STAT 480 Statistical Computing Applications

Unit 6. Model Selection

Lecture 2. Ridge Regression

Department of Statistics, Iowa State University

Spring 2019

Shortcomings of linear regression

- Prediction Accuracy
  - Recall that we can decompose prediction error into squared bias and variance.
  - Linear regression has low bias (zero bias) but suffers from high variance.
  - So it may be worth sacrificing some bias to achieve a lower variance.
- Interpretation
  - With a large number of predictors, it can be helpful to identify a smaller subset of important variables.
  - Linear regression doesn't do this.
- Linear regression is not defined when $p > n$.

Ridge regression

- Ridge regression is like least squares but shrinks the estimated coefficients towards zero.
- Given data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, the ridge regression coefficients $\hat{\beta}^{\text{ridge}}$ is the value that minimize:

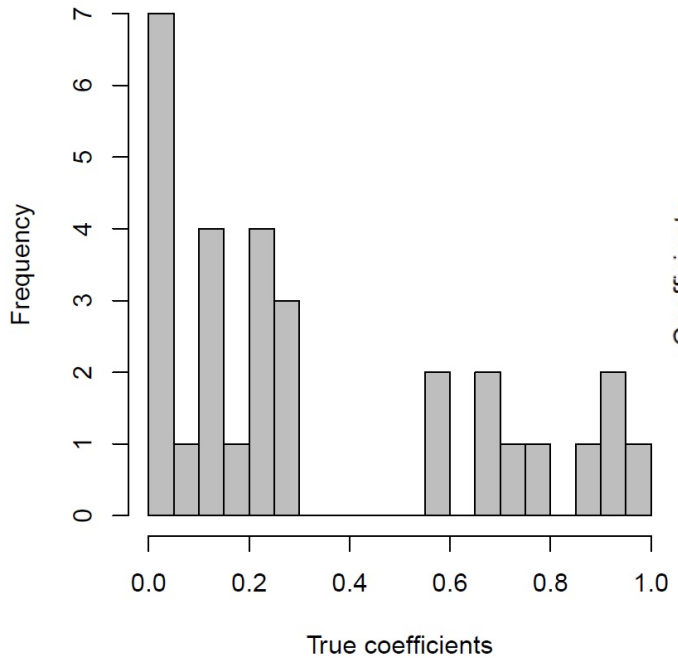
$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

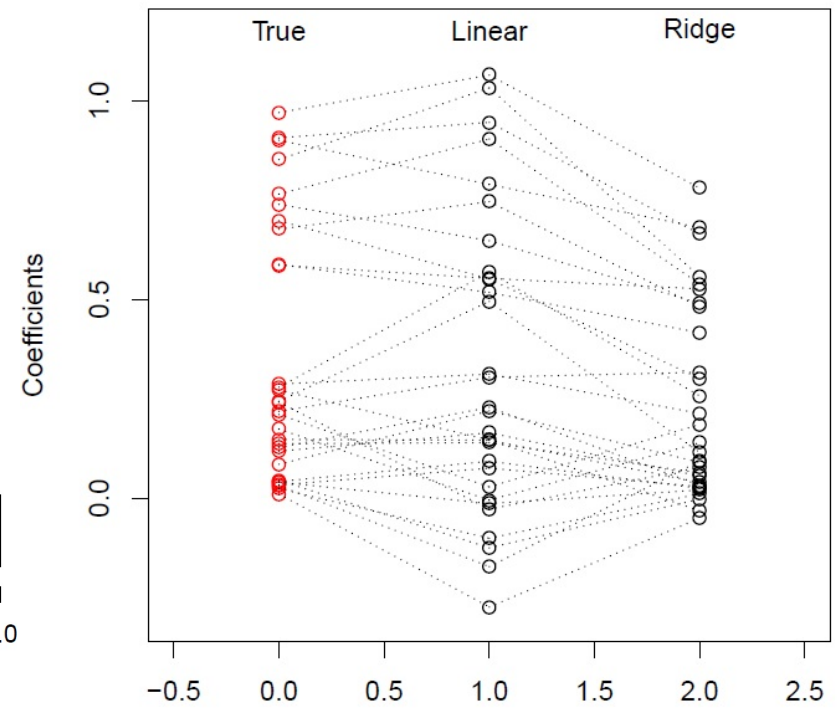
  Here λ is a tuning parameter, which controls the strength of the penalty term.

  - when $\lambda = 0$, we get the linear regression estimate;
  - when $\lambda = \infty$, we get $\hat{\beta}^{\text{ridge}} = 0$;
  - For λ in between, we are balancing two ideas: fitting a linear model of y on X, and shrinking the coefficients.

Visual Representation of Ridge Coefficients

Example 1: Simulation with $n = 50$ and $p = 30$. The entries of the predictor matrix X are all i.i.d. $N(0, 1)$. Here 10 coef's are large (between 0.5 and 1), 20 coef's are small (between 0 and 0.3)





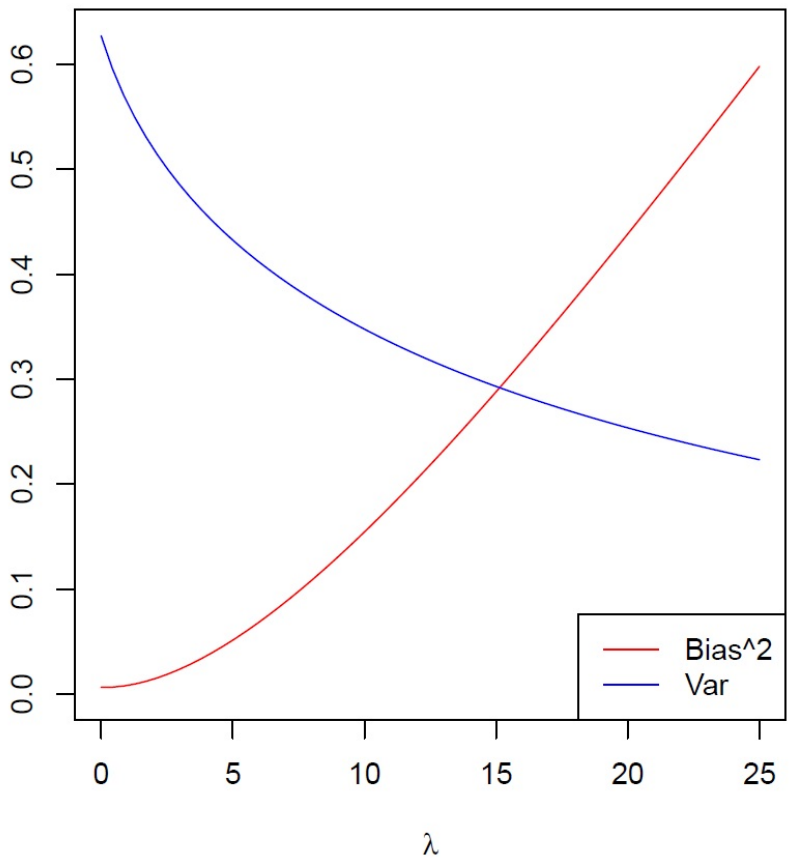
Some Important Remarks

- Note that when including an intercept term in the regression, we usually leave this coefficient unpenalized.
- If we center the columns of X, then the intercept estimate ends up being $\hat{\beta}_0 = \bar{y}$, so we usually just assume that y, X have been centered and don't include an intercept.
- Also, the penalty term $\sum_{j=1}^{p} \beta_j^2$ is unfair if the predictor variables are not on the same scale. (Why?) Therefore, if we know that the variables are not measured in the same units, we typically scale the columns of X (to have sample variance 1), then we perform ridge regression.

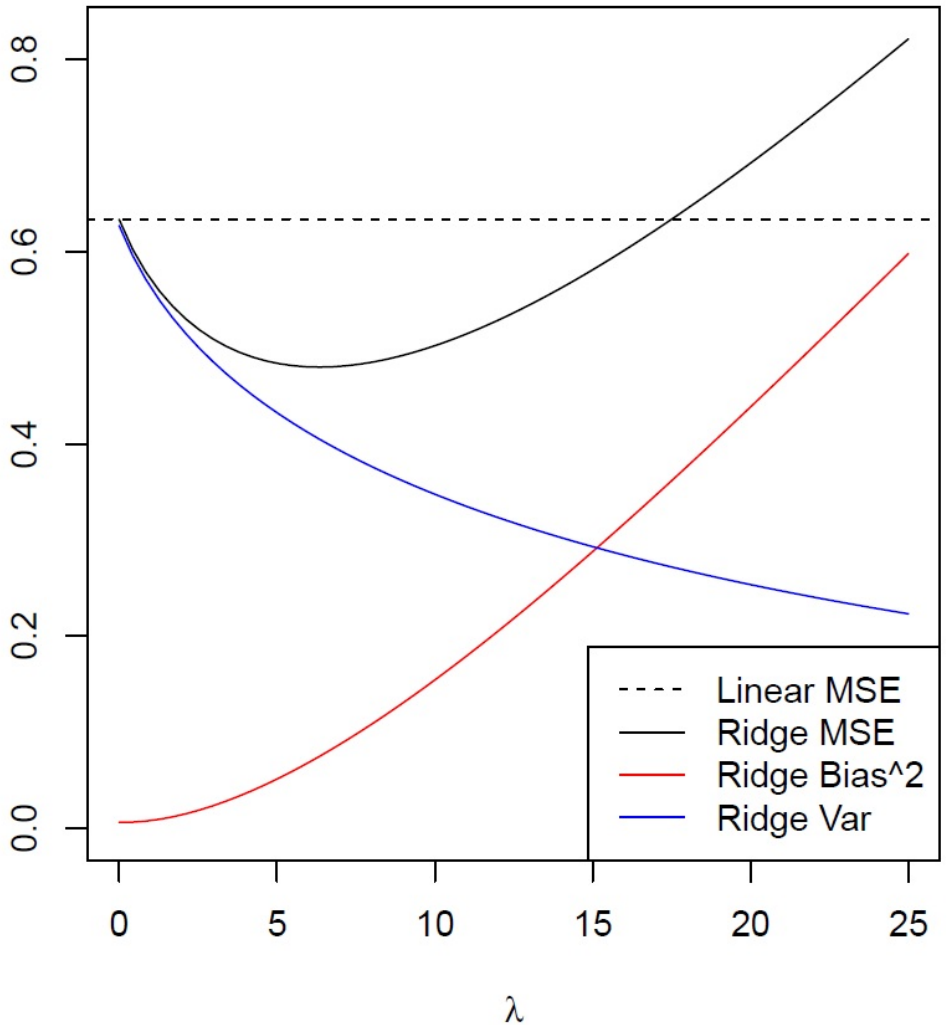
Bias and Variance of Ridge Regression

- The bias increases as λ (amount of shrinkage) increases
- The variance decreases as λ (amount of shrinkage) increases

Bias and variance for our last example ($n = 50$, $p = 30$; 10 large true coef's, 20 small):



MSE for our last example



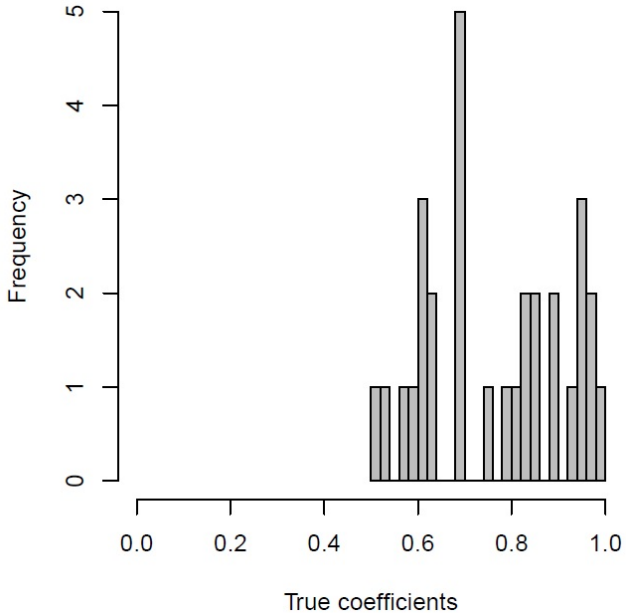
What you may (should) be thinking now

- **Thought 1:** "Yeah, OK, but this only works for some values of λ. So how would we choose λ in practice?"
  This is actually quite a hard question. We'll talk about this in detail later.
- **Thought 2:** "What happens when none of the coefficients are small?" In other words, if all the true coef's are moderately large, is it still helpful to shrink the coef's estimates?

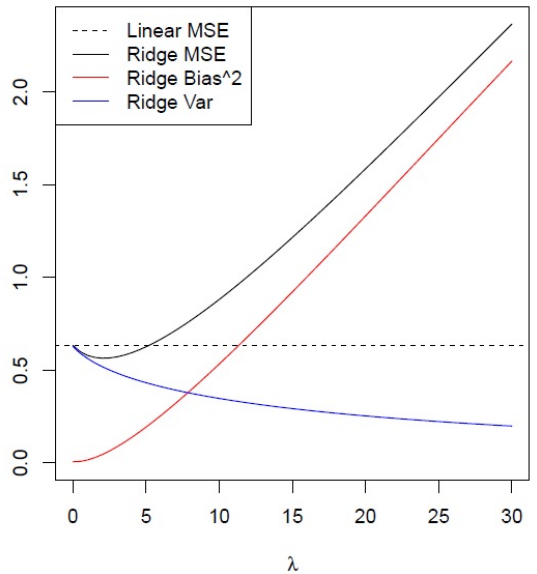
  The answer is (perhaps surprisingly) still "yes". But the advantage of ridge regression here is less dramatic, and the corresponding range for good values of λ is smaller.

Example 2: Moderate Regression Coefficients

Example 2: $n = 50$, $p = 30$. The true coef's are all moderately large (between 0.5 and 1).



(a) Histogram of true coefficients



(b) Mean-squared error

- Ridge regression can still outperform linear regression.
- Only works for $\lambda \approx 5$, otherwise it is very biased.

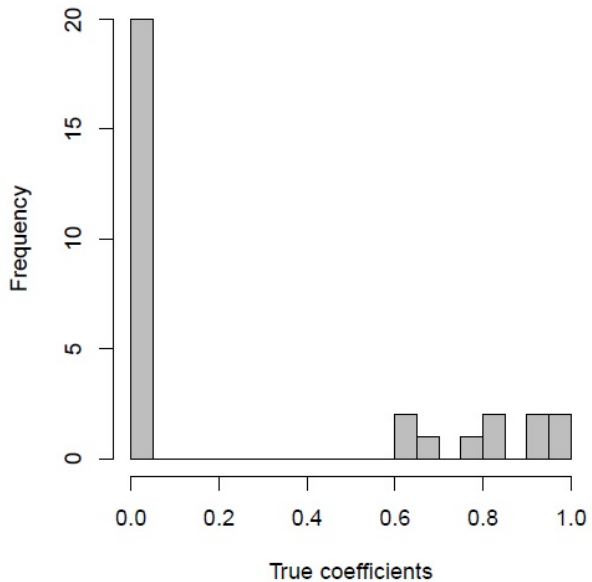
Variable selection

To the other extreme (of a subset of small coef's), suppose that there is a group of true coef's that are identically zero. This means that the mean response doesn't depend on these predictors at all; they are completely extraneous.

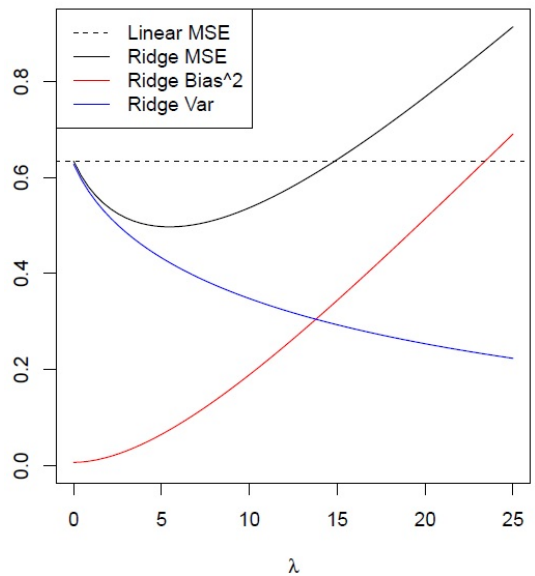
- **Thought 3:** "How does ridge regression perform if a group of the true coef's was exactly zero?"
  The answer depends whether on we are interested in prediction or interpretation. We'll consider the former first.

Example 3: Subset of Zero Coefficients

Example 3: $n = 50$, $p = 30$. Now, the true coef's: 10 are large (between 0.5 and 1) and 20 are exactly 0.



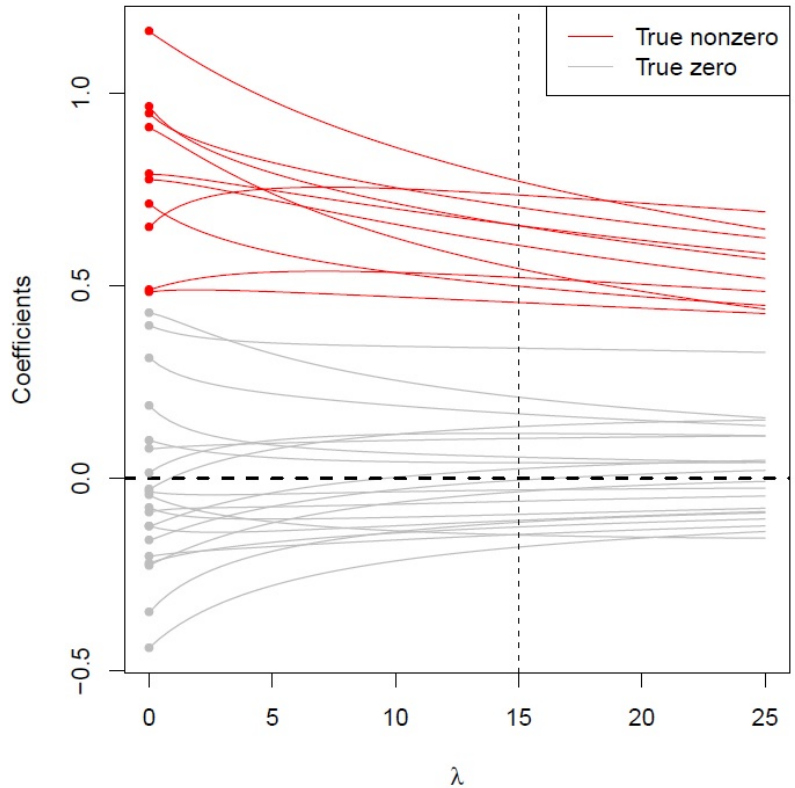
(a) Histogram of true coefficients



(b) Mean-squared error

- Ridge regression performs well in terms of mean-squared error.
- Why is the bias not as large here for large λ?

Remember that as we vary λ we get different ridge regression coef's, the larger the λ the more shrunken.



- Red paths: true nonzero coef's; gray paths: true zeros; vertical dashed line at $\lambda = 15$: the point above which ridge regression's MSE starts losing to that of linear regression.
- Note that the gray coef's paths are not exactly zero; they are shrunken, but still nonzero.

Ridge Regression Doesn't Perform Variable Selection

- We can show that ridge regression doesn't set coef's exactly to zero unless $\lambda = 1$, in which case they're all zero.
- Hence ridge regression cannot perform variable selection, and even though it performs well in terms of prediction accuracy, it does poorly in terms of offering a clear interpretation.

R code

- We will use the `glmnet` package in order to perform **ridge regression** and the **lasso** (will be introduced in the next lecture). The main function in this package is `glmnet()`.
- The `glmnet()` function has an `alpha` argument that determines what type of model is fit.
  - If `alpha=0`, then a ridge regression model is fit
  - if `alpha=1`, then a lasso model is fit (we will talk about the lasso model later).

R code: Details

- `x`: input matrix
- `y`: response vector
- `weights`: weight vector
- `alpha`: the elasticnet mixing parameter
- `nlambda`: the number of lambda values
- `lambda`: a user supplied lambda sequence
- `standardize`: logical flag for `x` variable standardization, default is `standardize=TRUE`
- `intercept`: should intercept(s) be fitted (`default=TRUE`)