LINEAR MODEL SELECTION AND REGULARIZATION

Chapter 06

Given a set of available features, how do we build the best set of features for our model?

- Subset Selection
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model
- Shrinkage Methods
  - Ridge Regression
  - The Lasso

Improving on the Least Squares Regression Estimates for models with many features

- Given a set of observations, in Linear Regression, the cost can be expressed as MSE or RSS or R^2
- Either least-squares fitting process or an iterative optimization picks coefficients that minimize this cost
- There are 2 reasons that coefficients selected using these cost estimates may not be ideal:
  1. Prediction Accuracy on non-training data
  2. Model Interpretability for features

Prediction Accuracy Problems

- The Linear Regression estimate has low variability especially when the relationship between Y and X is linear and the number of observations n is much larger than the number of predictors p $(n >> p)$

- But, when $n \approx p$, then the fit can have high variance and may result in overfitting and poor estimates on unseen observations – poor generalizability
- And, when $n < p$, then the variability of fit increases dramatically, and the variance of these estimates are unacceptable

Model Interpretability Problems

- When we have a large number of features in the model there will be many that have little or no influence on Y
- Leaving these variables in the model makes it harder to determine “important variables”
- The model would be easier to interpret by removing the unimportant variables

Solution Concepts

- Subset Selection
  - Identify a subset of all p predictors which best predict the response Y, and then fit the model using only this subset
  - Methods: *best subset selection* and *stepwise selection*
- Regularization through coefficient Shrinkage
  - Penalize the model (new cost function element) for having non-zero estimates of coefficients -> pushes coefficients towards zero
  - This shrinkage *reduces the variance* **WHY?**
  - Some of the coefficients may shrink to exactly zero – helps with variable selection/interpretation
  - Methods: *Ridge regression* and the *LASSO*
- Dimension Reduction
  - Project all p predictors into an M-dimensional space where $M < p$, and then fit a linear regression model
  - Example: Principle Components Regression

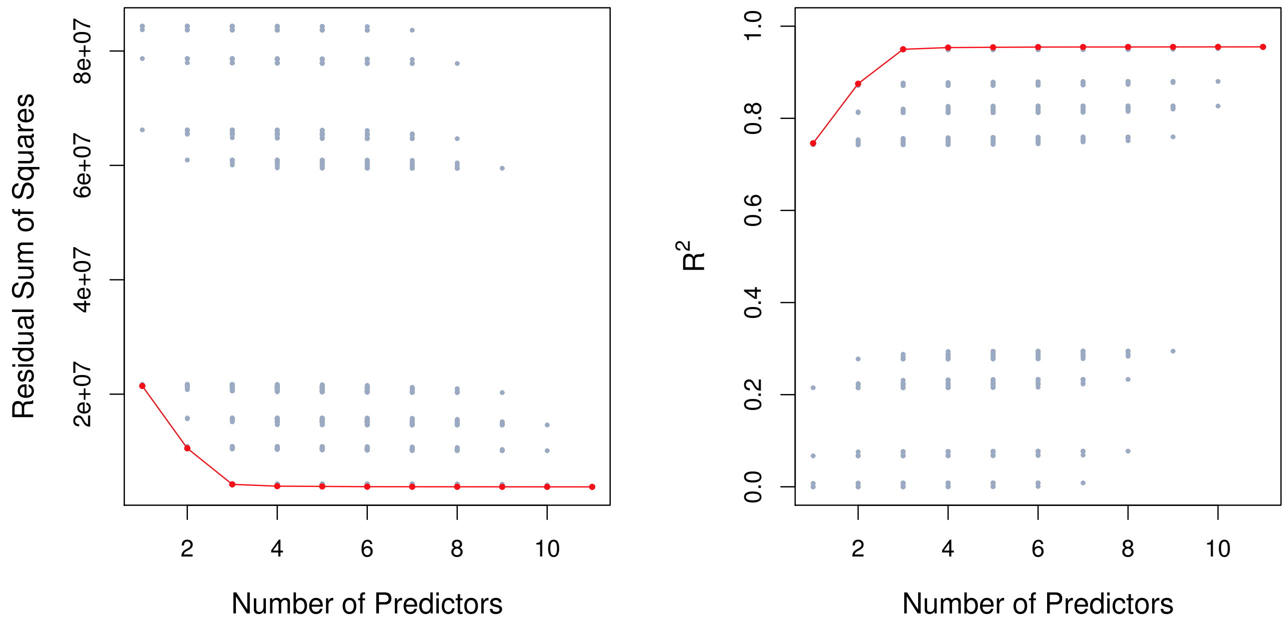
6.1 SUBSET SELECTION

6.6.1 Best Subset Selection

- Fit a linear regression model for each possible combination of the X predictors
- How do we judge which subset is the “best”?
- One simple approach is to take the subset with the smallest RSS or the largest R^2
- Unfortunately, one can show that the model that includes all the variables will always have the largest R^2 (and smallest RSS) **Why do you think this is?**

Credit Data: R^2 vs. Subset Size

- RSS will never increase (and R^2 will never decrease) as the number of variables increase - not very useful



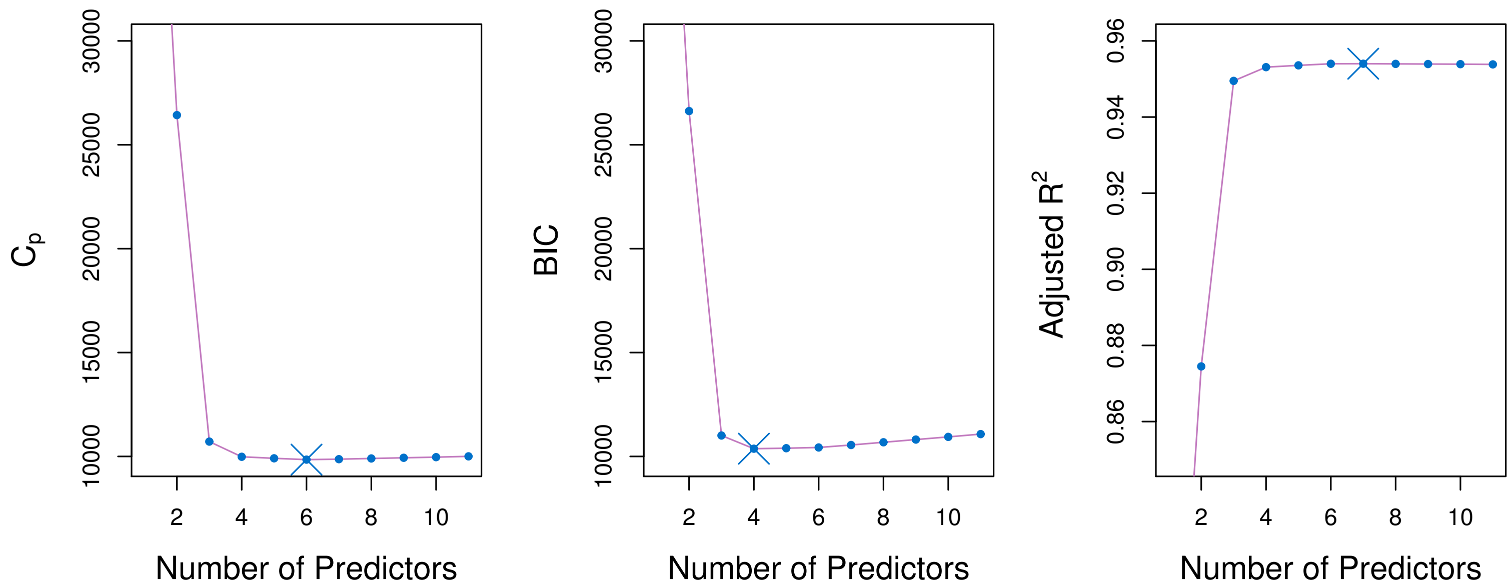
- Grey dots – actual performance of various subset models
- red line: the best model for a given number of predictors, according to RSS and R^2

Measures for *Estimating* model Performance on unseen data from *training* data fit

- To compare different models, adjust the RSS of the *training* data model fit based on some penalty for number of features
(p211-212):
  - Adjusted R^2
  - AIC (Akaike information criterion)
  - BIC (Bayesian information criterion)
  - C_p (Mallow's C_p: Proportional to AIC)
- These methods add penalty to RSS for the number of variables (i.e. complexity) in the model
- Estimates are made using model's fit of *training* data
- All are estimates…None are perfect

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

- A small value of C_p or BIC indicates a low error, and thus (hopefully) a better model
- A large value for the Adjusted R^2 indicates a better model



Feature Selection through Best Subset Selection

- Best Subset Selection considers all possible subsets of available features to find the optimal fit using validation data
- Select the model using the subset of features which yields the best performance on the (cross) validation data
  - E.g. best MSE or lowest classification error
- Concept Check: Compute O(·) for best subset selection as a function of p ...
  What is the number of possible feature subsets when there are p features available?

Feature Selection via Stepwise Selection

- Best Subset Selection is computationally intensive especially when we have a large number of predictors (large p)
- More computationally-attractive 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 occurs
  - Backward Stepwise Selection: Begins with the model containing all predictors, and then deleting one predictor at a time where the predictor chosen at each step is the *feature that causes the least degradation* to model performance when removed.
- Compute O(·) for these methods as a function of p :
  This can be thought of as a search (CSCE 523-style)
  What is the number of computations needed when there are p features available?

REGULARIZATION (Parameter Shrinkage) METHODS

6.2.1 Ridge Regression

- Ordinary Least Squares (OLS) estimates $\beta's$ 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 minimization equation which adds a term…

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

Math-Sense Check: Describe the influence of the last term in this equation

Ridge Regression Adds a Penalty on $\beta's$

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

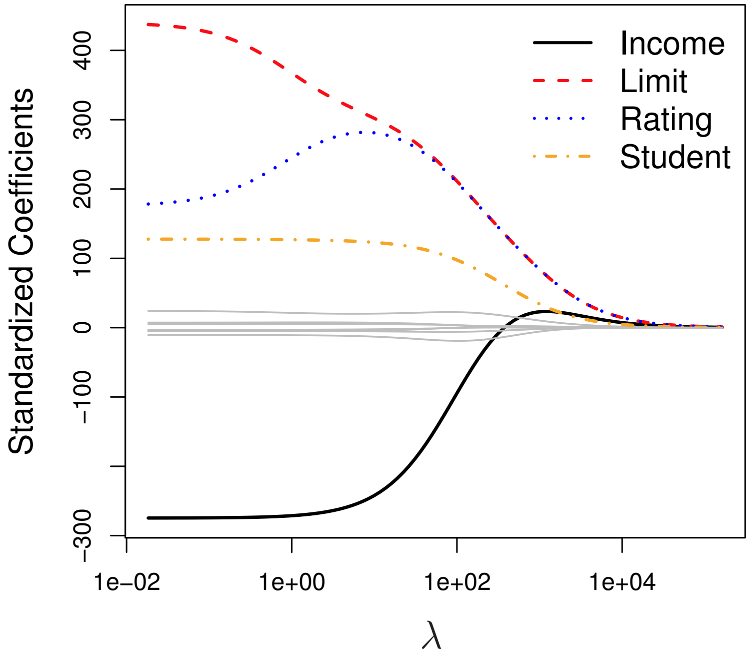
$$\lambda \sum_{j=1}^{p} \beta_j^2,$$

Where the tuning parameter λ is a positive value.

- This has the effect of “shrinking” large values of $\beta's$ towards zero.
- This penalty should improve the fit because shrinking the coefficients can significantly reduce their variance
- When $\lambda = 0$, we get the original RSS from Ordinary Least Squares

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

Credit Data: Ridge Regression



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

- As λ increases, the standardized coefficients shrinks *towards* zero.
- **Will coefficients ever reach zero?**
- **If not, what are the implications with model interpretability?**

Why can shrinking towards zero be a good thing to do?

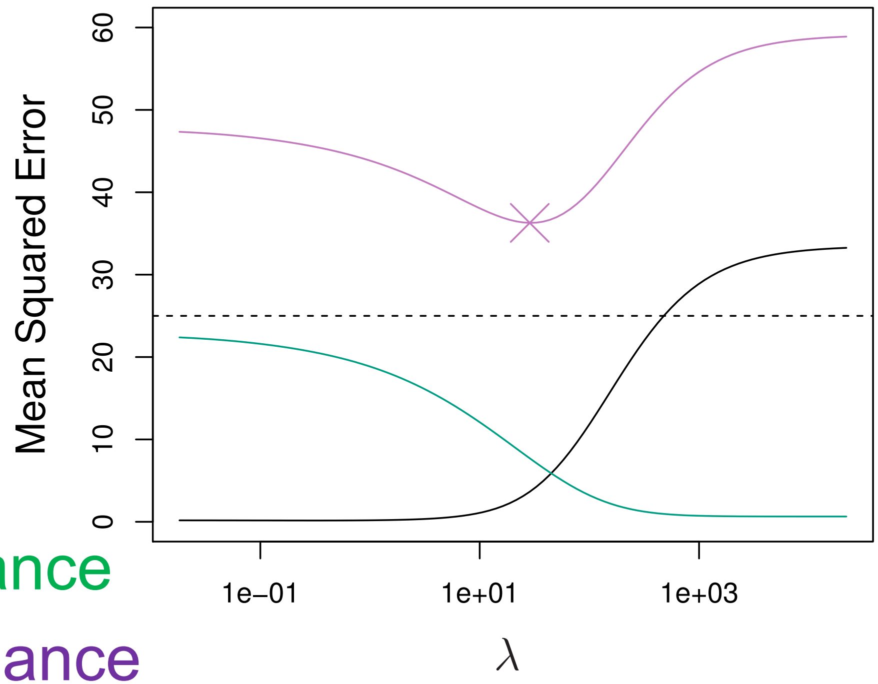
- It turns out that the parameter estimates generally have low bias but can be highly variable. In particular when n and p are of similar size or when $n < p$, then the OLS estimates will be extremely variable. **WHY?**

- The penalty term makes the ridge regression estimates biased but can also substantially reduce variance

- Thus, there is a bias / variance trade-off

Ridge Regression Bias / Variance

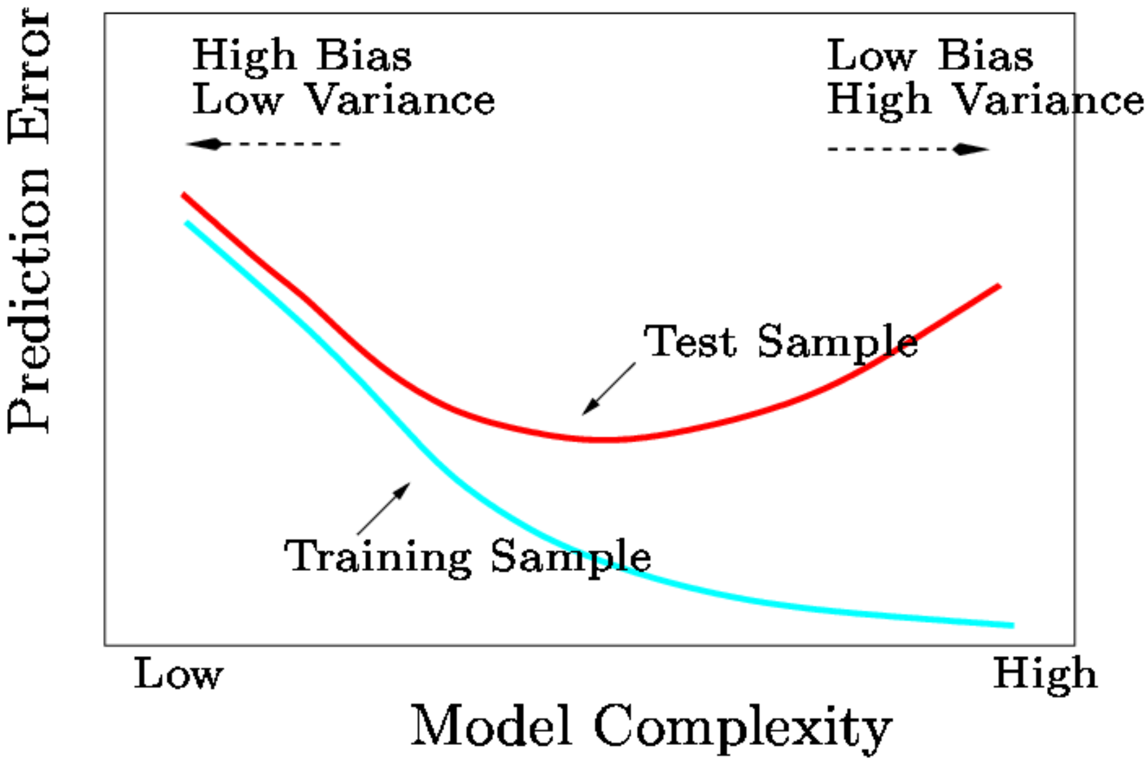
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2,$$



- Black: MSE due to Bias
- Green: MSE due to Variance
- Purple: MSE ~ Bias+Variance
- Increase in λ increases bias but decreases variance

Bias / Variance Trade-off

- In general, the ridge regression estimates will be more biased than the OLS ones but have lower variance
- Ridge regression will work best in situations where the OLS estimates have high variance



Computational Advantages of Ridge Regression

- If p is large, then using the best subset selection approach requires searching through multitudes of possible models
- With Ridge Regression, for any given λ , we only need to fit one model
- Ridge Regression can even be used when $p > n$, a situation where OLS fails completely

6.2.2. The LASSO

- Ridge Regression isn't perfect
- One significant problem is that the penalty term will never force any of the coefficients to be *exactly* zero. Thus, the final model will include all variables, which makes it harder to interpret
- A more modern alternative is the LASSO:
  Least **A**bsolute **S**hrinkage and **S**election **O**perator
- The LASSO works in a similar way to Ridge Regression, except it uses a different penalty term

Ridge Regression vs. LASSO: Penalty Term

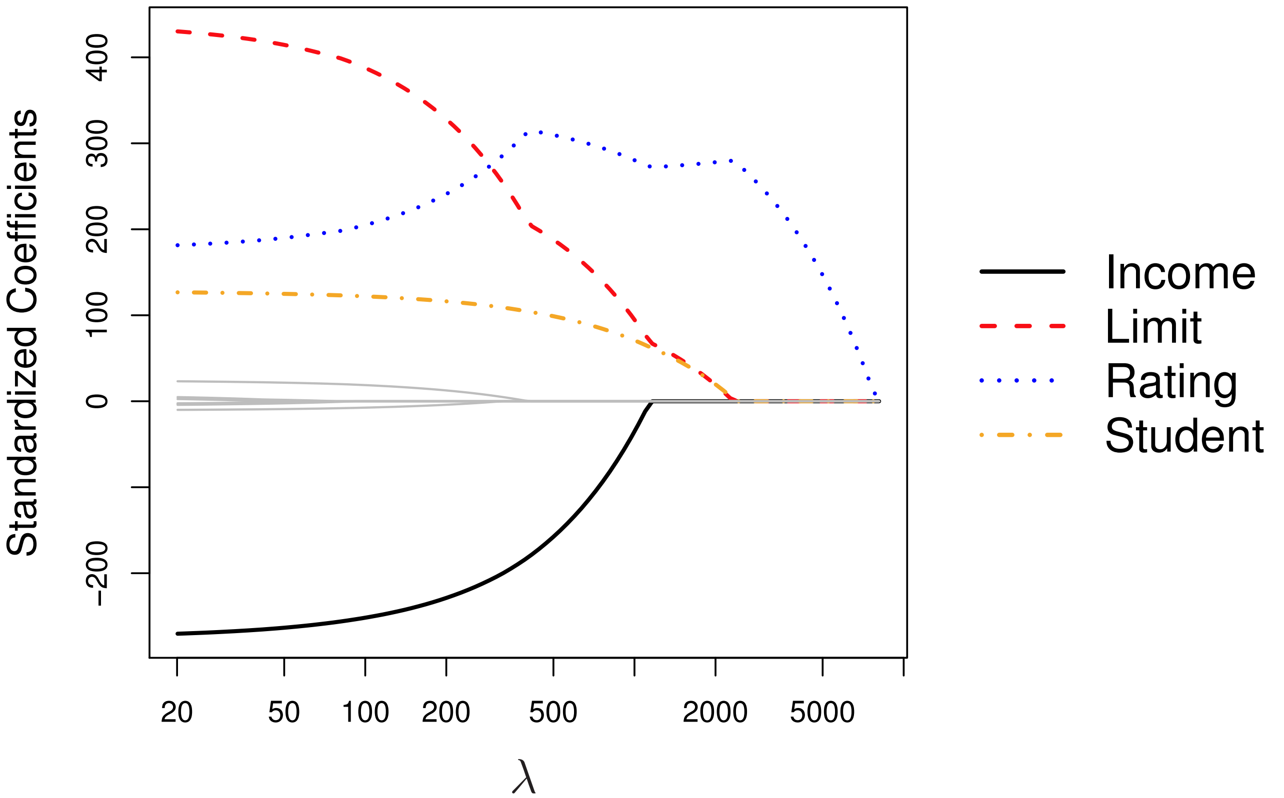
- Ridge Regression minimizes

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

- The LASSO estimates the $\beta's$ by minimizing

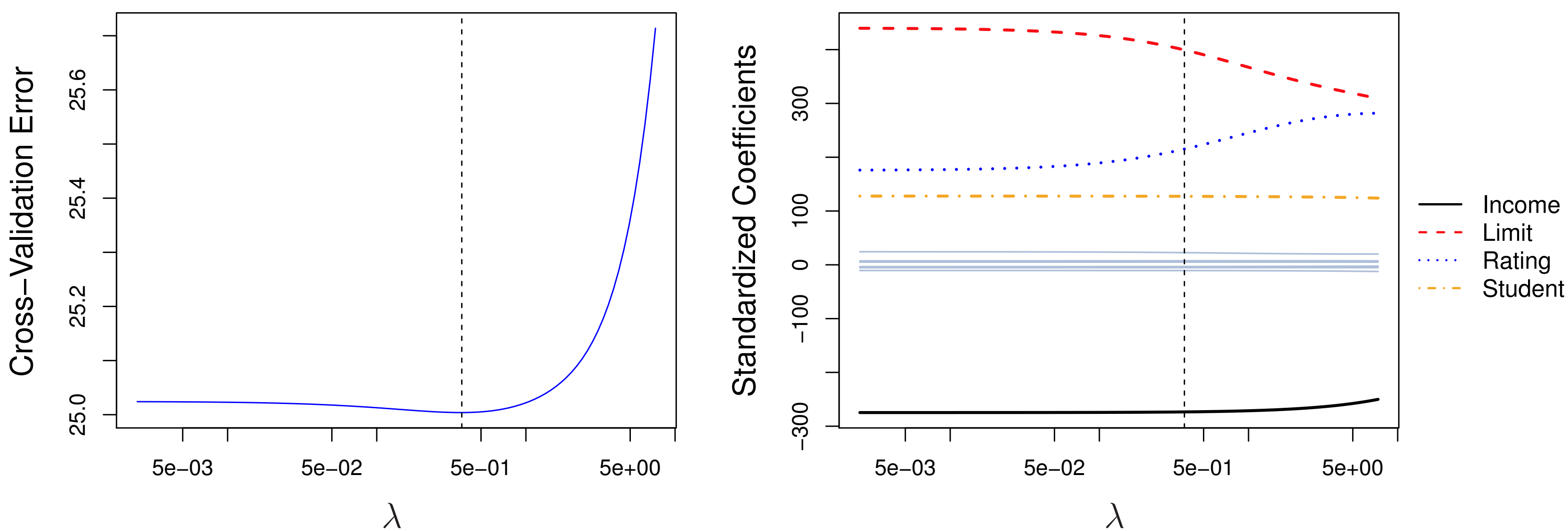
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Credit Data: LASSO



6.2.3 Selecting the Tuning Parameter for λ best performing model

- We need to decide on a value for λ
- Select a grid of potential values, use cross validation to determine error rate (for each value of λ) and select the lambda value that gives the lowest error rate



Benefits of LASSO

- 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 because some coefficients are driven to zero

-

- In this class we will show how to do this empirically
CLASS CODING EXERCISE (Regularization)