improving the Least Squares Estimates

- We want to improve the linear regression model, by replacing the ordinary least square (OLS) fitting with some alternative fitting procedure.
- Two reasons for improving the OLS model
  - Prediction Accuracy
  - Model Interpretability

1. Prediction Accuracy

- The least squares estimates have relatively low bias and low variability especially when
  - the relationship between Y and X is linear and
  - # observations is way bigger than # predictors ($n \gg p$)

- When $n \approx p$, OLS fit may have high variance, and result in overfitting and poor estimates on unseen observations.

- When $n < p$, OLS fit does not work, the variance of these estimates is infinite so this method cannot be used at all.

2. Model Interpretability

- When we have a large number of X variables in the model there will generally be many that have little or no effect on Y.
- Leaving these variables (terms) in the model makes it harder to see the “big picture”, i.e., the effect of the “important variables”.
- The model would be easier to interpret by removing the unimportant variables (i.e., setting the coefficients of those variables to zero).

Solution

- **Subset Selection**
  - Identifying a subset of the p predictors that we believe to be related to the response, and then fitting the model using this subset.
  - E.g., best subset selection and stepwise selection
- **Shrinkage**
  - Involves shrinking the estimates of coefficients toward zero
  - The shrinkage reduces the variance.
  - Some of the coefficients may shrink to exactly zero, and hence shrinkage methods can also perform variable selection.
  - E.g. Ridge regression and the Lasso
- Dimension Reduction
  - Involves projecting the p predictors into an M-dimensional space, where $M<p$, and then fitting regression model of Y on the projections.
  - E.g. Principle Components Regression

Subset Selection

Subset Selection

- From the p predictor variables, choose a subset of them such that the linear model with this subset as predictors has best **prediction accuracy**.

- Popular methods
  - Best subset selection
  - Stepwise selection

1. Best Subset Selection

- In this approach, we run a linear regression for each possible combination of the predictors.
- The set of possible models include: models that contain exactly one predictor, models that contain exactly two predictors, …, and the model with all the p predictors.
- There are $\binom{p}{k}, k = 1, \ldots, p$ models that contain exactly k predictors. For example,
  - There are $\binom{p}{1} = p$ models that contain exactly one predictor.
  - There are $\binom{p}{2} = p(p-1)/2$ models that contain exactly two predictors.

Procedure

- Step 1: find the best model among the models with the same number of predictors

  (Model_0) the model with no predictors
  (Model_1) the best model among the models with 1 predictor
  (Model_2) the best model among the models with 2 predictors
  ……
  (Model$p-1$) the best model among the models with $p-1$ preditors
  (Model$p$) the model with all the p predictors

- Step 2: find the overall best model among the models obtained in Step 1

The Algorithm of 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.

Model Selection Criteria

- **Algorithm-2(b):** use RSS or R^2 as criteria
- Be cautious: when the number of predictors in the model increases, RSS decreases monotonically, and R^2 increases monotonically.
- Therefore, if we use them to select *the single best model*, we will always end up with a model involving all the predictors.
- The problem is that a low RSS or a high R^2 indicates a model with low *training* error, not test error.

  (in other words, more predictors always fit data better in training.)

- **Algorithm-3:** will be discussed later.

Example: Credit Data

- Example: Credit data set

Response

balance: average credit card debt

Predictors

age

cards: number of credit cards

education: years of education

income: in thousands of dollars

limit: credit limit

rating: credit rating

gender: male/female (1 dummy variable)

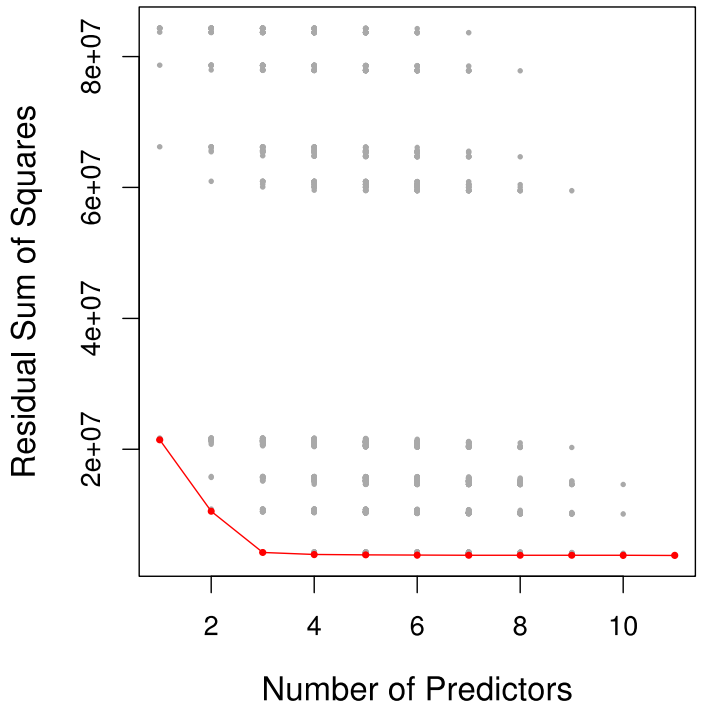
student: student status (1 dummy variable)

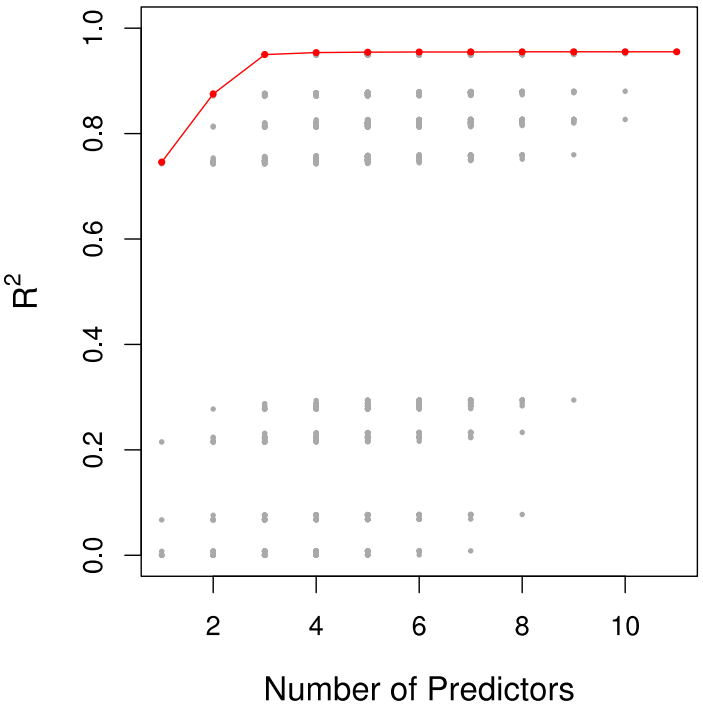
status: marital status (1 dummy variable)

Ethnicity: Caucasian/African American/Asian (2 dummy variables)

Example: RSS and R^2

- The red line tracks the best model for a given number of predictors, according to RSS and R^2.
- As expected, the RSS/R^2 will always decline/increase as the number of variables increases.
- From the three-predictor model on, adding predictors lead to little improvement.





Drawback of Best Subset Selection

- Best subset selection is a conceptually simple approach. But it is computationally intensive especially when we have a large number of predictors (large p).
- The number of possible models =2^p. If $p = 10$, that means we will consider 1024 possible models. If $p = 20$, there are over one million possibilities!
- This method becomes computationally infeasible for values of p greater than around 40, even with extremely fast modern computers.

2. Stepwise Selection

- Best subset selection searches the enormous space of possible models, whereas stepwise selection explores a far more restricted set of models, which are more computationally efficient.
- Two methods:
  - Forward stepwise selection: Begins with the model containing no predictor, and then adds one predictor at a time that improves the model the most until no further improvement is possible
  - Backward stepwise selection: Begins with the model containing all predictors, and then deleting one predictor at a time that improves the model the most until no further improvement is possible

Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest 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.

A Simple Example of Forward Stepwise Selection

- Assume predictors X_1, X_2, X_3, $p = 3$
- Algorithm-1

  The null model $\rightarrow$ $M_0 = Y \sim intercept$
- Algorithm-2
  1. Consider adding one predictor (X_1 or X_2 or X_3)

     3 models with one predictor: $Y \sim X_1$, $Y \sim X_2$, $Y \sim X_3$
  2. Choose the best among them $\rightarrow$ $M_1 = Y \sim X_2$
  3. Then consider adding one more predictor (X_1 or X_3)

     2 models with two predictors: $Y \sim X_2 + X_1$, $Y \sim X_2 + X_3$
  4. Choose the best among them $\rightarrow$ $M_2 = Y \sim X_2 + X_1$
  5. Consider adding one more predictor (X_3) $\rightarrow$ $M_3 = Y \sim X_2 + X_1 + X_3$
- Algorithm-3

  Choose the best model among M_0, M_1, M_2, M_3

Computational Advantage

- Best subset selection involves 2^p models. Forward stepwise selection involves $1 + p(p+1)/2$ models.
- When $p = 20$, best subset selection requires fitting 1,048,576 models.
- Forward stepwise selection requires fitting only 211 models.

Example: Credit Data

- However, since forward stepwise selection only searches part of the space of possible models, there is no guarantee that it will find the best possible model as best subset selection does.
- Results on Credit data set: the first four selected models from best subset selection and forward stepwise selection

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

Backward Stepwise Selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest 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.

A Simple Example of Backward Stepwise Selection

- Assume predictors X_1, X_2, X_3, $p = 3$
- Algorithm-1

  The full model $\rightarrow$ $M_3 = Y \sim X_1 + X_2 + X_3$
- Algorithm-2

  1. Consider removing one predictor

     3 models with two predictors: $Y \sim X_2 + X_3$, $Y \sim X_1 + X_3$, $Y \sim X_1 + X_2$
  2. Choose the best among them $\rightarrow$ $M_2 = Y \sim X_2 + X_3$
  3. Then consider removing one more predictor (X_2 or X_3)

     2 models with one predictor: $Y \sim X_2$, $Y \sim X_3$
  4. Choose the best among them $\rightarrow$ $M_1 = Y \sim X_2$
  5. Consider removing one more predictor (X_2) $\rightarrow$ $M_0 = Y \sim intercept$
- Algorithm-3

  Choose the best model among M_0, M_1, M_2, M_3

Properties about Backward Stepwise Selection

- Like forward stepwise selection, backward selection searches through only $1 + p(p + 1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection.
- Also, there is no guarantee to yield the best possible model.
- Backward vs. Forward: backward selection requires $n > p$, so that the full model can be fit. Forward selection can be used even when $n < p$. So forward selection is the only viable subset method when the number of predictors is very large.

Selecting the Single Best Model

- Subset selection results in a set of "good" models, each of which contains a subset of the p predictors. How do we determine which model is the "best" among them in terms of **prediction accuracy**?

- One simple approach is to take the model with the smallest RSS or the largest R^2.

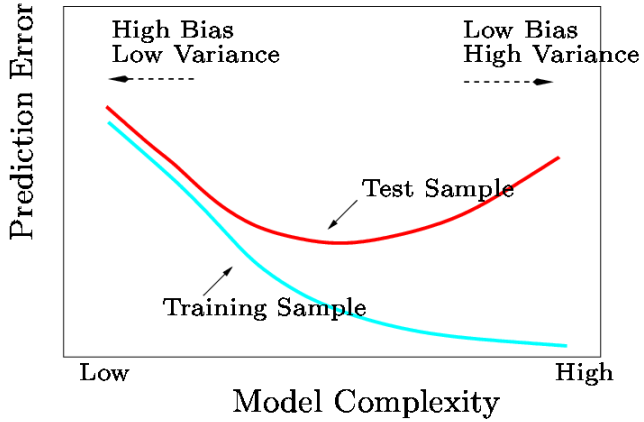
- As we discussed, unfortunately, the model that includes all the variables will always have the largest R^2 (and smallest RSS). This means that RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

Approaches

- Keep in mind: we want to find the model that has the best prediction accuracy (or lowest test error).

- Two common approaches

1. Indirectly estimate the test error by making an adjustment to the training error to account for the bias due to overfitting



2. Directly estimate the test error, using a validation set approach or cross validation

Approach 1

- These methods add penalty to RSS for the number of predictors (i.e. complexity) in the model (linear regression).
- AIC (Akaike information criterion)

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

- BIC (Bayesian information criterion)

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

- C_p (equivalent to AIC for linear regression)

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

- Adjusted R^2

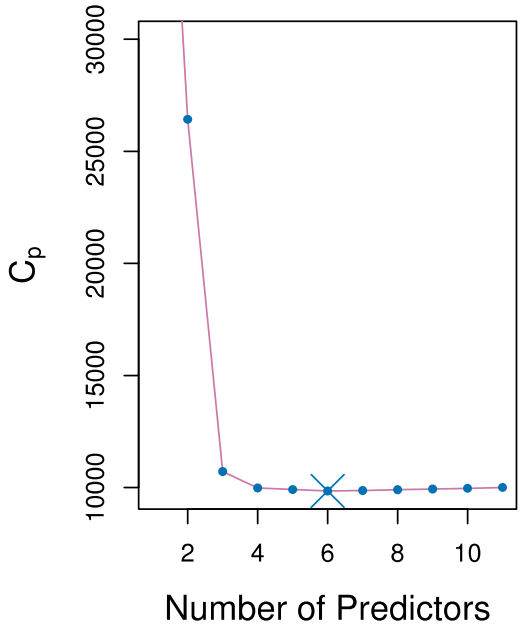
$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

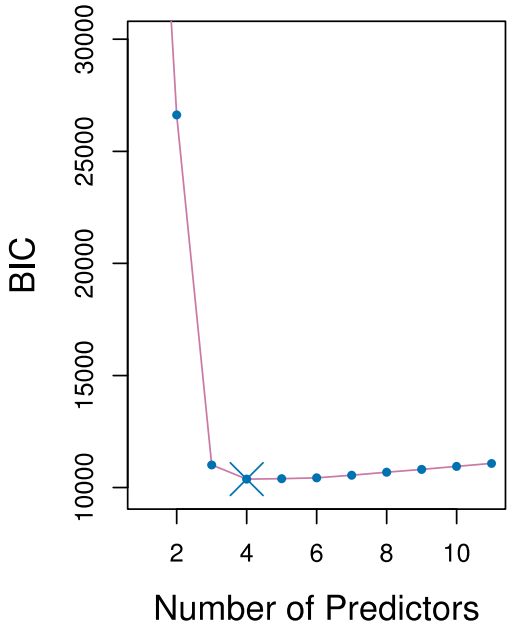
Approach 1

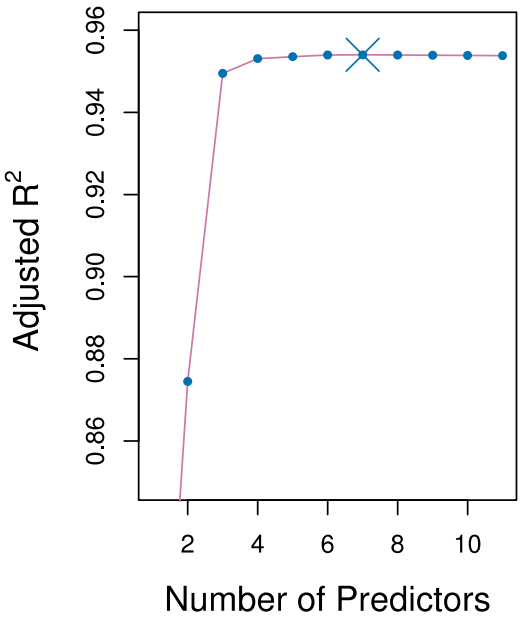
- For AIC, BIC and C_p, a smaller value indicates lower test error, and thus a better model.
- For adjusted R^2, a larger value indicates a better model.
- AIC vs. BIC: BIC places a heavier penalty on models with many predictors, hence results in the selection of smaller models.
- Adjusted R^2 is not as well motivated in statistical theory as other three.
- All of these measures are simple to use and compute.

Results of Credit Data: C_p, BIC and Adjusted R^2

- C_p (AIC) selects the 6-variable model containing income, limit, rating, cards, age and student.
- BIC selects the 4-variable model containing only income, limit, cards and student.
- Adjusted R^2 selects the 7-variable model containing those selected by C_p (AIC) and gender.





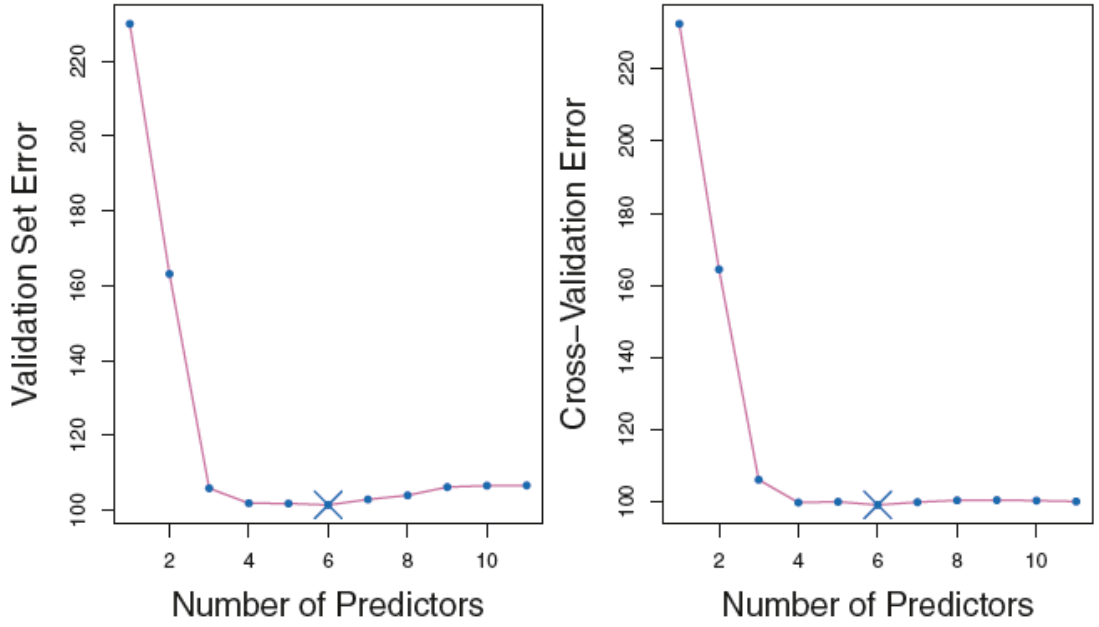


Approach 2

- Compute the validation set error or the cross-validation error for each model, and then select the model that results in smallest test error.
- Advantage over Approach 1: it provides a direct estimate of the test error, and makes fewer assumptions about the true underlying model (Approach 1 mainly applies to linear regression).
- In the past, performing cross validation was computationally prohibitive for problems with large p and/or n, so those simple measures, AIC, BIC, C_p and adjusted R^2, were often used. Nowadays, with fast computers, cross validation becomes more attractive.

Results of Credit Data: Cross Validation

- Validation set and cross validation methods both select the 6-predictor model.
- The two methods suggest that the 4-, 5- and 6-predictor models are roughly equivalent in terms of test error.



Shrinkage Methods

Shrinkage Methods

- Subset selection methods fit a linear model that contains a subset of the predictors through least squares.
- Shrinkage methods fit a model containing all p predictors using a technique that shrinks the coefficient estimates towards zero.
- Two popular shrinkage methods: Ridge regression and LASSO.

1. Ridge Regression

- Ordinary Least Squares (OLS) estimates βs by minimizing

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

- Ridge Regression uses a slightly different equation

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

$$\lambda \geq 0 \text{ is a } \textit{tuning parameter}$$

Ridge Regression Adds a Penalty on βs !

- The effect of this equation is to add a penalty of the form

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

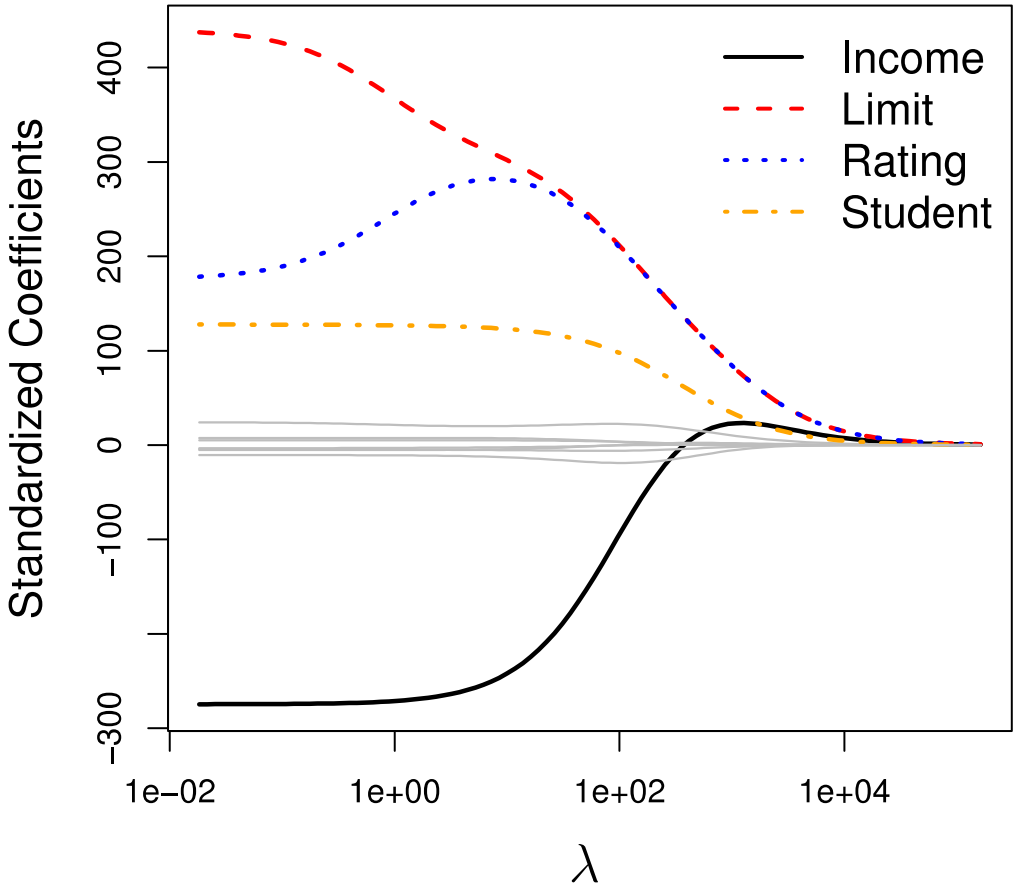
where the tuning parameter λ is a positive value.

- This has the effect of "*shrinking*" large βs towards zero.

- Notice that when $\lambda = 0$, we get the OLS!

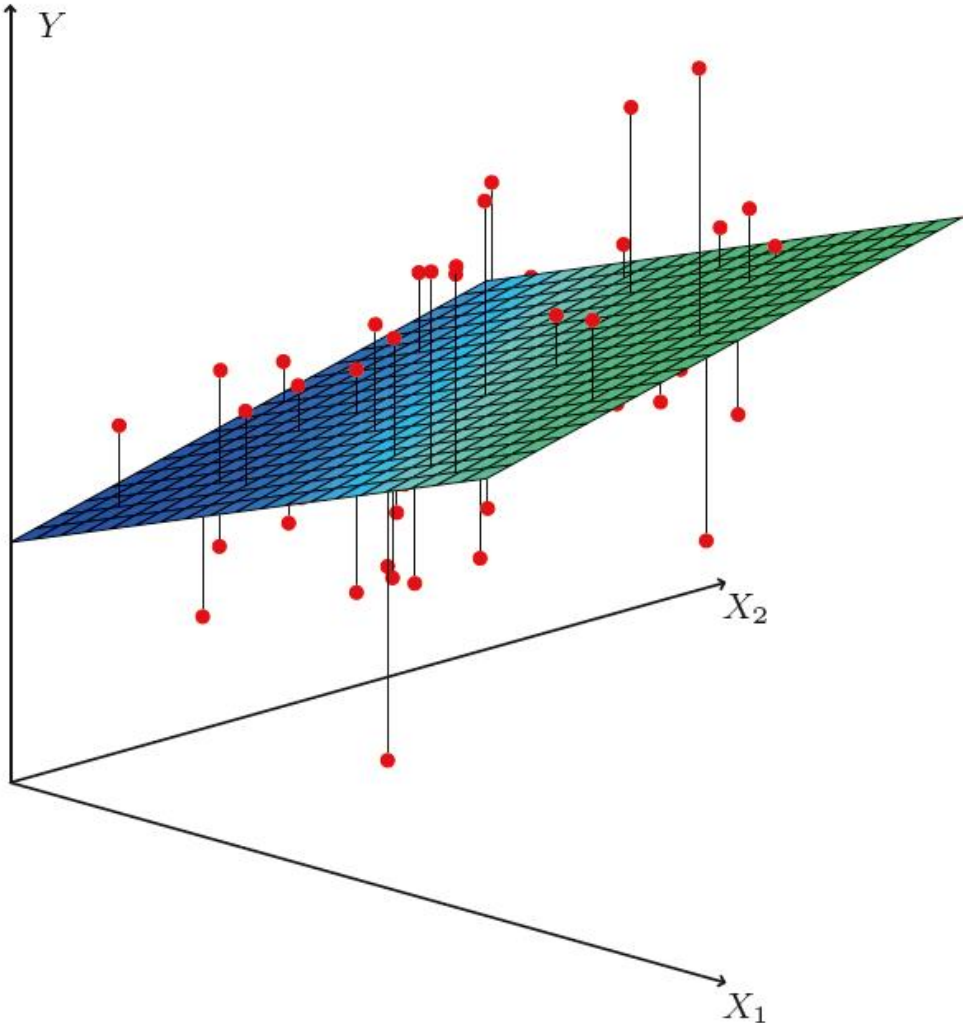
Credit Data: Ridge Regression

- As λ increases, the coefficients shrink towards zero.



What Makes the Method Work?

- The OLS estimates to minimize RSS



Another Formulation for Ridge Regression

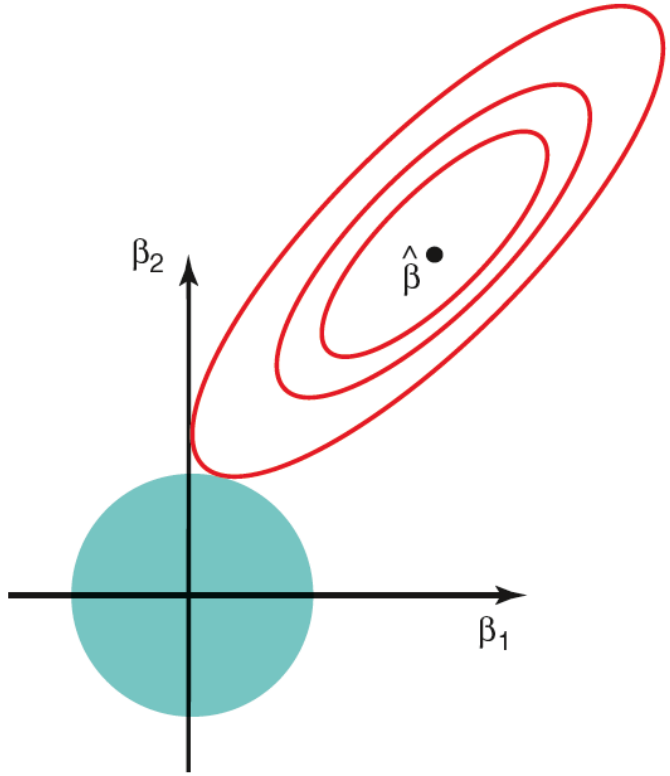
- The two formulations are equivalent

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

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

Ridge Regression Estimates

- Ellipses that are centered around $\hat{\beta}$ represent regions of constant RSS (i.e., all the points on a given ellipse share a common value of the RSS).
- The solid blue area is the constraint region of ridge regression, which is $\beta_1^2 + \beta_2^2 \leq s$ in the two-dimension case.
- The ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint region.



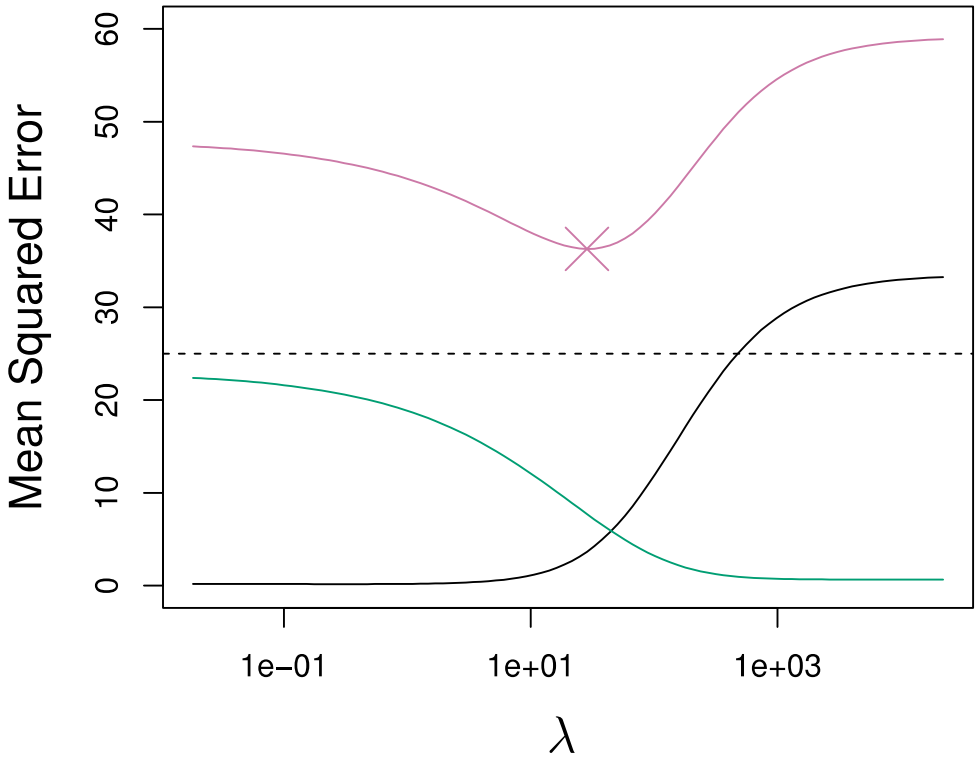
Why Shrinking towards Zero Is A Good Thing?

- OLS estimates
  - Generally have low bias but can have high variance
  - When n and p are of similar size or $n < p$, the OLS estimates will be extremely variable

- The penalty term in the ridge regression estimates
  - Increases bias
  - But can substantially reduce variance

Simulation Study

- Black: Bias
- Green: Variance
- Purple: Test MSE
- Increase in λ increases bias but decreases variance



Computational Advantages

- If p is large, the best subset selection approach requires searching through enormous numbers of possible models.
- With Ridge Regression, for any given λ, we only need to fit one model and the computation turns out to be very simple.
- Ridge Regression can even be used when $n < p$, a situation where OLS fails completely!

2. LASSO

- Ridge Regression isn't perfect.
- One significant problem: the penalty term will never force any of the coefficients to be exactly zero. That means the final model will include all variables, which makes it hard to interpret.
- A more modern alternative is the LASSO.
- The LASSO works in a similar way to Ridge Regression, except that it uses a different penalty term.

LASSO's Penalty Term

- Ridge Regression minimizes

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

- The LASSO minimizes

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

What's the Big Deal?

- This seems like a very similar idea but there is a big difference.
- Using this penalty, it could be proven mathematically that some coefficients end up being set to exactly zero.
- With LASSO, we can produce a model that has high predictive power and it is simple to interpret.

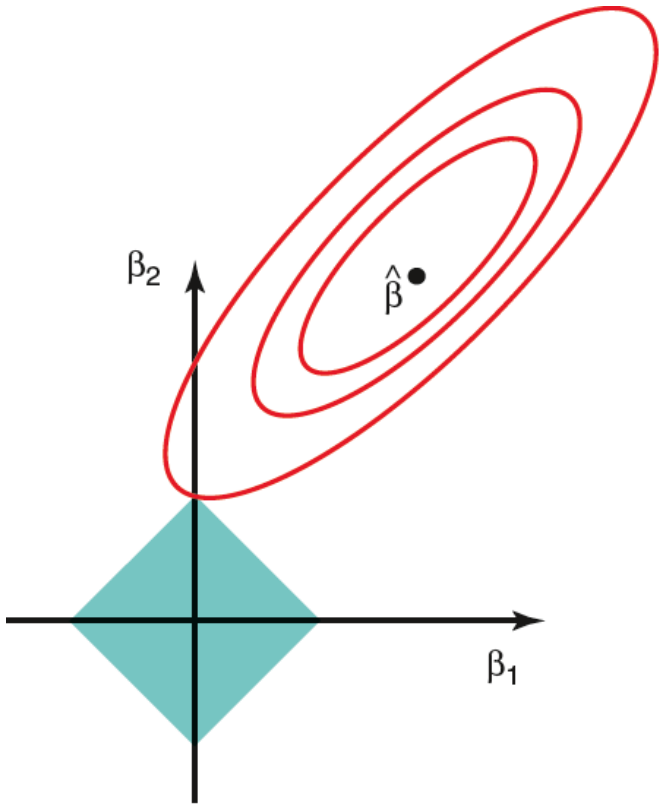
Another Formulation for LASSO

- The two formulations are equivalent

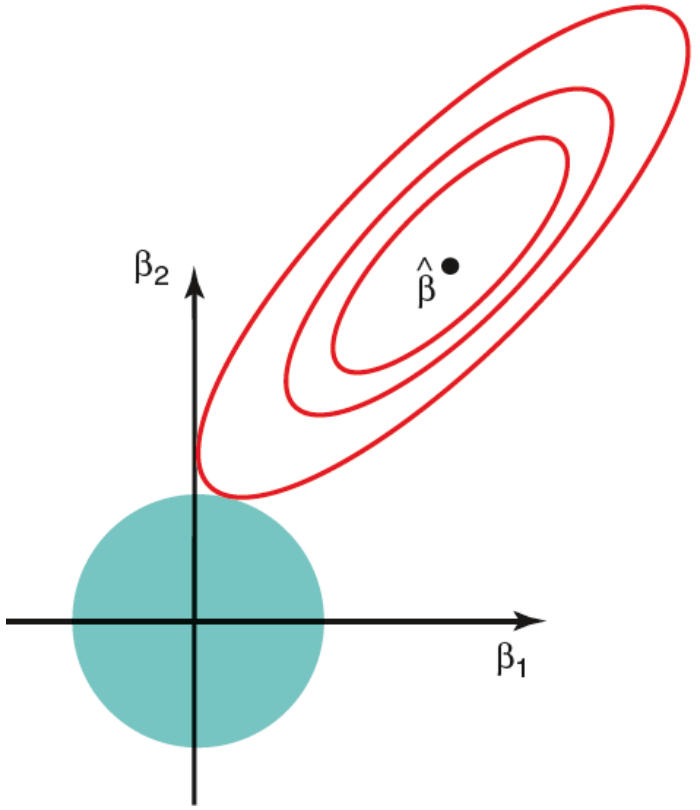
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

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

Reason for the Magic of LASSO



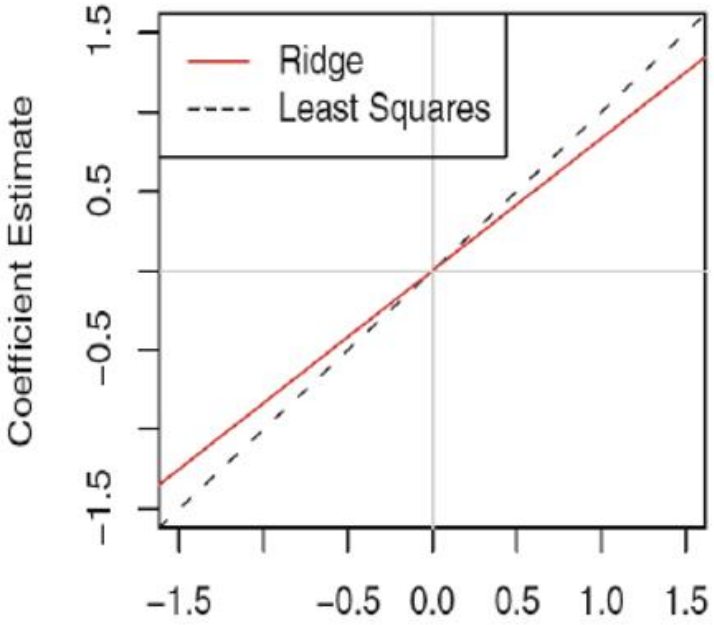
$|\beta_1| + |\beta_2| \leq s$

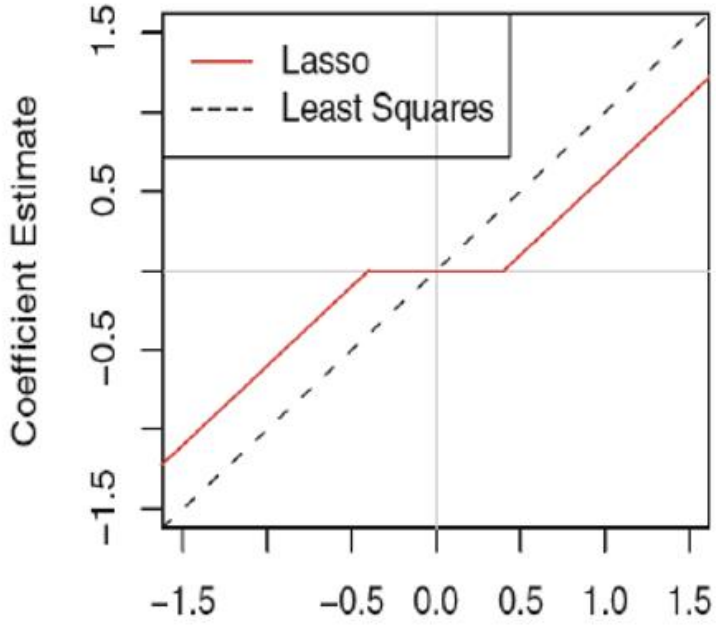


$\beta_1^2 + \beta_2^2 \leq s$

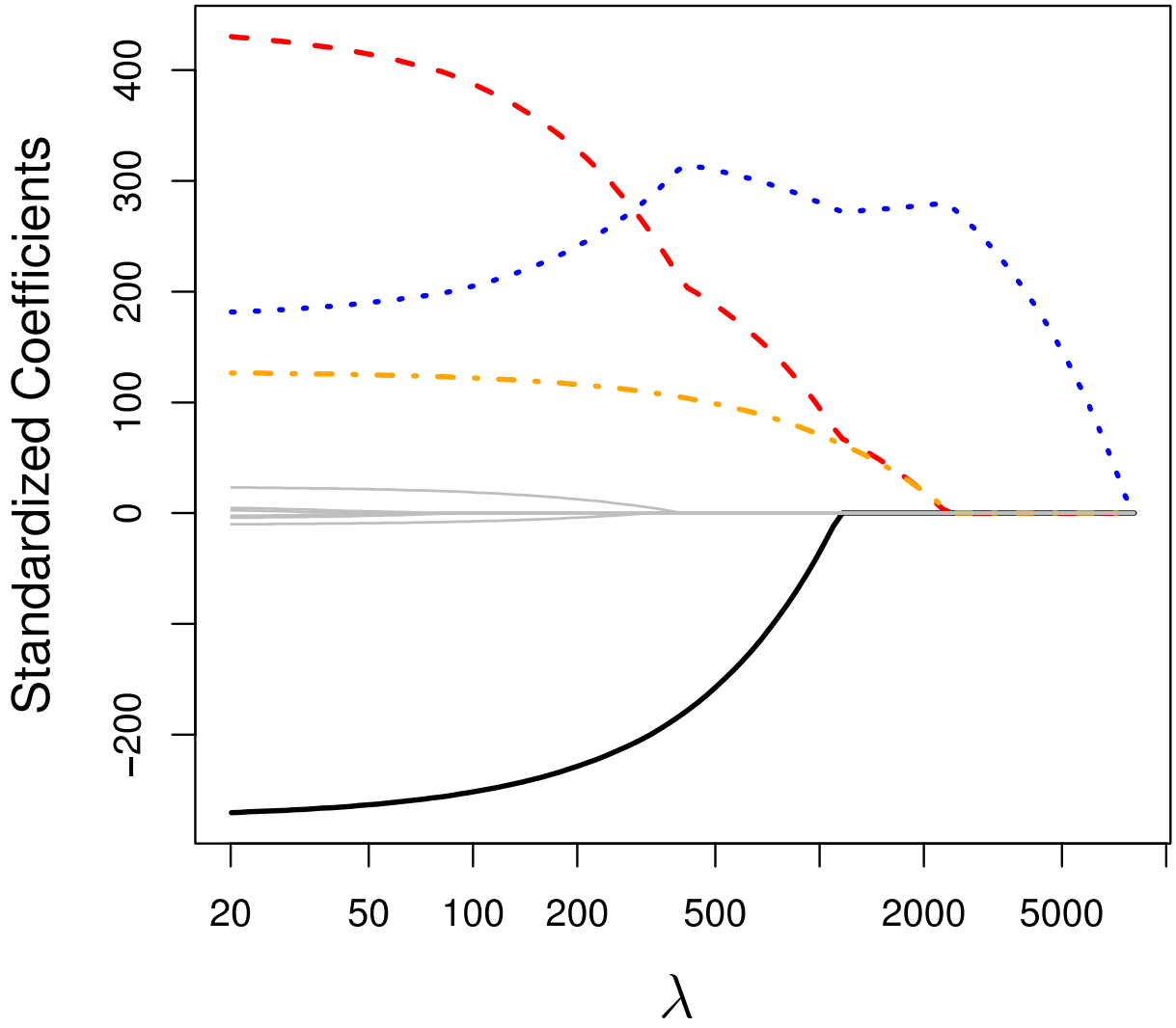
An Orthogonal Case

- Consider a simple case with $n = p$ and $\mathbf{X} = \mathbf{I}_p$, then $\hat{\beta}_j^{ols} = y_j$,
  - Ridge regression multiplies $\hat{\beta}_j^{ridge}$ by a constant, $\hat{\beta}_j^{ridge} = y_j/(1+\lambda)$.
  - Lasso truncates $\hat{\beta}_j^{ridge}$ towards zero by a constant, $\hat{\beta}_j^{lasso} = sign(y_j)\left(|y_j| - \lambda/2\right)_+$.





Credit Data: LASSO



Ridge Regression vs. LASSO

- Neither ridge regression nor the lasso will universally dominate the other in terms of prediction accuracy.
- Ridge regression performs better when the response depends on many predictors, all with coefficients of roughly equal size.
- LASSO performs better when a relatively small number of predictors have substantial coefficients, and the remaining predictors have very small or zero coefficients.

Selecting the Tuning Parameter

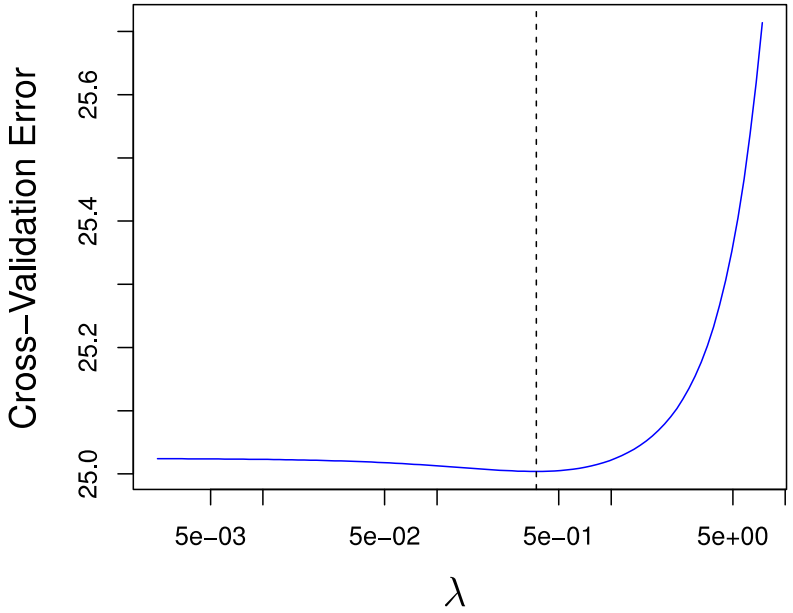
- In subset selection methods, we need to determine which model is the best.
- Similarly, in shrinkage methods, we need to select the optimal value for the tuning parameter λ .
- Different values of the tuning parameter will lead to a model with different levels of prediction accuracy.
- Question: Which value of λ will produce the model with lowest test error?

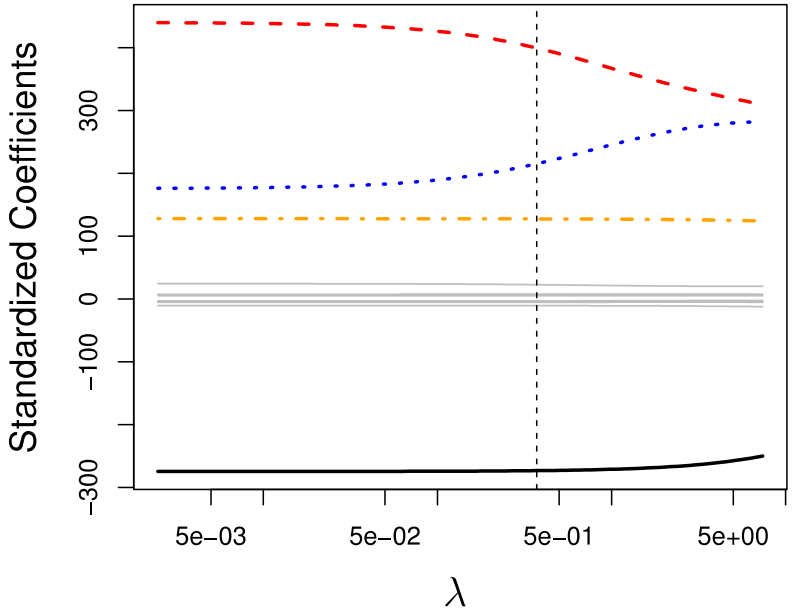
Approach — Cross Validation

- Select a grid of potential values for λ
- For each value of λ, use cross validation to estimate the test error rate
- Select the value that gives the lowest error rate

Credit Data: Selecting Tuning Parameter

- Leave-one-out cross validation (LOOCV) on Ridge regression





Other Extensions of LASSO

- Group lasso: if the p variables are partitioned into J groups, and then it is desirable to include or exclude the whole group

$$\min_{\beta} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} \left\| \vec{\beta}_j \right\|_2$$

where $\vec{\beta}_j$ is a coefficient vector for the jth group.

- Elastic net:

$$\min_{\beta} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

- Fused lasso:

$$\min_{\beta} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=2}^{p} \left\| \beta_j - \beta_{j-1} \right\|_1$$