Regularization and Regression

Ryan Miller

Grinnell College
Statistics

Introduction

Consider a basic linear regression model:

$$Y = w_0 + w_1 X_1 + w_2 X_2 + \ldots + w_p X_p + \epsilon$$

We'll focus more on these details later, but we typically find the optimal values of the model's weight parameters, $\{w_0, w_1, \ldots, w_p\}$, by minimizing a **cost function** that expresses the model's errors as a function of these parameters.

$$\text{Cost} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^T (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})$$

For quantitative outcomes, regression usually will use the *squared error* cost function (stated above).

Introduction (cont.)

- We've already seen how to *reduce the bias* of a linear model using feature expansion (splines, discretization, etc.)
  - But what if we'd like to *increase the bias* of our model to prevent overfitting?

Introduction (cont.)

- We've already seen how to *reduce the bias* of a linear model using feature expansion (splines, discretization, etc.)
  - But what if we'd like to *increase the bias* of our model to prevent overfitting?
- Regularized regression adds a penalty term to the cost function that shrinks weight estimates towards zero:

$$\text{Cost} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) + P_\alpha(\hat{\mathbf{w}})$$

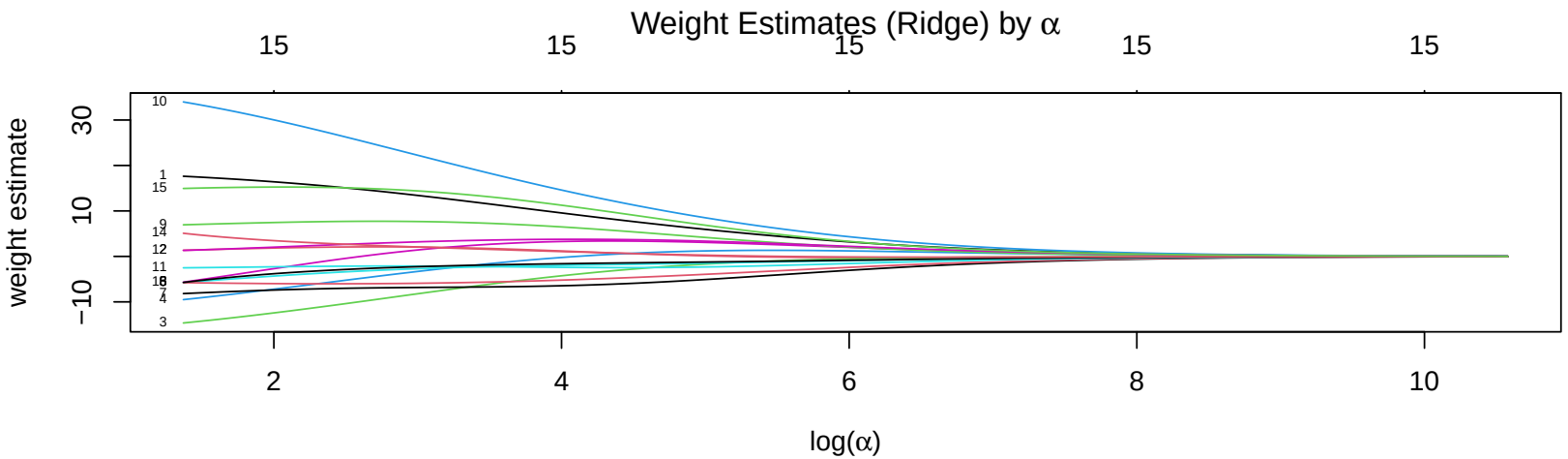
- $P()$ is a *penalty function* involving α, a **regularization parameter** that controls the trade-off between each term in the cost function

Types of Regularization

- **Ridge regression** uses the penalty function: $P_\alpha(\mathbf{w}) = \alpha \sum_{j=1}^{p} w_j^2$
  - This penalty function is commonly described as **L2 regularization**, as the penalty is applied to the L2 norm of the weight vector
- **Lasso regression** (least absolute shrinkage and selection operator) uses the penalty function: $P_\alpha(\mathbf{w}) = \alpha \sum_{j=1}^{p} |w_j|$
  - This is known as **L1 regularization**, as the penalty is applied to the L1 norm of the weight vector
  - L1 regularization promotes *sparsity*, a topic we'll explore in a bit
- Regardless of the type of regularization, input features should be *standardized* (re-scaled) to ensure the penalty is applied equally to all variables

Ridge Regression Example

We can plot the optimal weights (those that minimize the cost function) at every value of α:



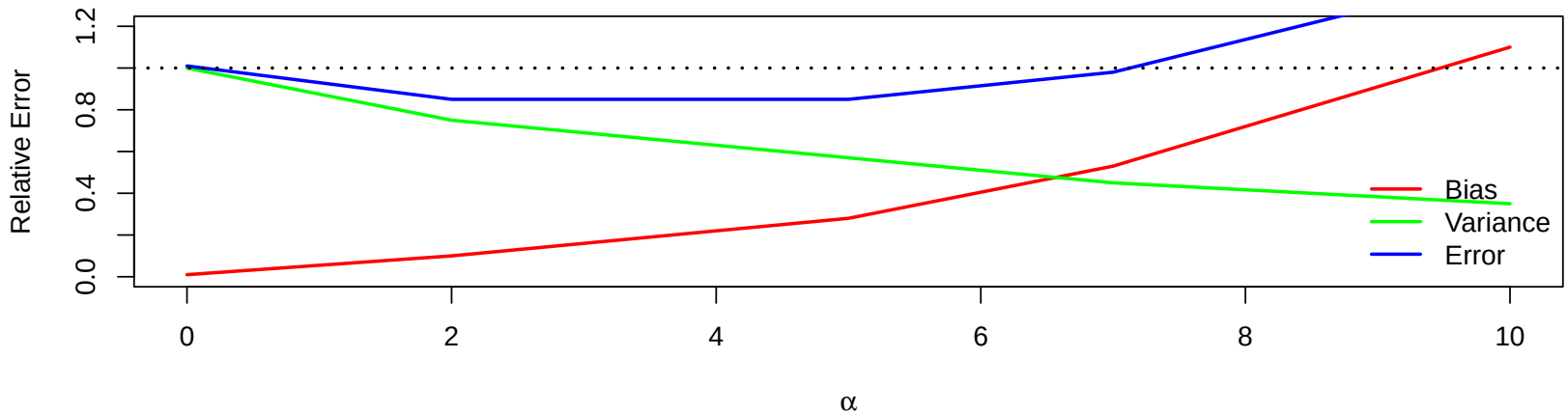
- When α, is large, the penalty term dominates the cost function and weights are estimated to be zero
- When α is zero, we have the ordinary least squares estimates

Benefits of Regularization

- The guiding philosophy of regularization is that when there are a large number of predictive features small weights, and small weights should be more likely than large weights
  - Thus, by encouraging smaller weight estimates, regularization should yield models that generalize better to new data (ie: lower out of sample error)
- In 1970, Hoerl and Kennard mathematically proved that regularized regression can always produce a lower *RMSE* than ordinary least squares regression

Benefits of Regularization (cont.)

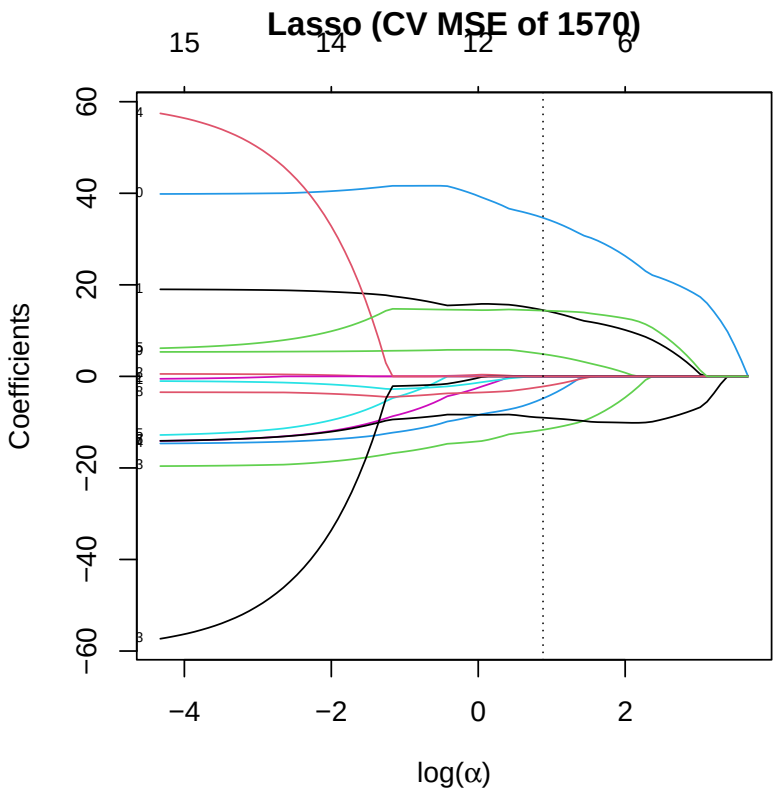
Mathematically, it's possible to decompose mean-squared error (MSE) into bias and variance terms. Here's a heuristic look at how these components might look as α is varied:

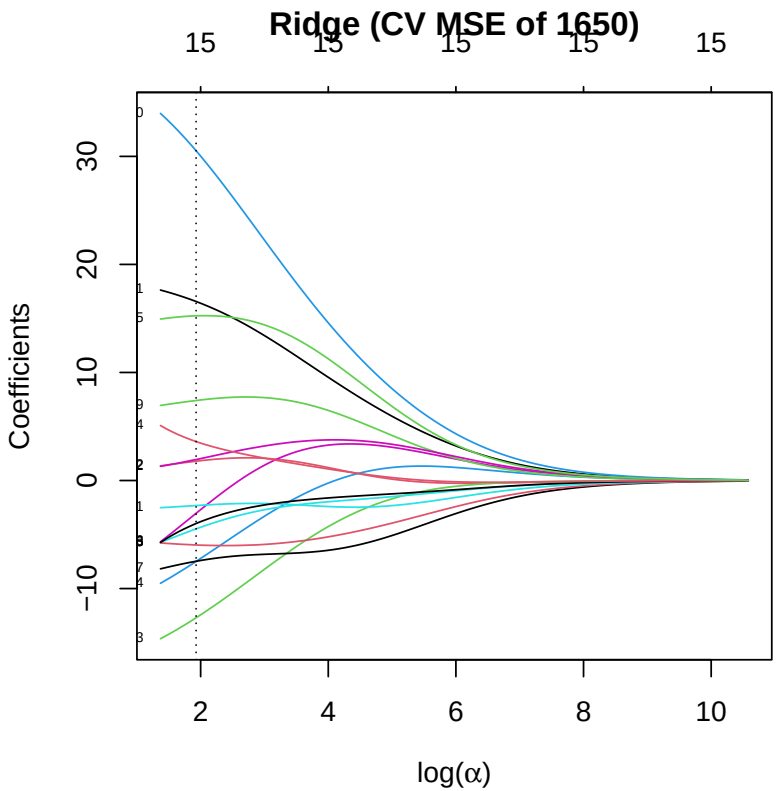


Essentially, variance will always decrease faster than bias increases (as α increases), thereby allowing a relative error rate < 1 for some α

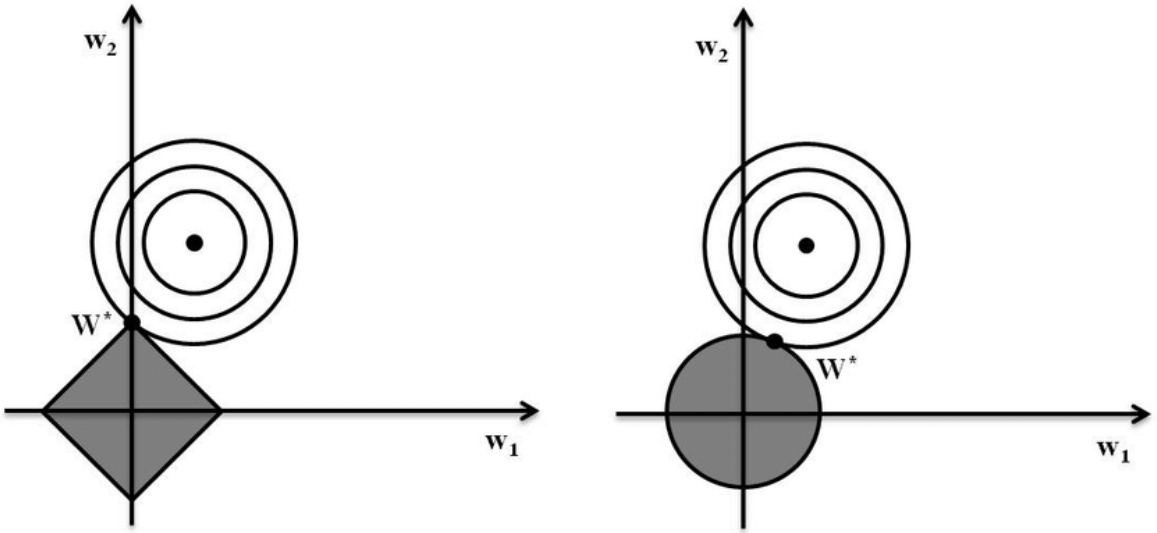
L1 vs. L2 Regularization

L1 Regularization encourages weight estimates of *exactly zero* (sparsity):





L1 vs. L2 Regularization (cont.)

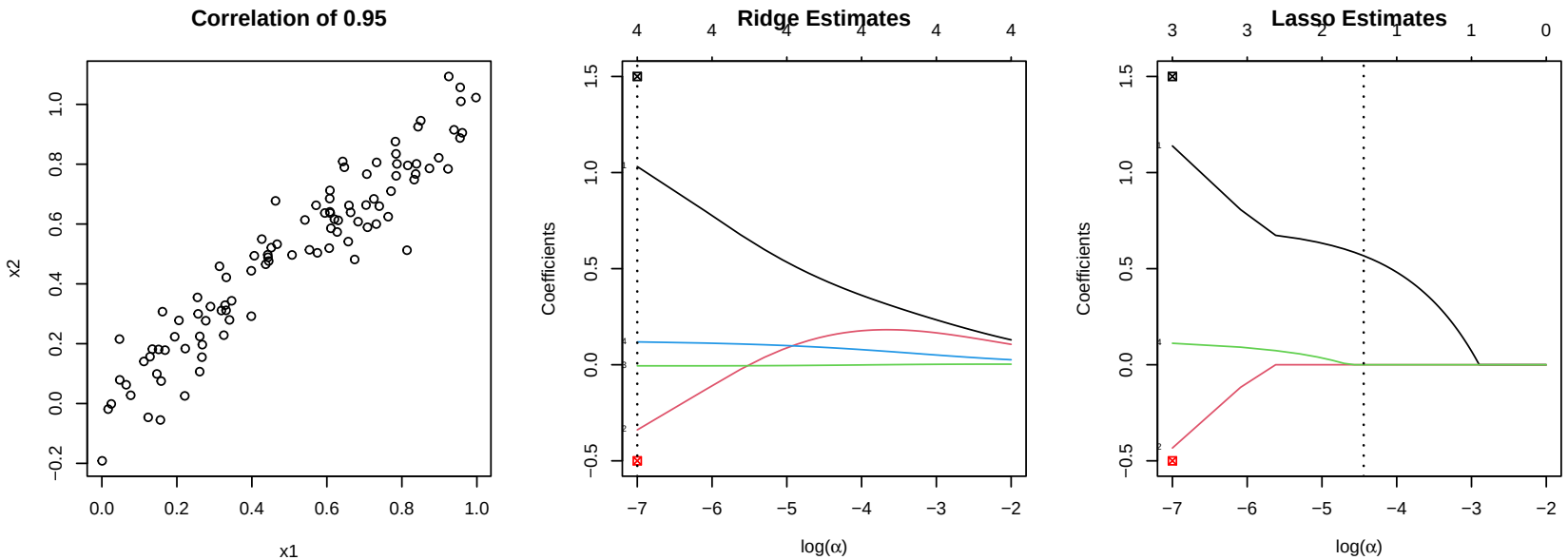


In two dimensions, weight estimates satisfying $\sum_{j=1}^{p} |w_j| < c$ exist within a diamond, while those satisfying $\sum_{j=1}^{p} w_j^2 < c$ exist within an ellipse. The former is likely to intersect contours of the squared error cost function at a corner (a weight estimate of exactly zero).

image credit: https://www.researchgate.net/figure/Plot-demonstrating-the-Sparsity-caused-by-the-LASSO-Penalty-The-plot-shows-the_fig1_317357840

L1 vs. L2 Regularization (cont.)

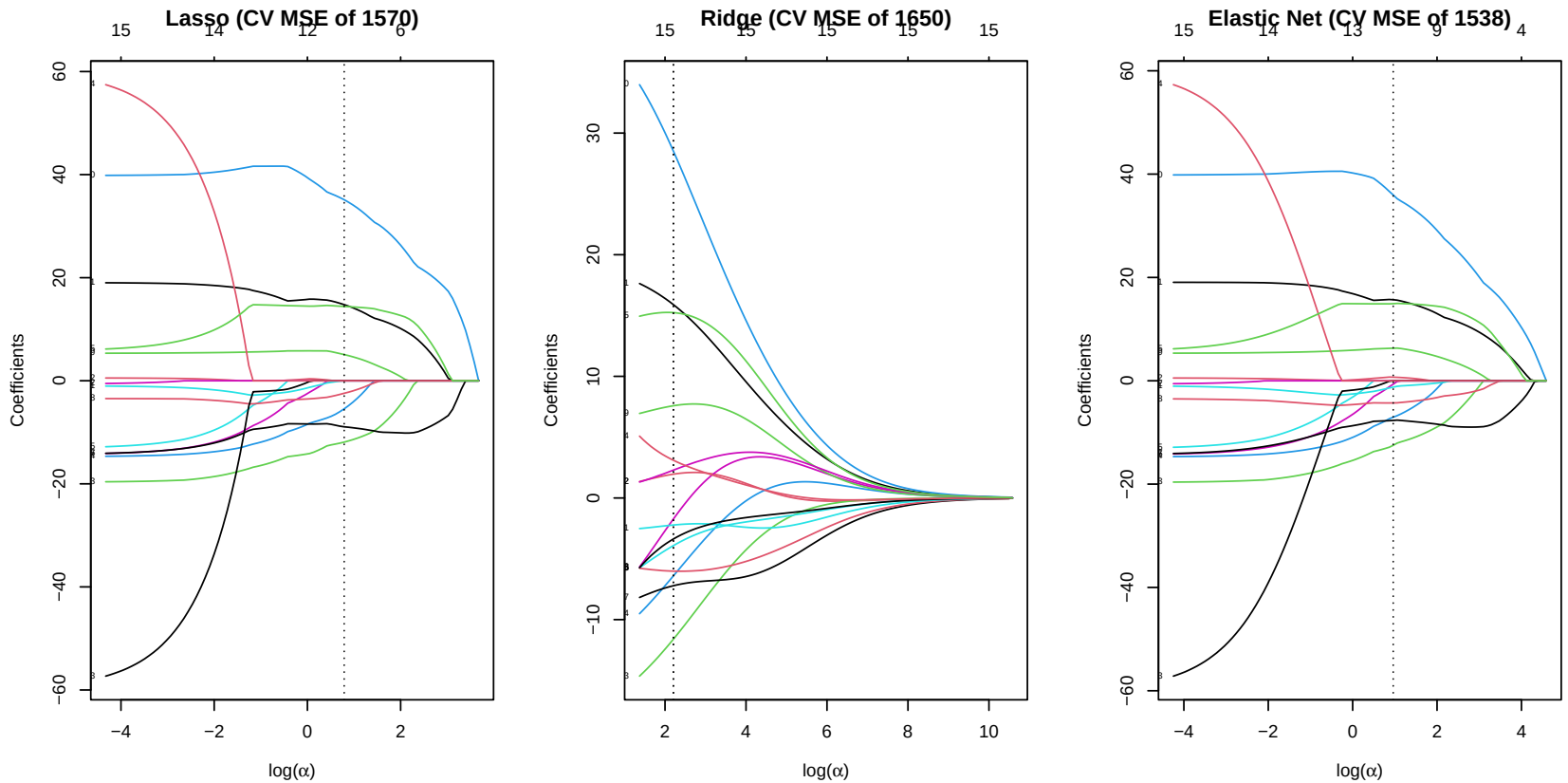
In the presence of *multicollinearity*, lasso favors a single representative, while ridge will distribute importance across weights:



Ridge regression's behavior can add stability (decrease variance) without increasing bias, making L2 regularization desirable when there are many correlated features.

Elastic Net

Elastic net models combine L1 and L2 penalties, seeking to get the benefits of sparsity (L1) and handling of multicollinearity (L2):



What's in a Name?

You don't need to know this, but understanding how methods get their names is good historical information:

- Ridge regression's name is from its closed form solution, which is similar to ordinary least squares, but with a "ridge" added to the covariance matrix
  - $\hat{\mathbf{w}}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ vs. $\hat{\mathbf{w}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- Lasso is an acronym: "least absolute shrinkage and selection operator"
- Elastic net was named by its creators (Zou and Hastie 2005), who say:
  . . . the elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. It is like a stretchable fishing net that retains 'all the big fish'.

What to Know for the Next Quiz?

- The basic idea of regularization as a strategy to reduce the variance (increase the bias) of a model by using a penalty function to shrink weight estimates towards zero
- Be familiar with regularization path plots of weight estimates in response to α for L1 and L2 penalties
  - As α increases, weight estimates approach zero for all features, and as α approaches zero weight estimates approach their ordinary least squares solution
  - L1 regularization encourages exact zeros on the path plot
  - L1 regularization selects representative features from blocks of correlated variables, while L2 regularization distributes importance across all features in the block
- Know that standardization/scaling is an essential step when using regularized regression