STA 4241 Lecture, Week 7

Overview of what we will cover

- Extensions to the linear model
  - What happens if we have many covariates?
- Variable selection procedures
  - Best subset selection
  - Stepwise variable selection
- Shrinkage approaches
  - Ridge regression

The linear model

- We have seen previously the standard linear model

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

- Widely used and powerful approach
  - Easy to perform inference
  - Performs well in many real data problems
    - Linear approximation is a good one
- We will discuss extensions to this over the next few weeks

The linear model

- This model is generally fit with least squares, where we find the coefficients that minimize

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{X_i}\beta)^2 = (\boldsymbol{Y} - \boldsymbol{X}\beta)^T(\boldsymbol{Y} - \boldsymbol{X}\beta)$$

- This week (and next) we will focus on improving this model in a number of ways
- In particular, we will focus on cases where the number of covariates p is relatively large
- In this case, simply performing least squares may not be the best approach

The linear model

- If p is small, then least squares will perform quite well
  - Unbiased, low variance
- If p starts approaching n then least squares estimates start becoming unstable
  - Still low bias, but high variance
- If $p > n$, then there is no unique least squares solution and we must perform alternative procedures
- Alternative procedures might have better predictive performance in terms of MSE

The linear model

- Another reason least squares may not be optimal is model interpretability
- When p is large, many covariates may not be associated with the outcome
- We are interested in knowing which predictors are associated with the outcome
  - Variable selection / feature selection
- More interpretable models if they have fewer predictors

Overview of what we will cover

- Extensions to the linear model
  - What happens if we have many covariates?
- **Variable selection procedures**
  - Best subset selection
  - Stepwise variable selection
- Shrinkage approaches
  - Ridge regression

Model selection approaches

- Today we will cover a few different approaches to choosing the best model
  - Best subset selection
  - Forward stepwise regression
  - Backward stepwise regression
- These approaches are based on the idea that you first try to find the best model
- Then you estimate this best model via least squares

Model selection criteria

- Before we decide what is the best model, we need to define the criteria by which we judge models
- RSS $= \sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i\beta)^2$
- $R^2 = \text{cor}(Y, \widehat{Y})^2$
- Note that both of these measures are susceptible to overfitting
  - Appear good on training data, but don't translate to good testing data performance
  - RSS always decreases and R^2 always increases as we include more covariates
    - Whether they are important or not

Model selection criteria

- Other measures are better for assuring the model is not overfit
- One commonly used measure is the AIC

$$\text{AIC} = -2\log(\hat{L}) + 2p$$

- BIC is similarly defined as

$$\text{BIC} = -2\log(\hat{L}) + p\log n$$

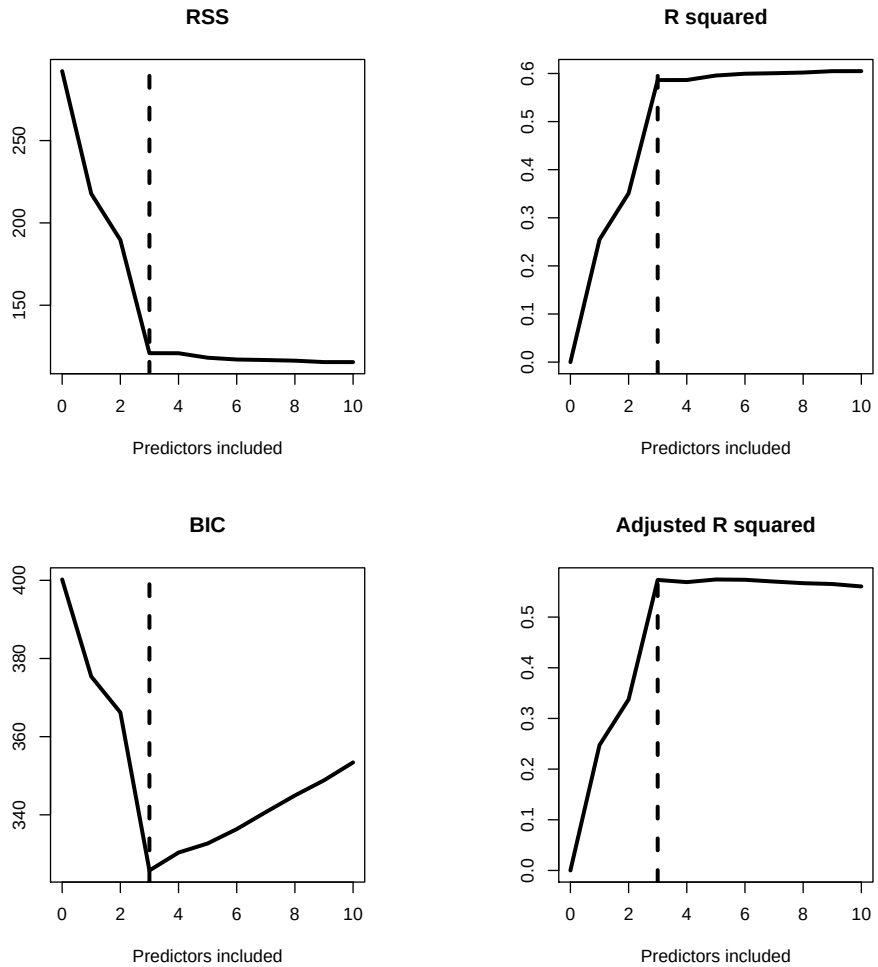
- $\log(\hat{L})$ is the log of the likelihood for the estimated model

Model selection criteria

- Adjusted R^2 can be used
  - Penalizes R^2 for having additional parameters
  - Not guaranteed to go up as more predictors are added
- Cross-validated error
  - Directly evaluates prediction performance on testing data

Model selection criteria

- Let's look at these measures as a function of variables included in a model
- True model has 3 predictors in it



Best subset selection

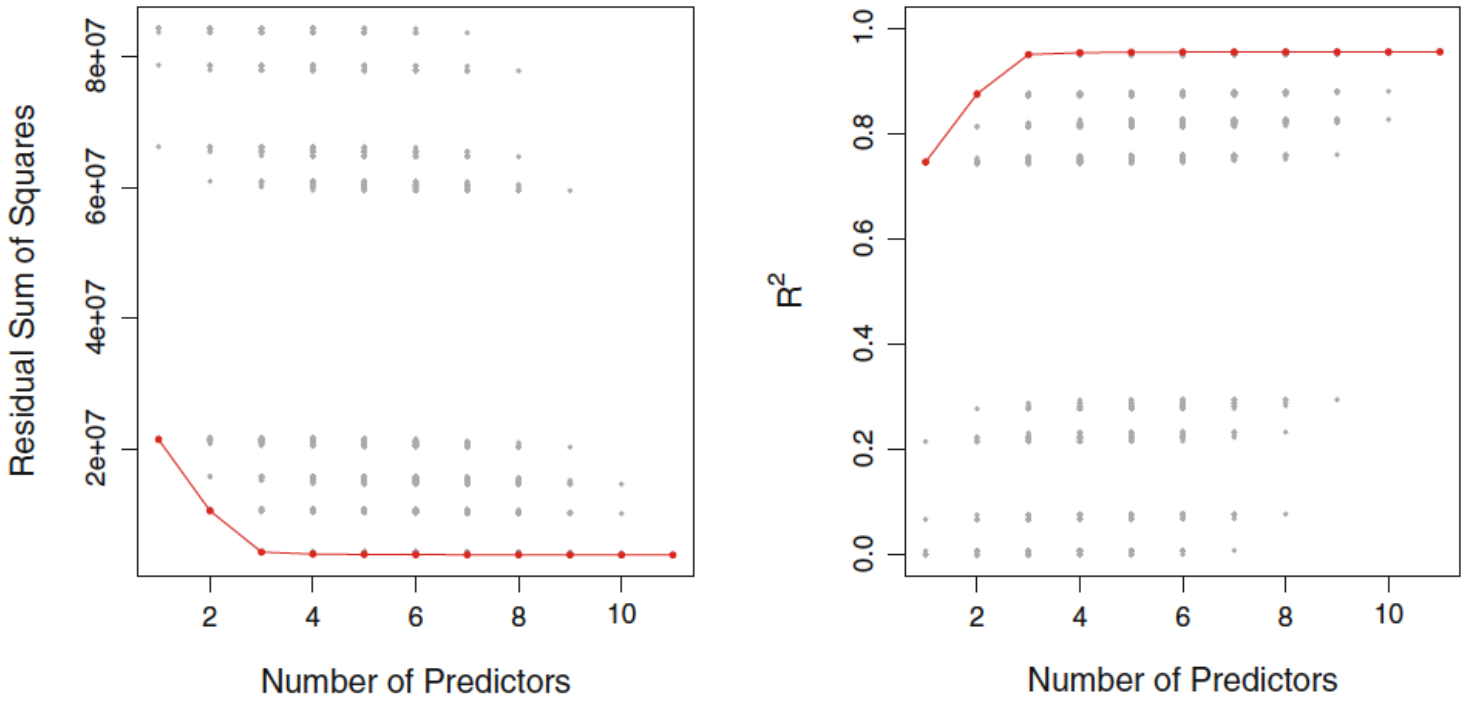
- The most obvious approach is just to compare all possible models!
- There are 2^p possible models
  - p covariates that can either be in or out
- We could apply any of these criteria to each model and find the best one

Best subset selection

- Let $\mathcal{M}_0$ be the model with no covariates
- The book provides an intuitive algorithm for doing subset selection
  1. For $k = 1, \ldots, p$:
     1. Fit all $\binom{p}{k}$ models with k predictors
     2. Pick the best model from this group and call it $\mathcal{M}_k$. Best here is defined as the model having smallest RSS or largest R^2
  2. Choose among the candidate models $\mathcal{M}_0$, $\mathcal{M}_1$, $\ldots$, $\mathcal{M}_p$ using cross validation or any other approach not susceptible to overfitting

Best subset selection

- An illustration of this on the Credit data set from the book shows this process



James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.

Best subset selection

- This approach will likely perform well for prediction
- The main drawback of this approach is computational
- The number of possible models, 2^p, gets out of hand pretty quickly
  - p=5: 32 models
  - p=10: 1024 models
  - p=20: 1048576 models
  - p=30: 1073741824 models
- It is infeasible to run this approach with even a moderate number of predictors
- If p is large, best subset selection might overfit
  - Just by chance one model will look good if you search so many

Stepwise regression

- An alternative to searching all models is to use a greedy search algorithm
- Intuitively, greedy algorithms are ones that separate the search into stages and they make optimal decisions at each stage, but are not guaranteed to be globally optimal
- The most common class of these for model selection are stepwise regression problems
  - Forward stepwise regression
  - Backward stepwise regression

Forward stepwise regression

- The simplest such algorithm is called forward stepwise regression
- Similar to the algorithm for best subset model selection, with a modification that reduces computation substantially
- For $k = 0, \ldots, p-1$:
  1. Consider the $p - k$ models that involve adding a single predictor to $\mathcal{M}_k$
  2. Pick the best model from this group and call it $\mathcal{M}_{k+1}$. Again best is defined as the model having smallest RSS or largest R^2
  3. Choose among the candidate models $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using cross validation or any other approach not susceptible to overfitting

Advantages of forward stepwise regression

- This algorithm has improved computation time substantially
- It turns out that the number of models considered with this approach is $1 + p(p+1)/2$
- Let's compare this to an approach that compares all models

	Best subset	Forward stepwise
$p = 5$	32	16
$p = 10$	1024	56
$p = 20$	1048576	211
$p = 30$	1073741824	466
$p = 100$	1267650600228229401496703205376	5051

Number of models to consider

Disadvantages of forward stepwise regression

- The computational gains are massive
- But at what cost? There's never a free lunch
- This approach will not necessarily find the best model out of the 2^p possible models
- Imagine a setting in which the best one variable model uses X_1, but the best two variable model uses (X_2, X_3)
  - Forward stepwise requires that X_1 be included in any two variable model
  - Won't find the best two variable model

Disadvantages of forward stepwise regression

- The book shows an example of this situation using the Credit data set
- Forward stepwise selection requires that rating be included in the best four variable model
  - Best subset selection does not

# 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`

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.

Backward stepwise regression

- An alternative to forward selection is backward stepwise selection
- Letting $\mathcal{M}_p$ be the full model that includes all covariates, the algorithm is as follows:

  For $k = p, p-1, \ldots, 1$:

  1. Evaluate all models that remove one predictor from $\mathcal{M}_k$
  2. Choose the best of these models using RSS or R^2 and call it $\mathcal{M}_{k-1}$

- Choose among the candidate models $\mathcal{M}_0$, $\mathcal{M}_1$, $\ldots$, $\mathcal{M}_p$ using cross validation or any other approach not susceptible to overfitting

Backward stepwise regression

- The pros/cons of backward stepwise selection are similar to those for forward stepwise
  - Fast computationally
  - Won't necessarily find the optimal model
- Backwards has an additional issue, which is that it can't be used if $p > n$
  - Can't fit the full model using least squares
- Forward stepwise regression can be used if $p > n$

Stepwise regression

- Some of the issues of forward/backward selection can be alleviated by considering both addition and subtraction of covariates
- A hybrid version follows a similar strategy as forward stepwise regression but allows for removal of covariates that have already been included if they are no longer deemed important
  - More computationally expensive than forward regression, but still less than evaluating all models
  - Might get closer to the global optimal solution than forward/backward stepwise regression

Treat models chosen by data carefully

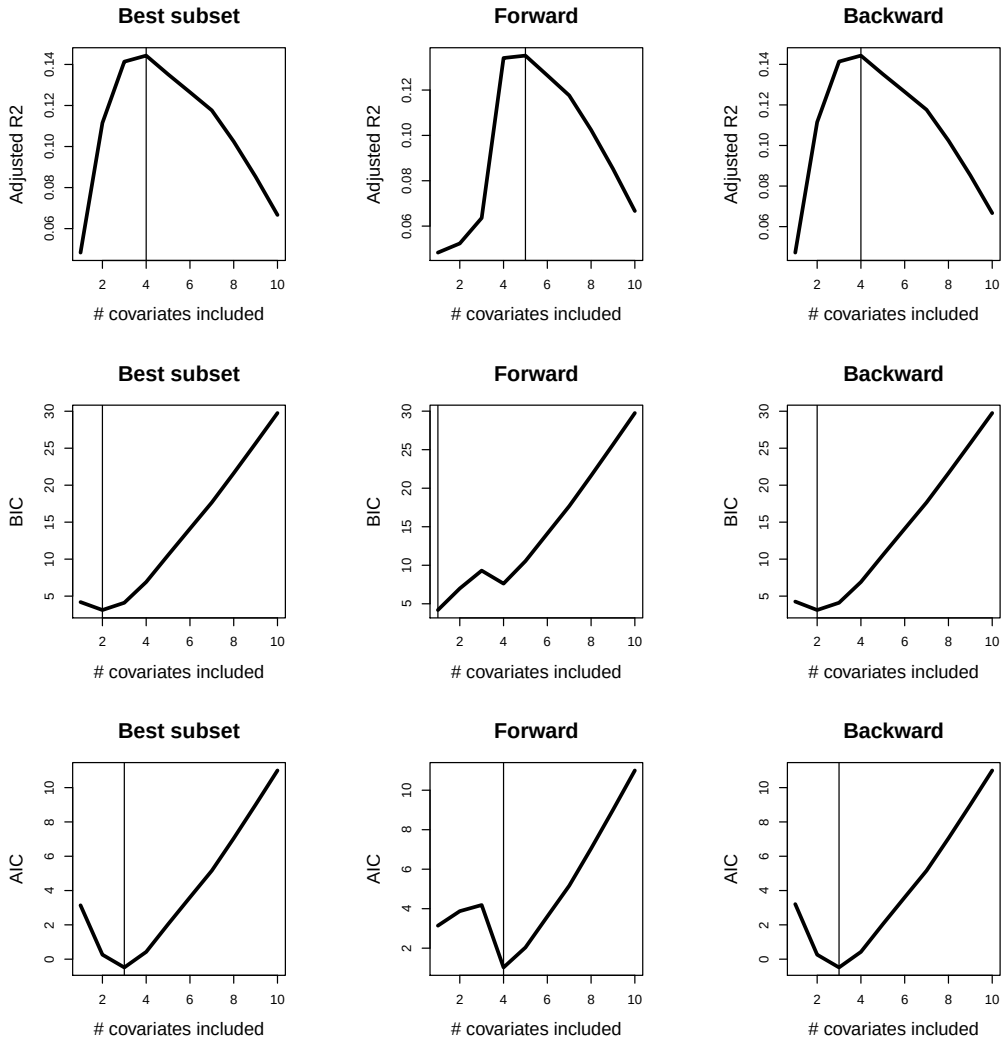
- Coefficients from models estimated after model selection tend to have certain issues
  - Ignore uncertainty stemming from model selection procedure
- Researchers have found ways to account for the model selection process when performing inference
  - Beyond the scope of this class
- The main takeaway is that you can't simply look at the output from your final chosen model and use those p-values like you normally would

Example on mice data

- Now let's compare each of these approaches to choosing a model
  - Best subset selection
  - Forward/backward stepwise
- We will use the mice data from the spls package in R
- Data contains 145 predictors and only 60 samples
- For now, we will focus on a subset of 10 predictors

Example on mice data

- Here are three different model criteria applied to the best models found by each approach



Example on mice data

- Substantial disagreement across the board!
- There was disagreement across model criteria
  - (AIC, BIC, etc.)
- Disagreement across the model searching tools
  - Forward, backward, best subset
- Not to mention disagreements across model searching tools on which model was best for a given number of covariates

Example on mice data

- How do we choose among all of these models if the approaches can't agree?
- If your goal is prediction, I recommend using cross-validation as your model criteria
  - Pick the model with the lowest CV
- If your interest is in finding a parsimonious model or finding the most important predictors, a conservative approach such as BIC could work
- There is no agreed upon way that everyone uses for model selection

Overview of what we will cover

- Extensions to the linear model
  - What happens if we have many covariates?
- Variable selection procedures
  - Best subset selection
  - Stepwise variable selection
- **Shrinkage approaches**
  - Ridge regression

Linear regression

- Before we get into shrinkage methods, let's review a couple things about the linear model and least squares
- The least squares estimate is given by

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

- It can be shown that this estimator has a very nice property in that it is the best estimator among all unbiased, linear estimates of $\boldsymbol{\beta}$

Gauss-Markov Theorem

- The Gauss-Markov Theorem says that any estimator $\tilde{\boldsymbol{\beta}}$, which is linear in $\boldsymbol{y}$ and unbiased for $\boldsymbol{\beta}$ has larger variance than $\widehat{\boldsymbol{\beta}}$, i.e.

$$\text{Var}(\tilde{\boldsymbol{\beta}}) \succeq \text{Var}(\widehat{\boldsymbol{\beta}})$$

  where $A \succeq B$ simply means that $A - B$ is positive definite

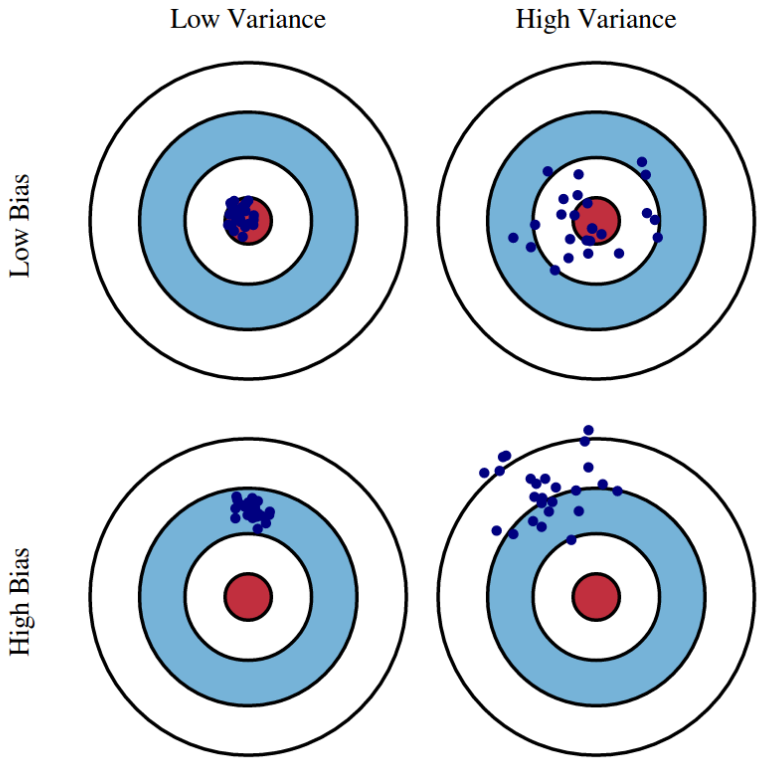
- This means that the least squares estimator is the best linear, unbiased estimator (BLUE) of $\boldsymbol{\beta}$
  - Smallest variance (and therefore MSE) among *unbiased* estimators

Linear regression

- This is great, but are we being overly restrictive by only considering unbiased estimators?
- As we have seen throughout class, there is typically a bias-variance trade-off that exists in many of the approaches we use
- Maybe we can induce some bias in the linear model estimates, but improve the MSE
  - Only incur small amount of bias
  - Big reductions in variance

Bias-variance Trade-off

- The least squares estimate could live in the top-right panel of this plot
- Might be better to increase the bias just a little bit



Taken from the blog of Scott Fortmann-Roe

Shrinkage methods

- Shrinkage methods are a general class of methods that "shrink" the least squares estimates towards zero
  - Regularize, constrain, or penalize the coefficients
- We will learn about two very popular approaches to regularization
  - Ridge regression
  - Lasso regression
- Both of these approaches bias coefficients towards zero in distinct ways

Shrinkage methods

- Why does this work?
- Why would forcing coefficients towards zero improve prediction error?
- It turns out that this shrinkage can be done in a way that ends up with desirable bias-variance trade-offs
- All of these approaches will bias estimates, but also reduce their variance, and the key is finding the optimal balance between these two competing issues

Penalized approaches to regression

- Remember that least squares finds the value of $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2$$

- Penalized least squares estimators find the value that minimizes

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2 + P_\lambda(\boldsymbol{\beta})$$

where $P_\lambda(\boldsymbol{\beta})$ is called a penalty term

Penalized approaches to regression

- Penalty terms are positive functions that are larger for values of $\boldsymbol{\beta}$ that are farther from zero
- Many different penalties one could use
  - $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \beta_j^2$
  - $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|$
  - $P_\lambda(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|$
- We will look closely at the first of these two, but many choices exist
  - Each penalty implies different structure on the estimated coefficients

Penalized approaches to regression

- How does the penalty work?
- We want values of β that fit the data well, which makes

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i\beta)^2$$

  smaller. The penalty term, however, favors small values of β that may not fit the data well
- Solution to the penalized least squares function is the value of β that balances competing interests of the RSS and the penalty term

Ridge regression

- Ridge regression is the value of $\boldsymbol{\beta}$ that minimizes

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

- Refer to this solution as $\widehat{\boldsymbol{\beta}}_\lambda^R$
- Note that this solution is indexed by λ
  - Unique solution for each λ value
  - λ controls the bias-variance trade-off

Importance of λ

- The performance of ridge regression is heavily dependent on λ
  - Will be chosen via cross-validation
- Also note that we do not penalize the intercept term, β_0
- The intercept simply measures the mean of the outcome when all covariates are zero
- We only penalize the effect of each covariate on the response

Ridge regression

- From now on, we will assume that both the outcome and covariates have been shifted to be mean zero
  - This doesn't change the true value of $\beta_1, \ldots, \beta_p$
  - Forces that $\beta_0 = 0$
- So now $\boldsymbol{X}$ is an $n \times p$ matrix of predictors (with no intercept) and $\boldsymbol{\beta}$ is a vector of length p representing $\beta_1, \ldots, \beta_p$
- Let's investigate the ridge regression solution further

Ridge regression

- We can show that $\widehat{\boldsymbol{\beta}}_{\lambda}^{R} = (\boldsymbol{X}^{T}\boldsymbol{X} + \lambda \boldsymbol{I}_{p})^{-1}\boldsymbol{X}^{T}\boldsymbol{Y}$

Ridge regression

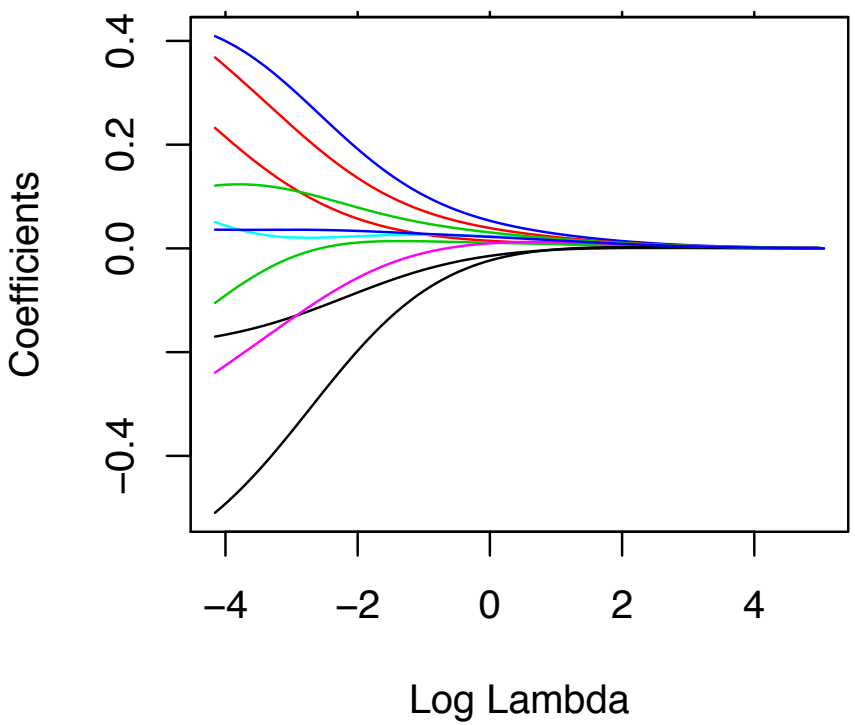
- Now let's show that ridge regression is biased

Ridge regression

- This closed form for the estimate is useful because it provides us with some intuition about λ
- If $\lambda = 0$, then $\widehat{\boldsymbol{\beta}}_\lambda^R = \widehat{\boldsymbol{\beta}}^{ols}$
- As $\lambda \to \infty$, then $(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \approx \lambda^{-1}\boldsymbol{I}_p$ and the ridge solution converges to $\widehat{\boldsymbol{\beta}}_\lambda^R = \mathbf{0}_p$
- Larger values of λ lead to more bias, but smaller variance

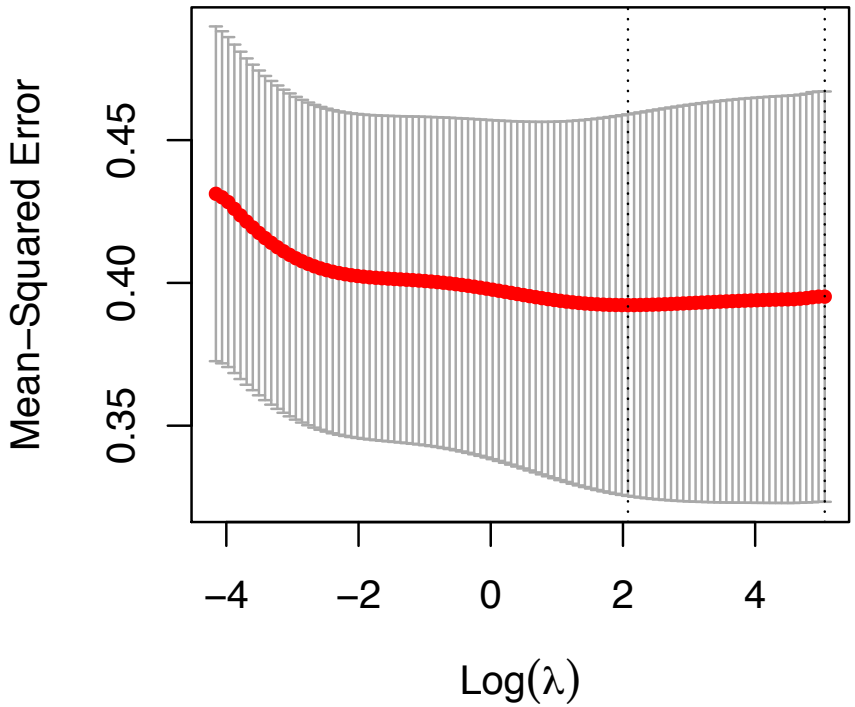
Mice example

- Returning to the mice data, below are the estimates of $\beta_1, \ldots, \beta_{10}$ as a function of λ
- As we increase λ, they all are shrunk towards zero
- Results are quite sensitive to the choice of λ it appears



Mice example

- let's run cross-validation to see which λ is best
- Below is the CV curve for λ
  - Bars denote $\pm$ standard error of CV estimate



Ridge regression

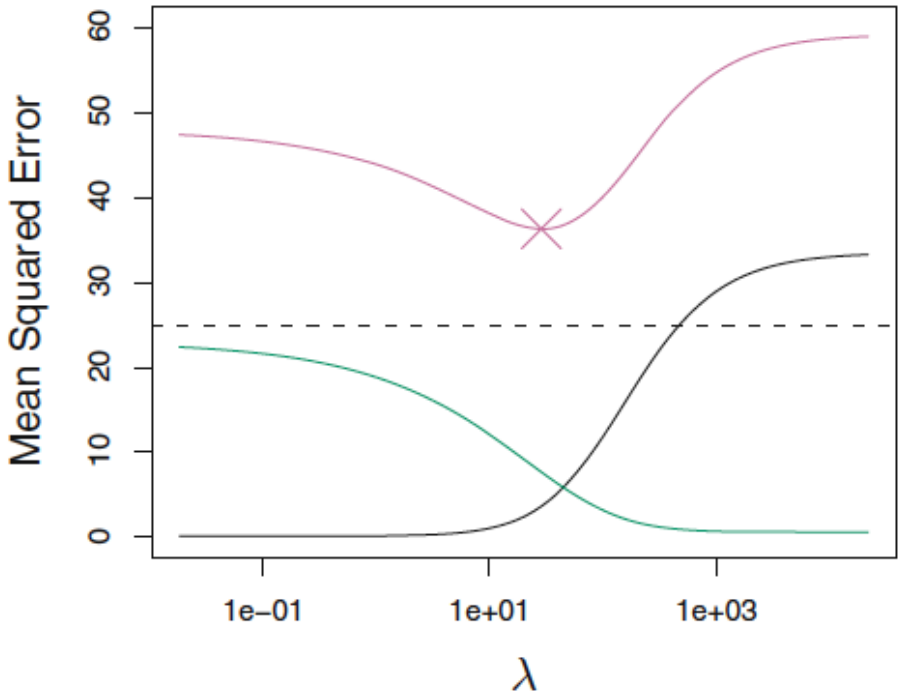
- A really interesting property of ridge regression is that there always exists a $\lambda > 0$ such that

$$\text{MSE}(\widehat{\boldsymbol{\beta}}_\lambda^R) < \text{MSE}(\widehat{\boldsymbol{\beta}}^{ols})$$

- We can always beat least squares in terms of MSE!
- Importantly this relies on the correct choice of λ

Ridge regression

- The book highlights this with a simulated example
  - Purple line is MSE, variance is green, and squared bias is black
- The optimal λ has lower MSE than $\lambda = 0$, which corresponds to least squares



James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.

Standardizing covariates

- One important feature of ridge regression is that your covariates should be standardized
- Each coefficient is penalized the same amount
- If your predictors are on very different scales, this can lead to bad performance
- General rule of thumb is to standardize predictors to have mean zero and variance one

High-dimensional data

- High-dimensional settings are when $p > n$
- Ridge regression, and other penalized estimators, become increasingly important in high-dimensional settings
- In high dimensions, it turns out that there are an infinite set of solutions that minimize

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

- The least squares solution doesn't exist

High-dimensional data

- $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ does not exist when $p > n$
  - Matrix is singular, i.e non-invertible
- Ridge regression solves this problem by adding a positive value to the diagonal $(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_p)^{-1}$
  - Makes the solution unique
  - Stabilizes the coefficients and reduces their variance
- Ridge regression is still useful when $p < n$
  - Can improve MSE of predictions
  - Improves estimates of $\boldsymbol{\beta}$ when the covariates are highly correlated