OUTLINE

- Regularization or Shrinkage Methods
  - Ridge Regression
  - The Lasso

IMPROVING LINEAR REGRESSION

- Multiple linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

$$\sum_{k=1}^{n} \left(y_k - \beta_0 - \beta_1 x_{1k} - \cdots - \beta_p x_{pk}\right)^2 \to min$$

- We want to improve the Linear Regression model, by replacing the least square fitting with some alternative fitting procedure.
- There are 2 reasons we might not prefer to just use the ordinary least squares (OLS) estimates
  1. Prediction Accuracy
  2. Model Interpretability

PREDICTION ACCURACY

- The least squares estimates have relatively low bias and low variance when the relationship between Y and X is linear and the number of observations n is bigger than the number of predictors p $(n \gg p)$.
- When $n \approx p$, then the least squares fit can have high variance and may result in overfitting and poor estimates on unseen observations.
- When $n < p$, then the variability of the least squares fit increases dramatically. There is no unique least squares coefficient estimate.

MODEL INTERPRETABILITY

- When we have a large number of variables X in the model there will be many that have little or no effect on Y
- Utilization of these variables 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 (i.e. setting the coefficients to zero) the unimportant variables

SOLUTION I

- Subset Selection
  - Identifying a subset of all predictors that we believe to be related to the response Y, and then fitting the model using this subset
  - E.g. best subset selection

SOLUTION II

- Shrinkage
  - Involves shrinking of the estimates of coefficients towards zero
  - This 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

SOLUTION III

- Dimension Reduction
  - Involves projecting all p predictors into an M-dimensional space where $M < p$, and then fitting linear regression model
  - E.g. Principle Components Regression

SHRINKAGE METHODS

Ridge Regression

The Lasso

RIDGE REGRESSION

- Ordinary Least Squares (OLS) estimates the coefficients by minimizing the RSS

$$RSS = \sum_{k=1}^{n} \left(y_k - \beta_0 - \beta_1 x_{1k} - \cdots - \beta_p x_{pk}\right)^2 \rightarrow min$$

- Ridge Regression uses a slightly different equation

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \rightarrow min$$

where $\lambda \geq 0$ is a tuning parameter. Selecting a good value for λ is critical.

RIDGE REGRESSION ADDS A PENALTY

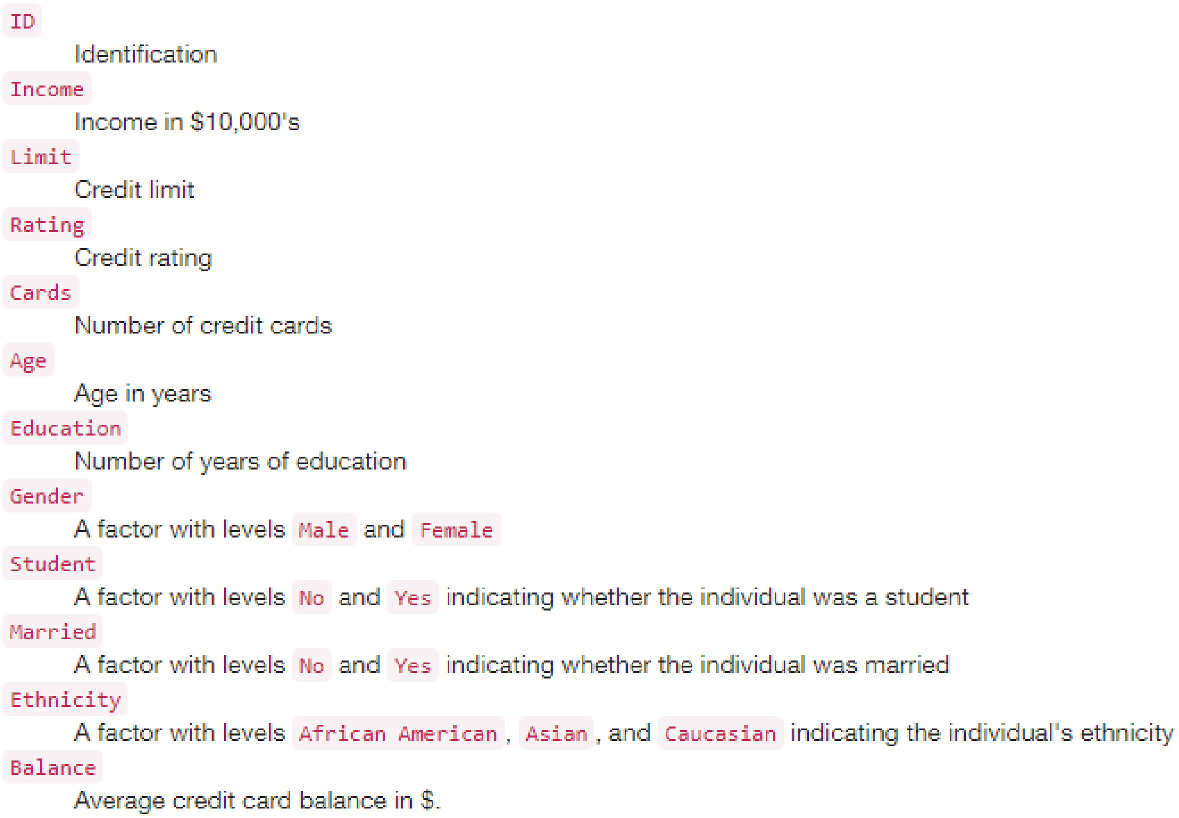
- The effect of this equation is to add a penalty of the form

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

- This has the effect of "shrinking" large values of the coefficients towards zero.
- It turns out that such a constraint should improve the fit, because shrinking the coefficients can significantly reduce their variance.
- Notice that when $\lambda = 0$, we get the OLS!

CREDIT DATA

- A simulated data set containing information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt



`ID`

Identification

`Income`

Income in $10,000's

`Limit`

Credit limit

`Rating`

Credit rating

`Cards`

Number of credit cards

`Age`

Age in years

`Education`

Number of years of education

`Gender`

A factor with levels `Male` and `Female`

`Student`

A factor with levels `No` and `Yes` indicating whether the individual was a student

`Married`

A factor with levels `No` and `Yes` indicating whether the individual was married

`Ethnicity`

A factor with levels `African American`, `Asian`, and `Caucasian` indicating the individual's ethnicity

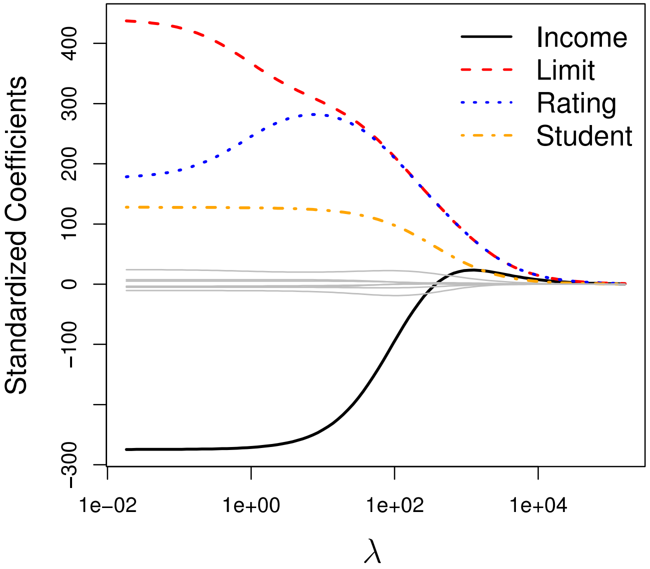
`Balance`

Average credit card balance in $.

CREDIT DATA: RIDGE REGRESSION

- As λ increases, the standardized coefficients shrinks towards zero
- Standardized means that all variables are normalized (scaled) by dividing to their standard deviations before performing a regression

$$Balance \sim .$$

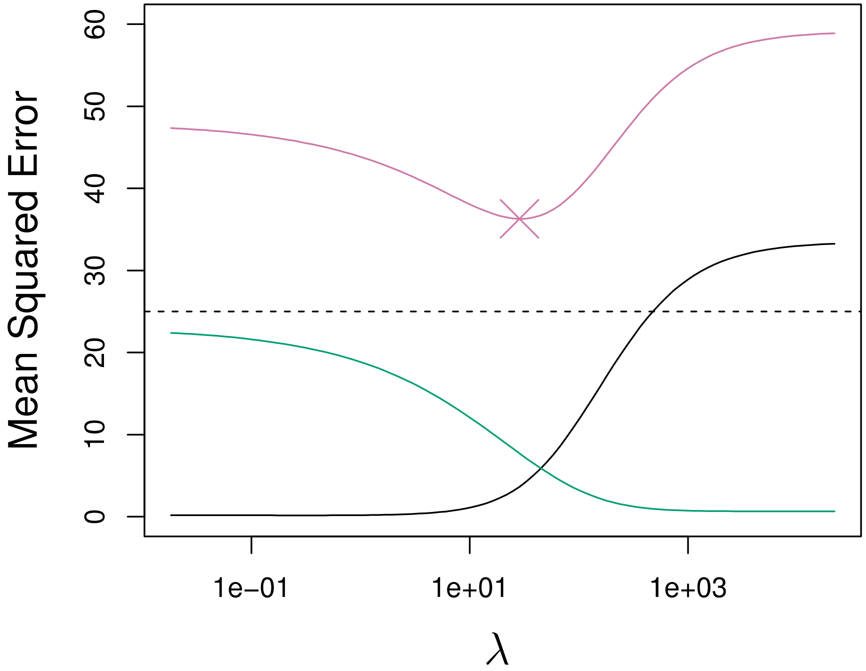


WHY CAN SHRINKING TOWARDS ZERO BE A GOOD THING TO DO?

- It turns out that the OLS 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
- The penalty term makes the ridge regression estimates biased but can also substantially reduce variance
- As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias
- Thus, there is a bias/variance trade-off

RIDGE REGRESSION BIAS/VARIANCE

- Black: Squared Bias
- Green: Variance
- Purple: MSE
- Dashed: Minimum Possible MSE



COMPUTATIONAL ADVANTAGES OF RIDGE REGRESSION

- If p is large, then using 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 computations turn out to be very simple

- Ridge Regression can even be used when $p > n$, a situation where OLS fails completely!

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

- The LASSO works in a similar way as Ridge Regression, except it uses a different penalty term

LASSO'S PENALTY TERM

- Ridge Regression minimizes

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \rightarrow min$$

- The LASSO minimizes

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \rightarrow min$$

WHAT'S THE BIG DEAL?

- This seems like a very similar idea but there is a big difference

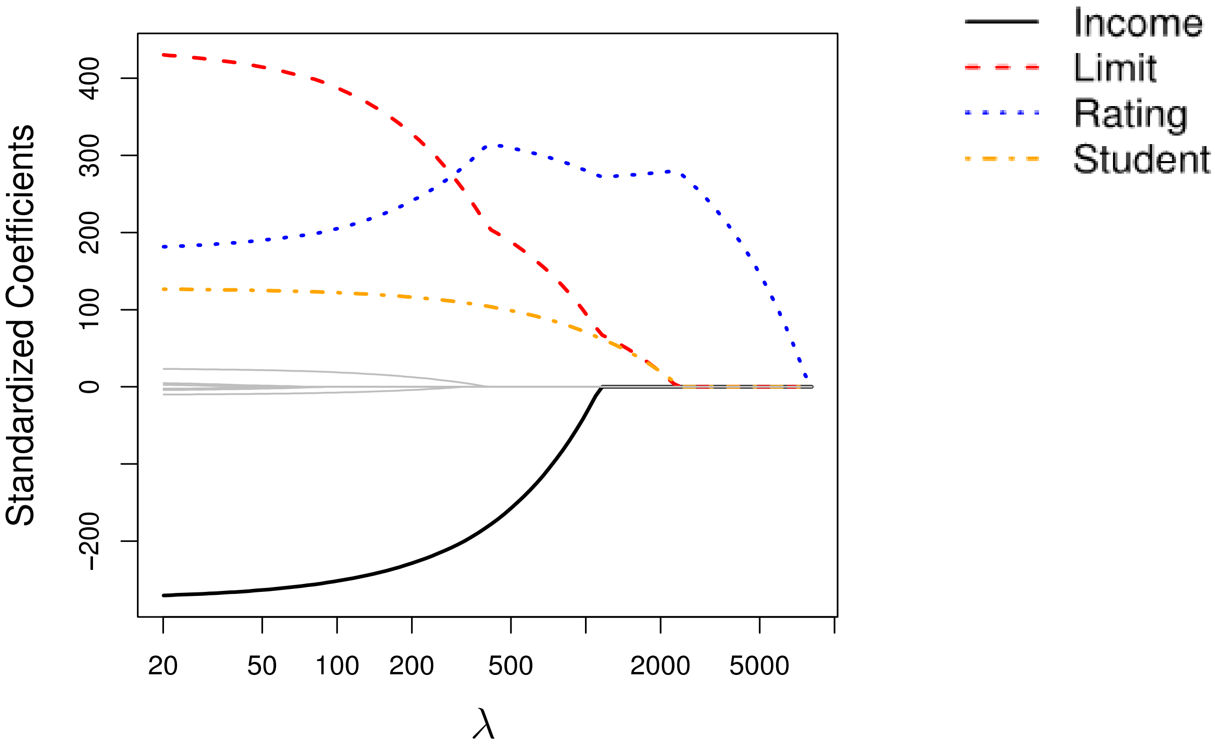
- Using this penalty, it could be proven that some coefficients end up being set to exactly zero

- With Lasso, we can produce a model that has high predictive power and it is simple to interpret

- The Lasso performs variable selection yielding sparse models

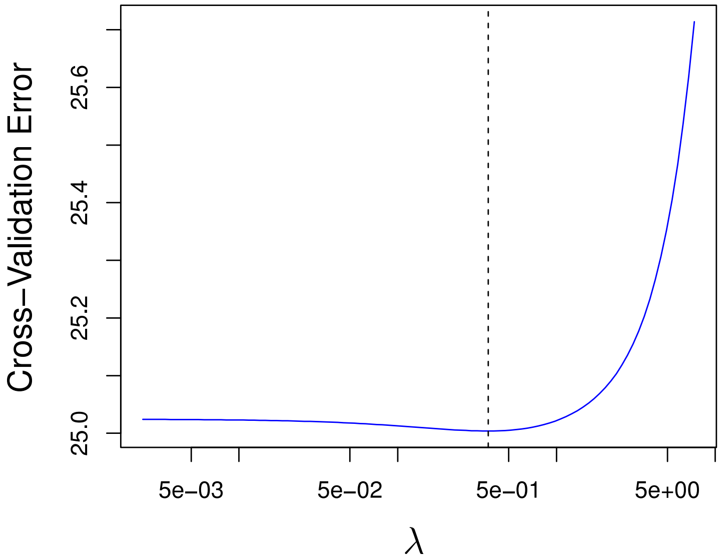
CREDIT DATA: LASSO

$Balance \sim .$

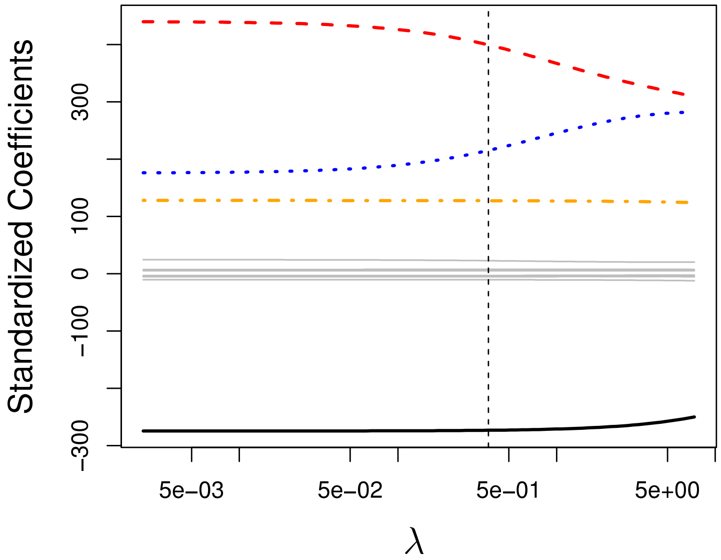


SELECTING THE TUNING PARAMETER λ

- We need to decide on a value for λ:
  - Select a grid of potential values
  - Use cross validation to estimate the error rate on test data (for each value of λ)
  - Select the value that gives the least error rate
- Below LOOCV for Credit data.



Ridge regression coefficients


Standardized Coefficients
5e−03
5e−02
5e−01
5e+00
λ