Chapter 6: Linear Model Selection and Regularization

Thiago G. Martins | NTNU & Verizon

Spring 2021

Recap

Standard Linear Models

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

or in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Least Square Fitting: Minimize the RSS

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

- Least squares and maximum likelihood estimator for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

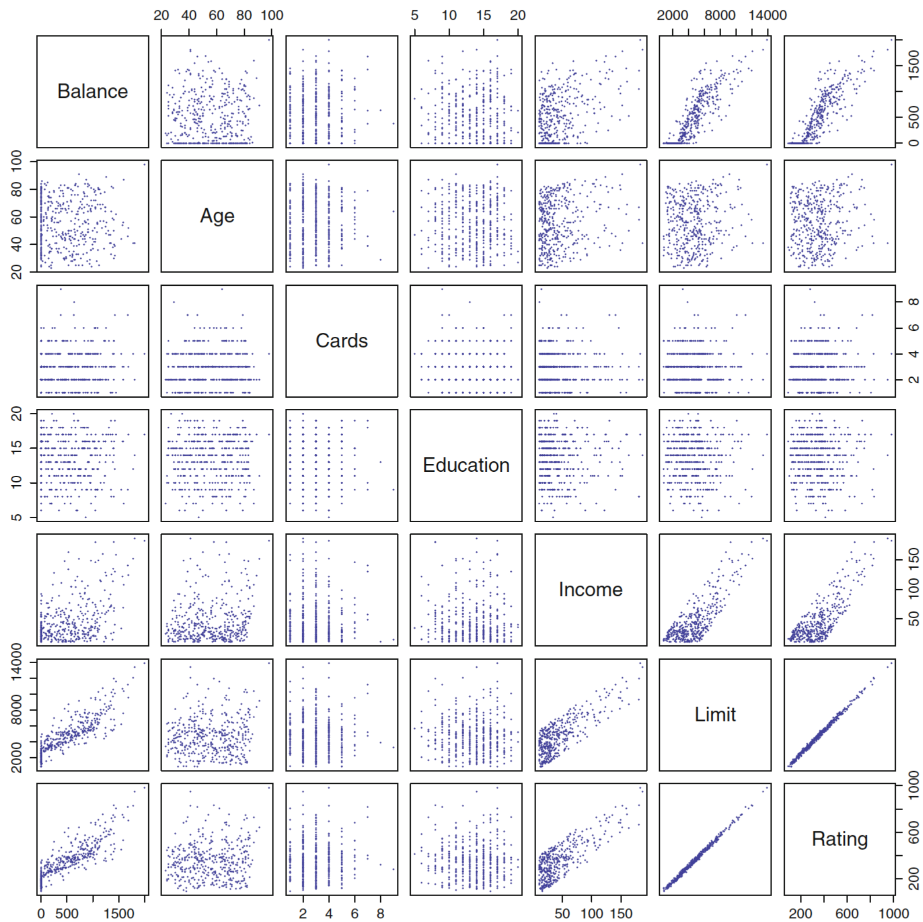
Recommended exercise 1

1. Show that the least square estimator of a standard linear model is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

2. Show that the maximum likelihood estimator is equal to the least square estimator for the standard linear model.

Credit Dataset



Recommended exercise 2

Write R code to create a similar representation of the Credit data figure of the previous slide. That is, try to recreate a similar plot in R.

Introduction

Objective of the module

Improve linear models **prediction accuracy** and/or **model interpretability** by replacing least square fitting with some alternative fitting procedures.

Prediction accuracy …

… when using standard linear models

Assuming true relationship is approx. linear: **low bias**.

- n >> p: **low variance**
- n not much larger than p: **high variance**
- n < p: multiple solutions available, **infinite variance**, model cannot be used.

By constraining or shrinking the estimated coefficients:

- often substantially reduce the variance at the cost of a negligible increase in bias.
- better generalization for out of sample prediction

Model Interpretability

- Some or many of the variables might be irrelevant wrt the response variable
- Some of the discussed approaches lead to automatically performing feature/variable selection.

Outline

We will cover the following alternatives to using least squares to fit linear models

- **Subset Selection**: Identifying a subset of the p predictors that we believe to be related to the response.
- **Shrinkage:** fitting a model involving all p predictors with the estimated coefficients shrunken towards zero relative to the least squares estimates.
- **Dimension Reduction**: This approach involves projecting the p predictors into a M-dimensional subspace, where $M < p$.

Subset Selection

Subset Selection

Identifying a subset of the p predictors that we believe to be related to the response.

Outline:

- Best subset selection
- Stepwise model selection

Best Subset Selection

1. Fit a least square regression for each possible combination of the p predictors.

2. Look at all the resulting models and pick the best.

Number of models considered:

$$\binom{p}{1} + \binom{p}{2} + \ldots + \binom{p}{2} = 2^p$$

Best Subset Selection (Algorithm)

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.

Penalties to training error

- Indirectly estimate test error by penalizing the training error
  - C_p, AIC, BIC and Adjusted R^2
  - All of the options above add a penalty to the training error that increase with the number of predictors in the model
  - All have rigorous theoretical justifications that are beyond the scope of this course.

Penalties to training error (C_p and AIC)

- C_p

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

  - C_p is an unbiased estimator of the test MSE, if $\hat{\sigma}^2$ is an unbiased estimator of σ^2
  - lower C_p is better

- AIC

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

  - AIC is proportional to C_p in the case of the standard linear model.

Penalties to training error (BIC and Adjusted R^2)

- BIC

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

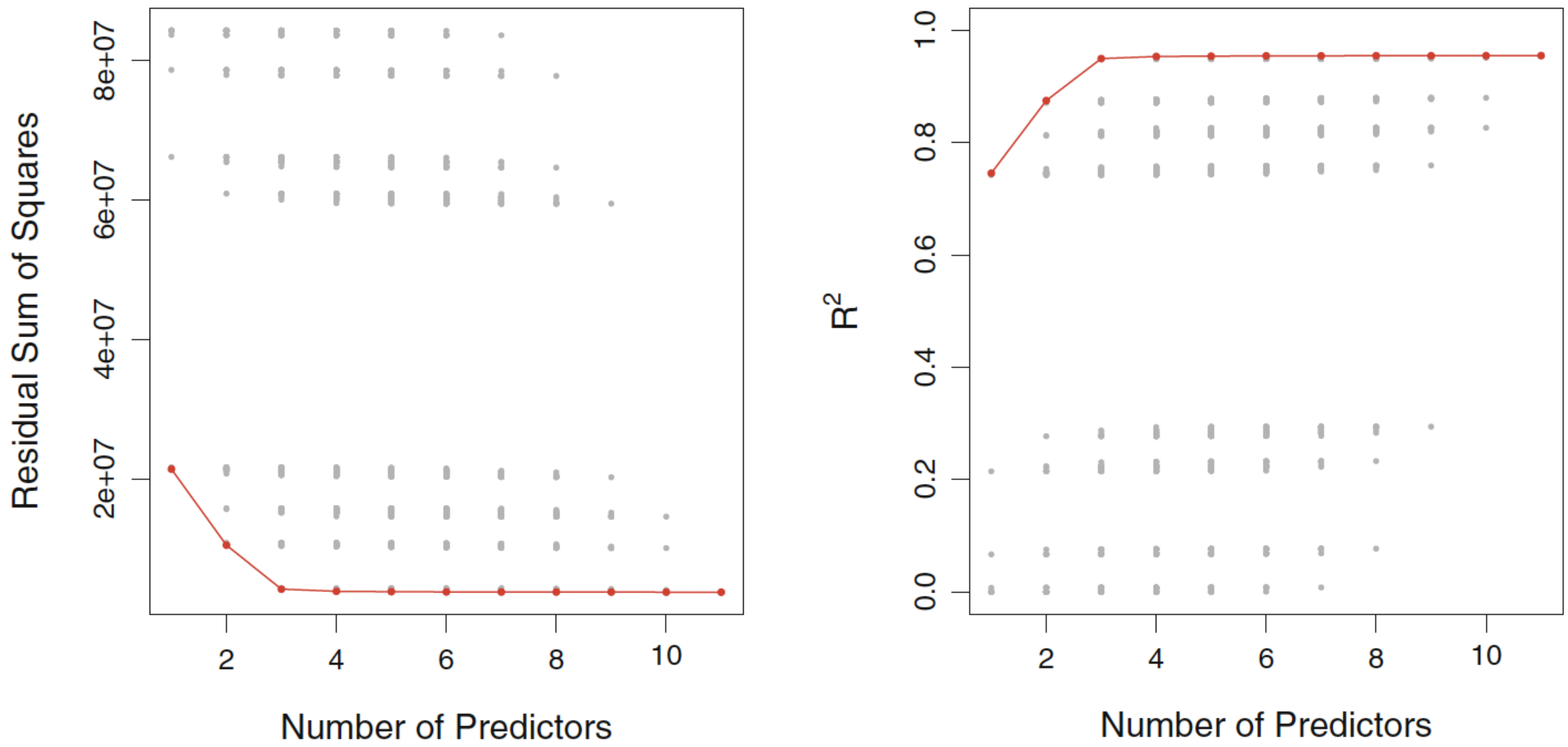
  - Derived from a Bayesian point of view.
  - the lower the better

- Adjusted R^2

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

  - the higher the better

Best Subset Selection (Credit Data Example)



Recommended exercise 3

1. For the Credit Dataset, pick the best model using Best Subset Selection according to C_p, BIC and Adjusted R^2
   - Hint: Use the `regsubsets()` of the `leaps` library, similar to what was done in Lab 1 of the book.
2. For the Credit Dataset, pick the best model using Best Subset Selection according to a 10-fold CV
   - Hint: Use the output obtained in the previous step and build your own CV function to pick the best model.
3. Compare the result obtained in Step 1 and Step 2.

Best Subset Selection (Drawbacks)

- Does not scale well -> the number of models to consider explode as p increases
  - $p = 10$ leads to approx. 1000 possibilities
  - $p = 20$ leads to over 1 million possibilities
- Large search space might lead to overfitting on training data

Stepwise selection

Add and/or remove one predictor at a time.

Methods outline:

- Forward Stepwise Selection
- Backward Stepwise Selection
- Hybrid approaches

Forward Stepwise Selection

- Starts with a model containing no predictors, $\mathcal{M}_0$
- Adds predictors to the model, one at time, until all of the predictors are in the model
    - $\mathcal{M}_1$, $\mathcal{M}_2$, ..., $\mathcal{M}_p$
- Select the best model among $\mathcal{M}_0$, $\mathcal{M}_1$, ..., $\mathcal{M}_p$

Forward Stepwise Selection (Algorithm)

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.

Forward Stepwise Selection (About the Algorithm)

- Goes from fitting 2^p models to $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models
- It is a guided search, we don't choose $1 + p(p+1)/2$ models to consider at random.
- Not guaranteed to yield the best model containing a subset of the p predictors.

- Forward stepwise selection can be applied even in the high-dimensional setting where $n < p$
    - By limiting the algorithm to submodels $\mathcal{M}_0, \ldots, \mathcal{M}_{n-1}$ only

Forward Stepwise Selection (Credit Data Example)

# 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`

Backward Stepwise Selection

- Starts with a model containing all predictors, $\mathcal{M}_p$.
- Iteratively removes the least useful predictor, one-at-a-time, until all the predictors have been removed.
  - $\mathcal{M}_{p-1}$, $\mathcal{M}_{p-2}$, …, $\mathcal{M}_0$
- Select the best model among $\mathcal{M}_0$, $\mathcal{M}_1$, …, $\mathcal{M}_p$

Backward Stepwise Selection (Algorithm)

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.

Backward Stepwise Selection (About the Algorithm)

- Similar properties to the Forward algorithm
  - Search $1 + p(p + 1)/2$ models instead of 2^p models
  - It is a guided search, we don't choose $1 + p(p + 1)/2$ models to consider at random.
  - Not guaranteed to yield the best model containing a subset of the p predictors.

- However, Backward selection requires that the number of samples n is larger than the number of variables p
  - So that the full model can be fit.

Hybrid Approach

- Similarly to forward selection, variables are added to the model sequentially.
- However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.
- Better model space exploration while retaining computational advantages of stepwise selection.

Recommended exercise 4

1. Select the best model for the Credit Data using Forward, Backward and Hybrid (sequential replacement) Stepwise Selection.
   - Hint: Use the `regsubsets()` of the `leaps` library
2. Compare with the results obtained with Best Subset Selection.

Shrinkage Methods

Shrinkage Methods

- 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.
- Reduce the number of effective parameters
  - While retaining the ability to capture the most interesting aspects of the problem.
- The two best-known techniques for shrinking the regression coefficients towards zero are:
  - the ridge regression.
  - the lasso.

Ridge regression

The ridge regression coefs β^R are the ones that minimize

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

with $\lambda > 0$ being a tuning parameter.

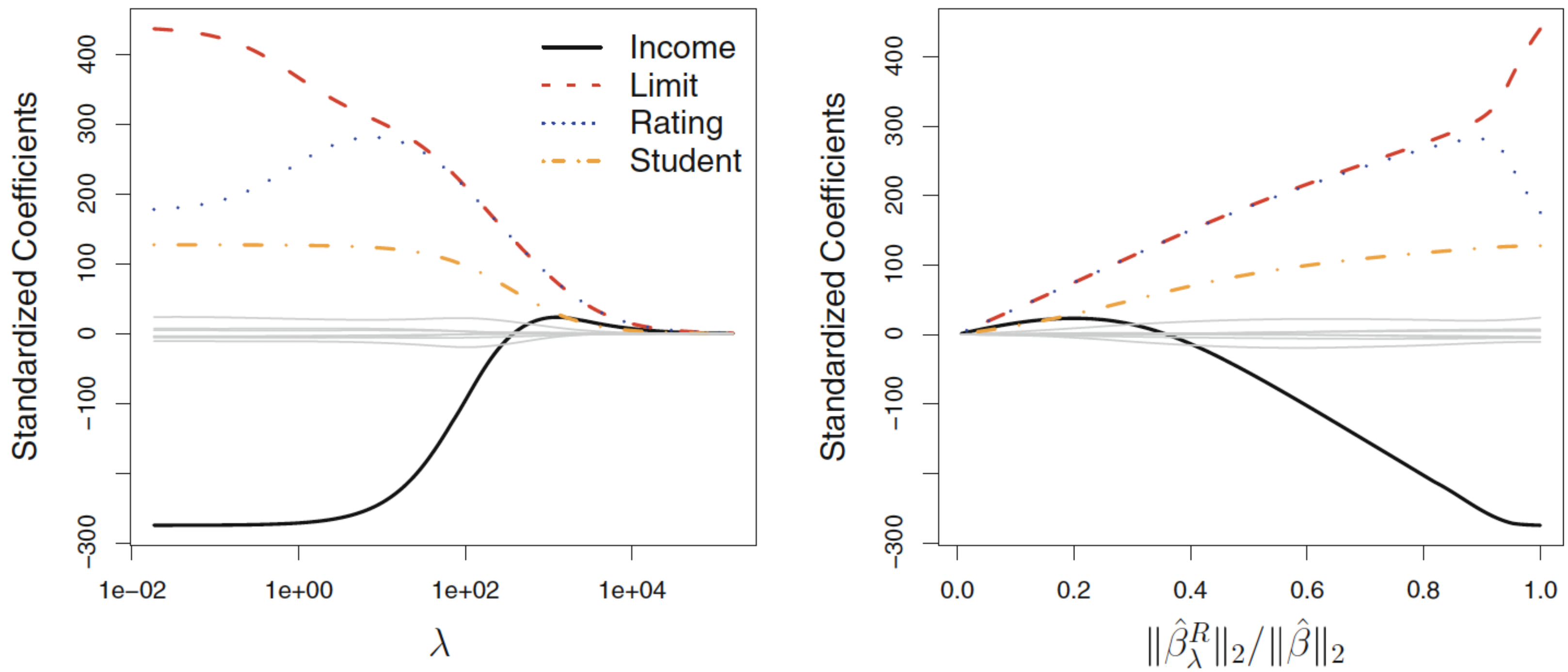
- Note that the penalty is not applied to the intercept, β_0.
  - If we included the intercept, β^R would depend on the average of the response.
  - We want to shrink the estimated association of each feature with the response.

Ridge regression

- Ridge regression are not scale-invariant
  - The standard least square are scale-invariant
  - β^R will not only depend on λ but also on the scaling of the jth predictor
  - Apply Ridge regression after standardizing the predictors

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

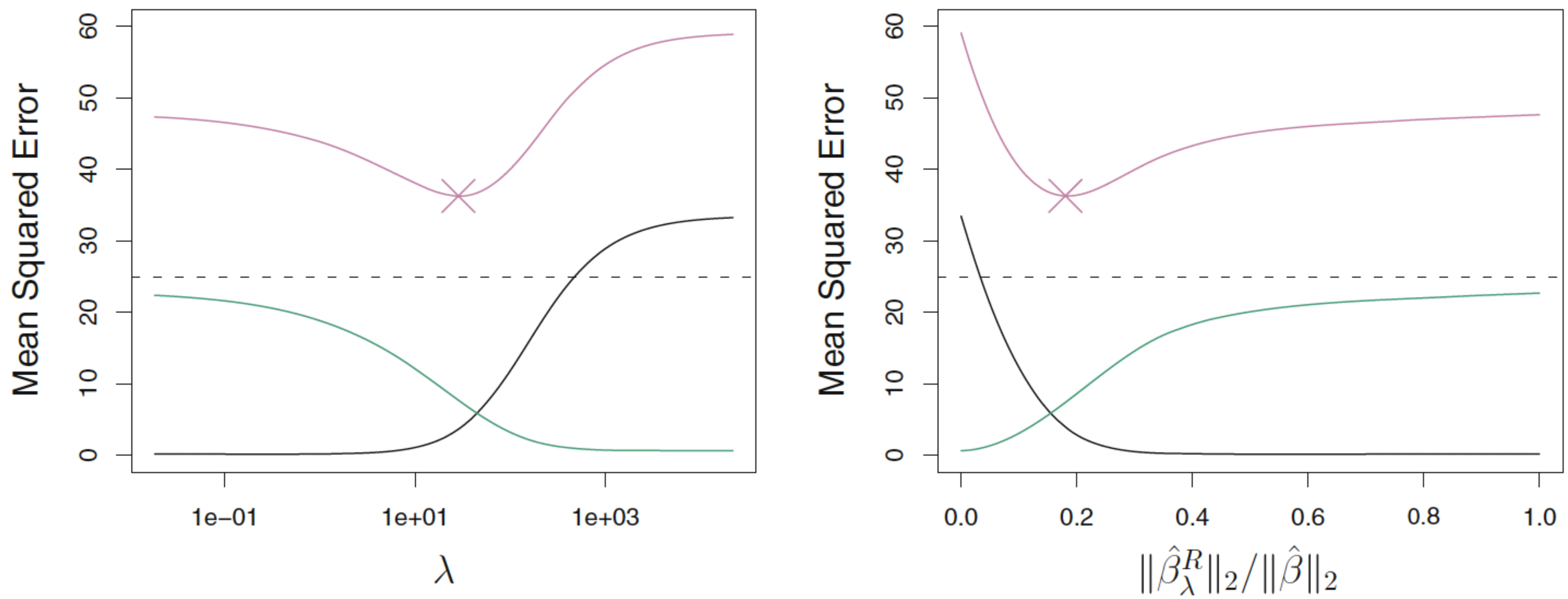
Ridge regression (Credit Data Example)



Ridge regression (Effectiveness)

- Why does it work?
  - As λ increase, the flexibility of the fit decreases.
  - Leading to a decrease variance but increased bias
- MSE is a function of the variance and the squared bias
  - Need to find sweet spot (see next Fig.)
- Therefore, ridge regression works best for the cases where
  - The relationship between covariates and response is close to linear (low bias)
  - And the least square estimates have high variance (high p in relation to n)

Ridge regression (MSE)



Ridge regression (Computationally efficient)

- The computations required to solve β_λ^R, simultaneously for all values of λ, are almost identical to those for fitting a model using least squares.
  - See (Friedman, Hastie, and Tibshirani 2010) and the references therein.

Ridge regression (Disadvantages)

- Unlike previous methods, ridge regression will include all p predictors in the final model.
  - The penalty λ 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 makes model interpretation hard for large p.

Recommended exercise 5

1. Apply Ridge regression to the Credit Dataset.
2. Compare the results with the standard linear regression.

Lasso

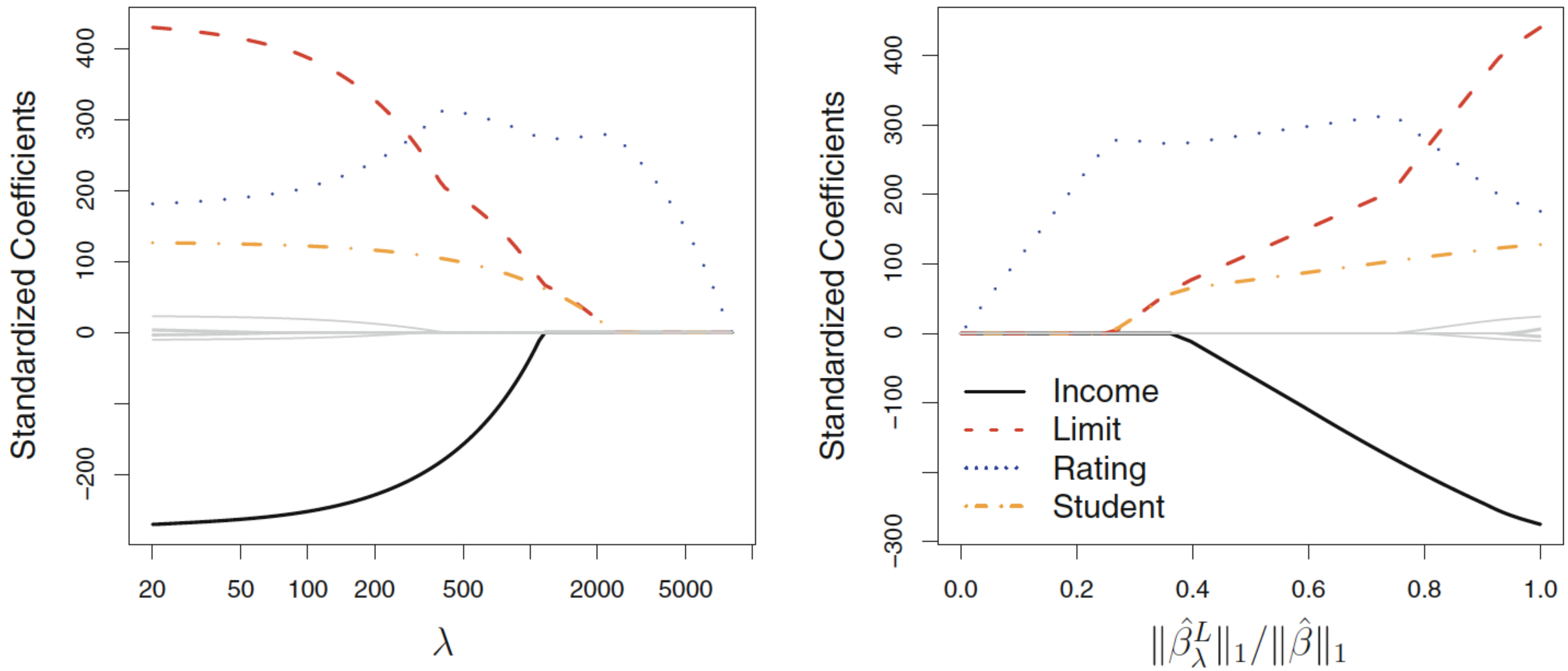
- The Lasso regression coefs β^L are the ones that minimize

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

with $\lambda > 0$ being a tuning parameter.

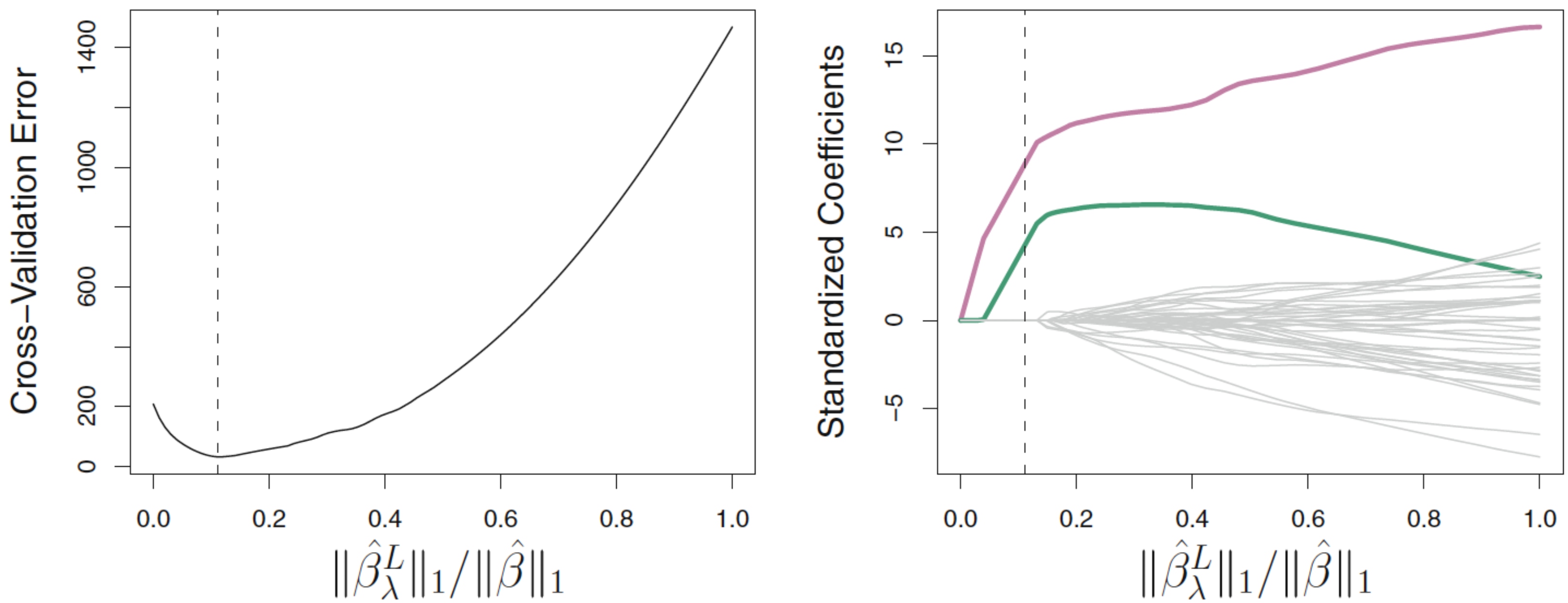
- Lasso also shrinks the coefficients towards zero
- In addition, the l_1 penalty has the effect of forcing some of the coefficients to be exactly zero when λ is large enough
- A geometric explanation will be presented in a future slide.

Lasso regression (Credit Data Example)



Lasso regression (Simulated Data Example)

- $p = 45$, $n = 50$ and 2 out of 45 predictors related to the response.



Ridge and Lasso (Different formulations)

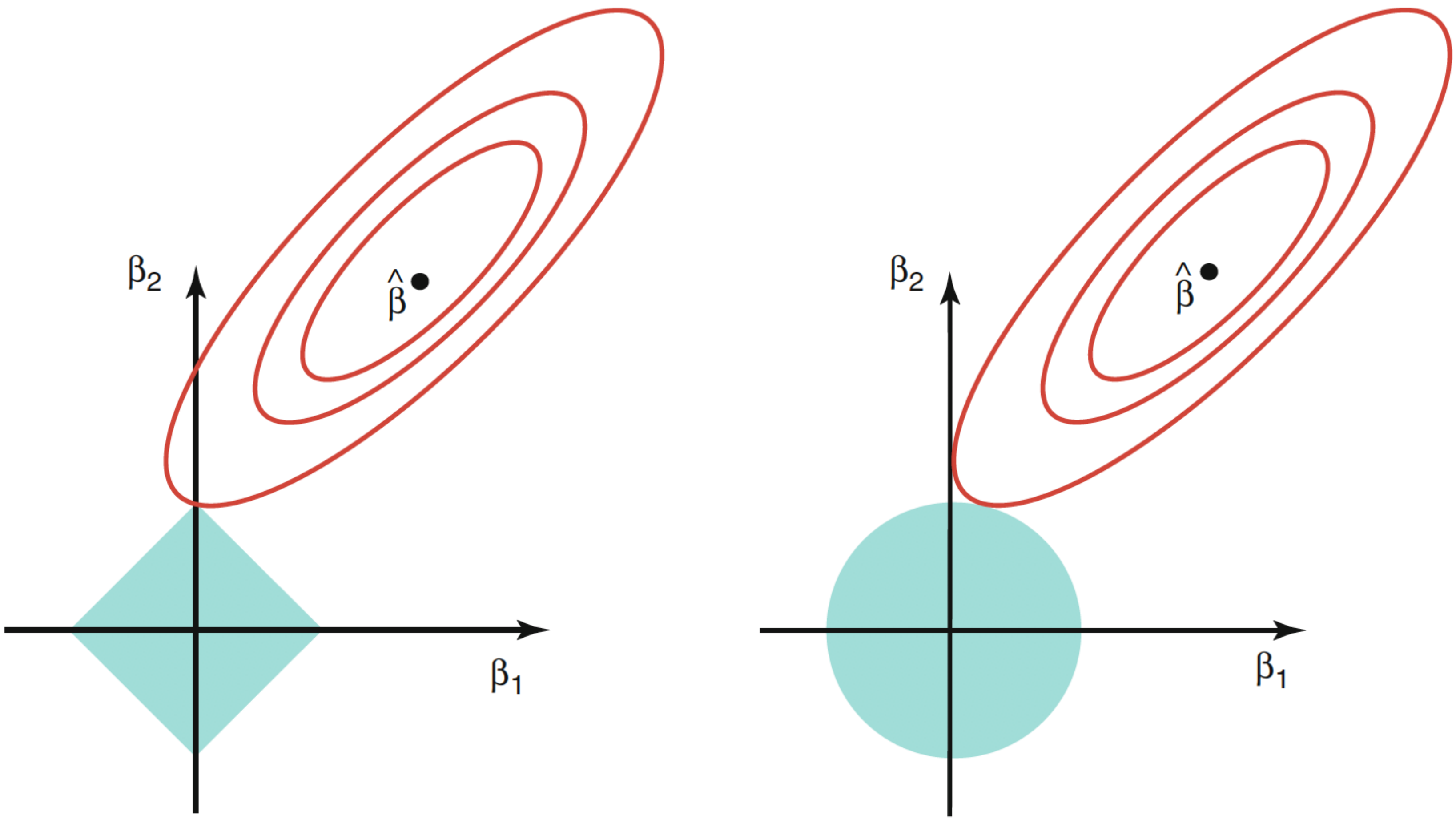
- Lasso

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

- Ridge

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

Ridge and Lasso (Geometric intuition)



Comparison between Ridge and Lasso

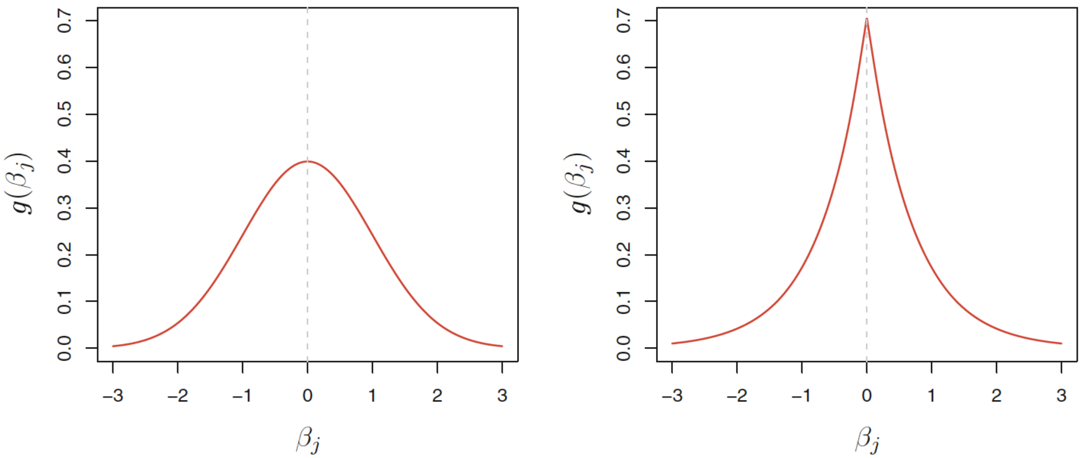
- Neither is universally better than the other
- One expects lasso to perform better for cases where a relatively small number of predictors have coefs that are very small or zero
- One expects ridge to be better when the response is a function of many predictors, all with roughly equal size
- Hard to know a priori, techniques such as CV required

Recommended exercise 6

1. Apply Lasso regression to the Credit Dataset.
2. Compare the results with the standard linear regression and the Ridge regression.

Bayesian interpretation

- Gaussian prior with zero mean and std. dev. as function of lambda
  - posterior mode is the ridge regression solution
- Laplace prior with zero mean and scale parameter as a function of lambda
  - posterior mode is the lasso solution



Selecting λ

- Pick λ for which the cross-validation error is smallest.
- re-fit using all of the available observations and the selected value of λ.

References

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software* 33 (1): 1.