Penalized Regression Methods

Cody Chiuzan, PhD

Department of Biostatistics
Columbia University
June 6, 2019

Remember Linear Regression?

- Linear regression aims to predict a response variable (Y) with a linear combination of predictor variables (X) and a normally distributed error term with variance σ^2:

- MLR model is given below:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_p X_{ip} + \varepsilon_i, i = 1, 2, \ldots, n.$$

- Assumptions:
  - Uncorrelated error terms with mean 0 and constant variance, $\varepsilon_i \sim N(0, \sigma^2)$
  - Linearity in parameters

- How to we find the model estimates?
  - Least Squares and Maximum Likelihood Methods

Remember Linear Regression?

- The goal is to find estimates for the true model parameters β by minimizing the criterion:

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

- Once we find the model estimates, $\hat{\beta}$, we need to measure the variability of the estimator (precision) and its bias (accuracy):

$$MSE(\hat{\beta}) = E[(\hat{\beta} - E(\hat{\beta}))^2] + E\left[\left(\hat{\beta}\right) - \beta\right]^2$$

$$MSE\left(\hat{\beta}\right) = \left(bias\ of\ \hat{\beta}\right)^2 + var\left(\hat{\beta}\right)$$

More about MSE

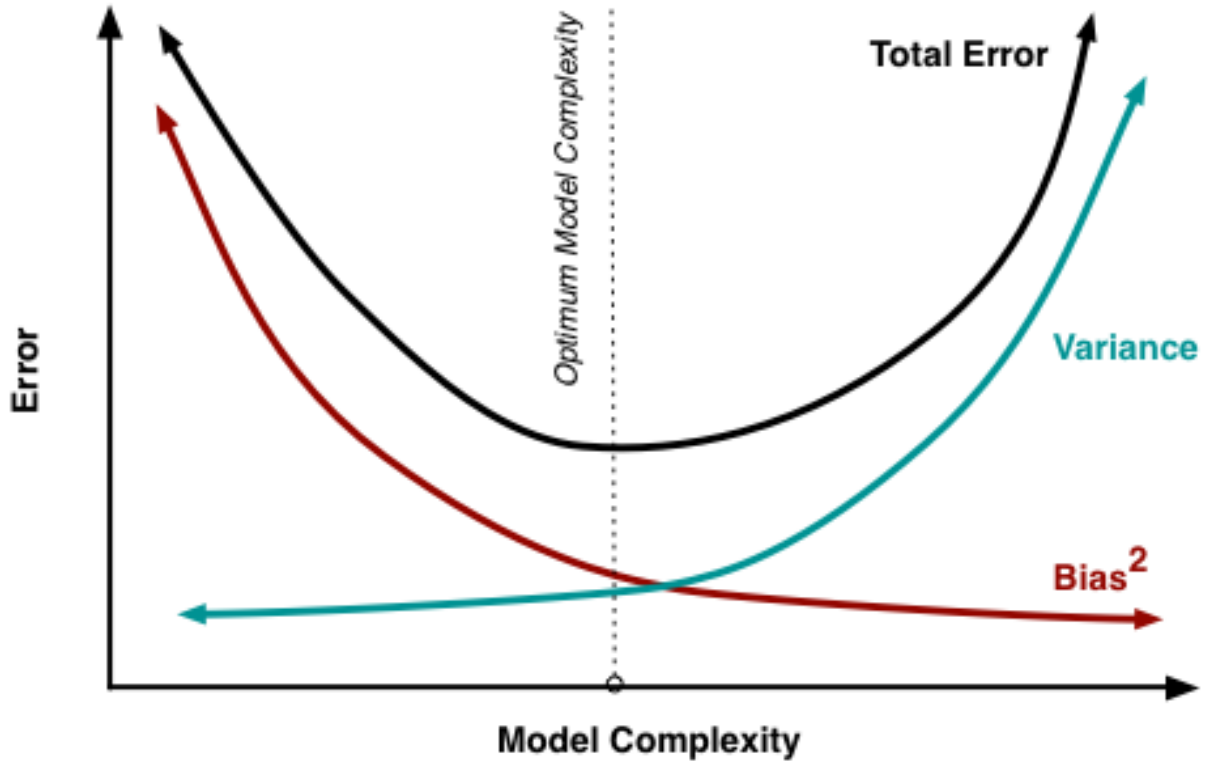
- Goal: minimize the MSE: zero bias and small variance for the estimators
- How does this translate to our regression model?

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-(p+1)} = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)}$$

- MSE is also an estimator of the error variance σ^2
- In regression an estimator function equal to the 'real' dependency of Y on X will have an MSE equal to the variance of the random error term

Shortcoming of Linear Regression

- Linear regression has low bias, but can have high variance. Trade-off?



Variable Selection

- Several variable selection techniques already exist, with caveats
  - Automated search procedures: forward, backward, stepwise, etc.
  - Criterion-based procedures: adjusted R^2, AIC, Mallow's C_p

- What if we consider many predictors in our model: $p \approx n \; or \; p > n$?
  - A large number of predictors increases the chances of multicollinearity
  - Overfitting – model fits the date well, but performs poorly for prediction
  - Linear regression is no defined for p > n (no unique solution)

- Examples of high-dimensional data: $p \gg n$
  - Precision Medicine – most 'omics' analysis
  - E.g., estimate the risk of cancer based on DNA sequence: n=1,000 patients and p=300,000 variables

Penalized Regression Methods

- Goal: decrease the model complexity, reduce the variability and improve the accuracy of the linear regression models

- Automated procedures (forward/backward) do not tell anything about the removed variables' effect on response

Solutions?

- 1. Ridge Regression: penalizes the coefficients if they are too far from zero, forcing them to be small in a continuous way; this shrinks the coefficients towards zero, but it does not set them to be exactly zero

- 2. Lasso Regression: also penalizes the coefficients that can be set to be exactly zero

- 3. Elastic Net: a combination of Ridge and Lasso (get the best of two worlds)

Ridge Regression

- Ridge regression is similar to (regular) regression, but the coefficients are estimated by minimizing a different quantity

- We are not only minimizing the sum of squared residuals, but also the size of the estimates

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

- $\lambda \geq 0$ is called the *tuning parameter*

- $\lambda \sum_{j=1}^{p}{\beta_j}^2 \geq 0$ is called a *shrinkage penalty term*, it has the effect of shrinking the estimates towards zero.

More about Penalty

- We use Lagrange multipliers to turn constrained optimization problems into unconstrained but penalized ones
- The nature of the penalty term reflects the sort of constraint we put on the problem
  - Shrinkage (Ridge)
  - Sparsity (Lasso - later)
- When $\lambda = 0$, the penalty has no effect, will get the least squares (LS) estimates
- When $\lambda \to \infty$, the shrinkage penalty grows and estimates will approach zero

Ridge Regression

- To shrink or not to shrink?
- Ridge performs well when there is a subset of true coefficients that are small or even zero
- Ridge is still helpful for moderately large coefficients (b/w 0.5 - 1), but its advantage is not as dramatic (and smaller range for λ)
- Ridge cannot perform variable selection; the penalty term will shrink all the coefficients towards zero, but not exactly zero
- The penalty term is unfair if the predictors are not on the same scale
  - Better to center and scale the predictors if not measured in the same units, and then perform Ridge regression
  - R automatically centers and scales variables in Ridge regression

The Lasso

- Lasso = Least Absolute Shrinkage and Selection Operator
- Proposed by **Robert Tibshirani**, 1996, *"Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88.*
    - *Lasso can be applied in linear, logistic, Cox regression*
- Ridge model always includes all the declared predictors. Lasso overcomes this disadvantage and performs variable selection by using a different penalty term.

The Lasso

- Again, we are not only minimizing the sum of squared residuals, but also the size of the estimates:

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \quad (3)$$

- High values of λ will set the coefficients to zero

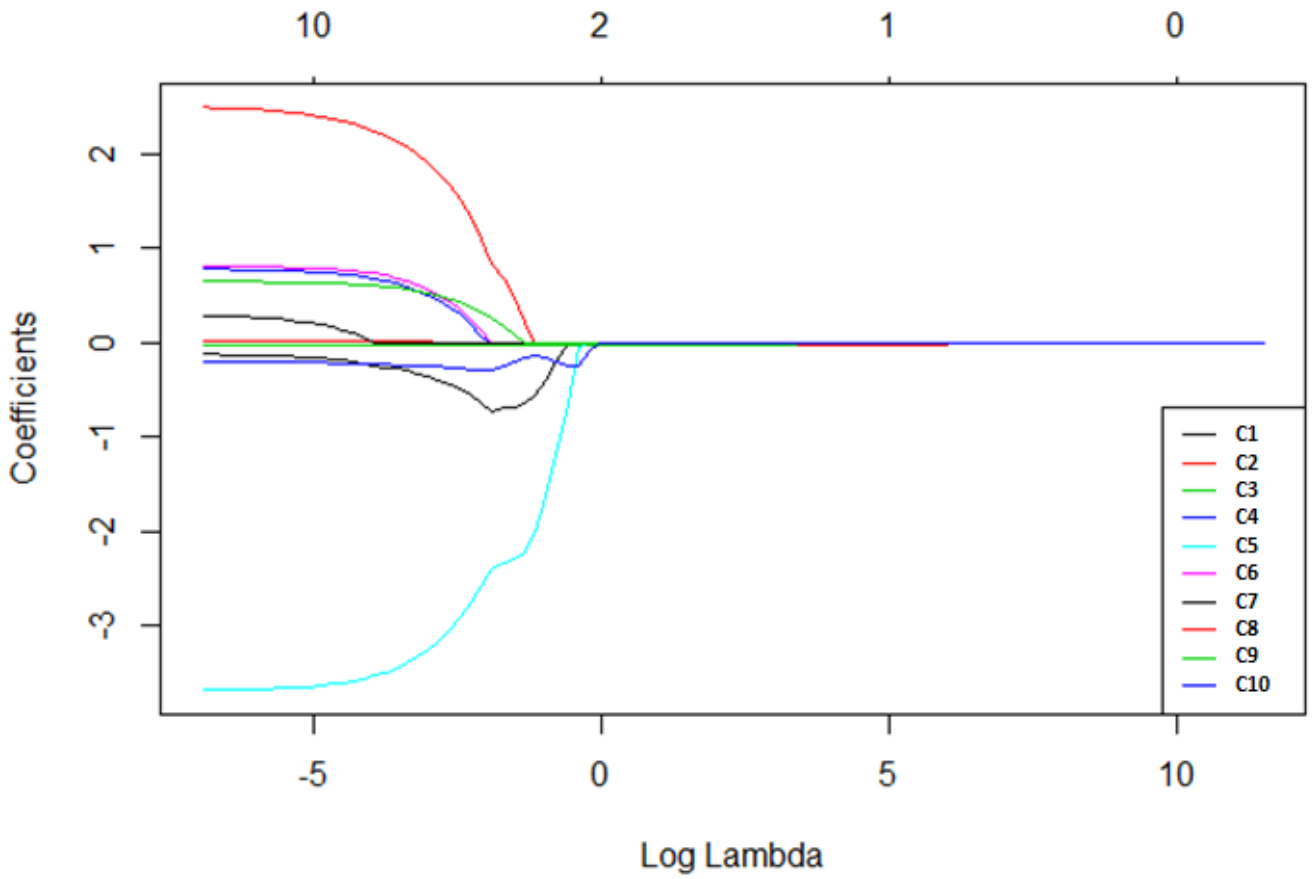
- In the final model the response variable will only be related to a small subset of predictors with nonzero coefficients

The Lasso

- Ridge and Lasso use different penalties, but they are similar
  - λ is non-negative tunning parameter that controls model complexity
  - $\lambda = 0$, we get LS estimates
  - $\lambda \to \infty$, the shrinkage penalty grows and estimates will be zero
- Lasso yields sparse models, i.e., models that involve only a subset of the variables
- Advantage? Lasso models are easier to interpret than those produced by Ridge regression

Lasso Coefficients

Notice how increasing lambda shrinks the model coefficients:



Another formulation: Lasso vs Ridge

Ridge:

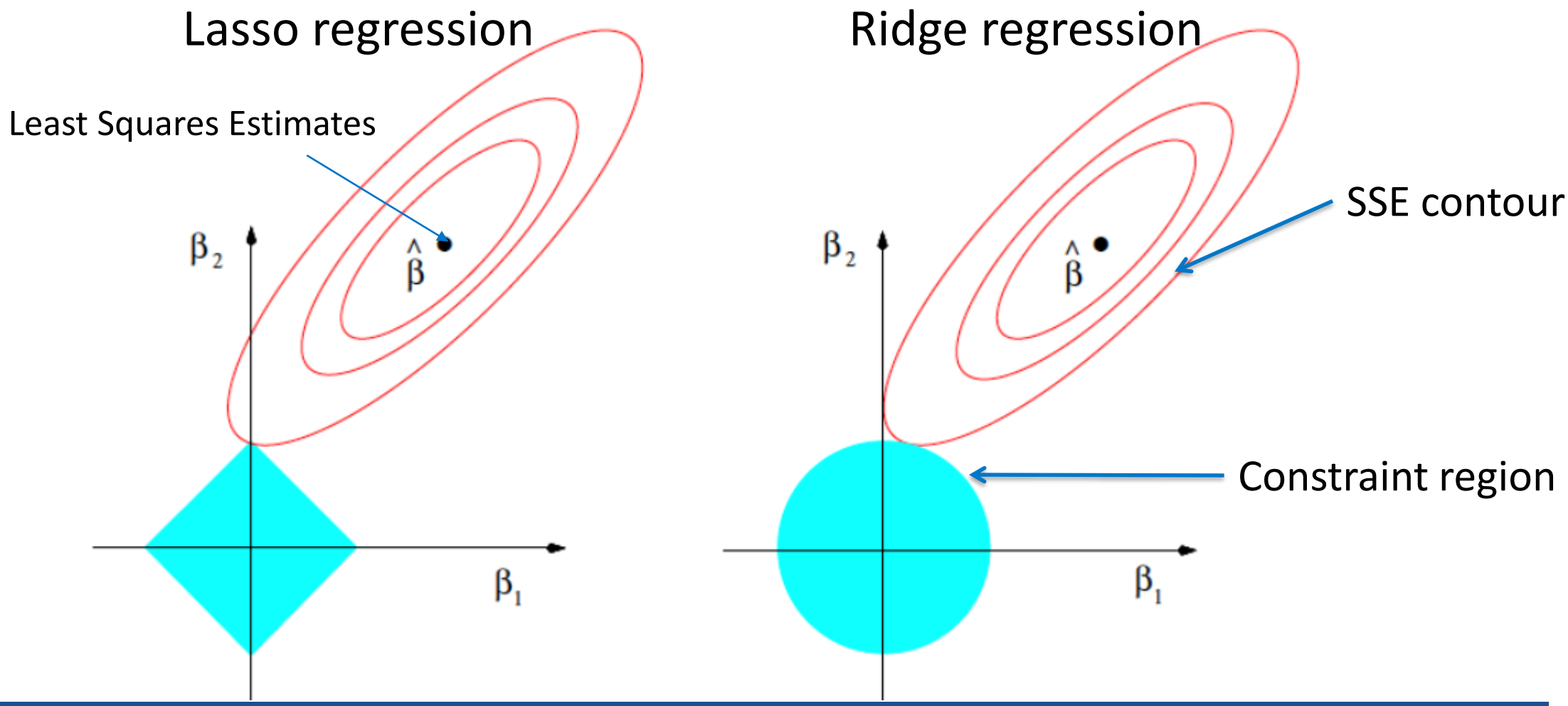
$$minimize \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 subject\ to \sum_{j=1}^{p}{\beta_j}^2 \leq s \ (4)$$

Lasso:

$$minimize \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 subject\ to \sum_{j=1}^{p}|\beta_j| \leq s \ (5)$$

- For every value of $\boldsymbol{\lambda}$, there is some $\boldsymbol{s}$ such that equations:
  (2) and (4) will give the same Ridge estimates
  (3) and (5) will give the same Lasso estimates

Sum of Squared Residuals (SSE) and Constraints



Lasso vs Ridge

- Lasso produces simpler and more interpretable models involving only a set of predictors
- The minimum MSE of Ridge regression is slightly lower than Lasso (Why?)
- Ridge will perform better when all the variables/features are associated with the outcome/response
- Lasso will perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining are very small or close to zero

Lasso Limitations

- If there is a group of variables among which the pairwise correlations are very high, then the Lasso tends to select only one variable or a few and shrinks the rest to 0
    - Not always ideal: e.g., microarray analysis – several correlated genes may all contribute to the biological process
- When n > p, and there are high correlations between predictors, Lasso can be inferior to Ridge in terms of prediction performance

(Some) Versions of Lasso

- Adaptive Lasso
  - Use adaptive weights for penalizing coefficients to reduce the estimation bias and improve accuracy
  - Higher penalty for zero coefficients and lower for non-zero coefficients

- Relaxed Lasso
  - Sometimes it might be desirable to estimate the coefficients of all selected variable without shrinkage (hard-thresholding)
  - Relaxed Lasso controls model selection and shrinkage estimation using two separate parameters
  - Relaxed lasso is performing better when just a few variables are carrying signal

Choice of Penalty/Tunning

- We rarely know the constraint level or tunning parameter λ
    - The bias of the estimated coefficients $\hat{\beta}$ increases as penalty term decreases
    - The variance decreases as penalty increases
- How to pick the λ value, that is, the 'best' level of model complexity?
    - Classical approach: use an information criterion such some information criterion, e.g., AIC or BIC, is the smallest **(focuses on model fit)**
    - Machine learning approach: perform cross-validation and select the value of λ that minimizes the cross-validated sum of squared residuals **(focuses on model predictive capability)**

Choice of Penalty/Tunning

- **Cross-validation** is a popular and simple data-driven technique for estimating prediction error
- Widely used for estimating the appropriate regularization parameter λ:
  - Use cross-validation to select the optimal λ value, e.g., λ which extrapolates best on average
  - Choose a grid of λ values and compute cross-validation for each value
  - Select the optimal value for λ (e.g., gives the minimum average cross-validated prediction error)
  - Perform Ridge/Lasso on the full data, using the chosen λ value

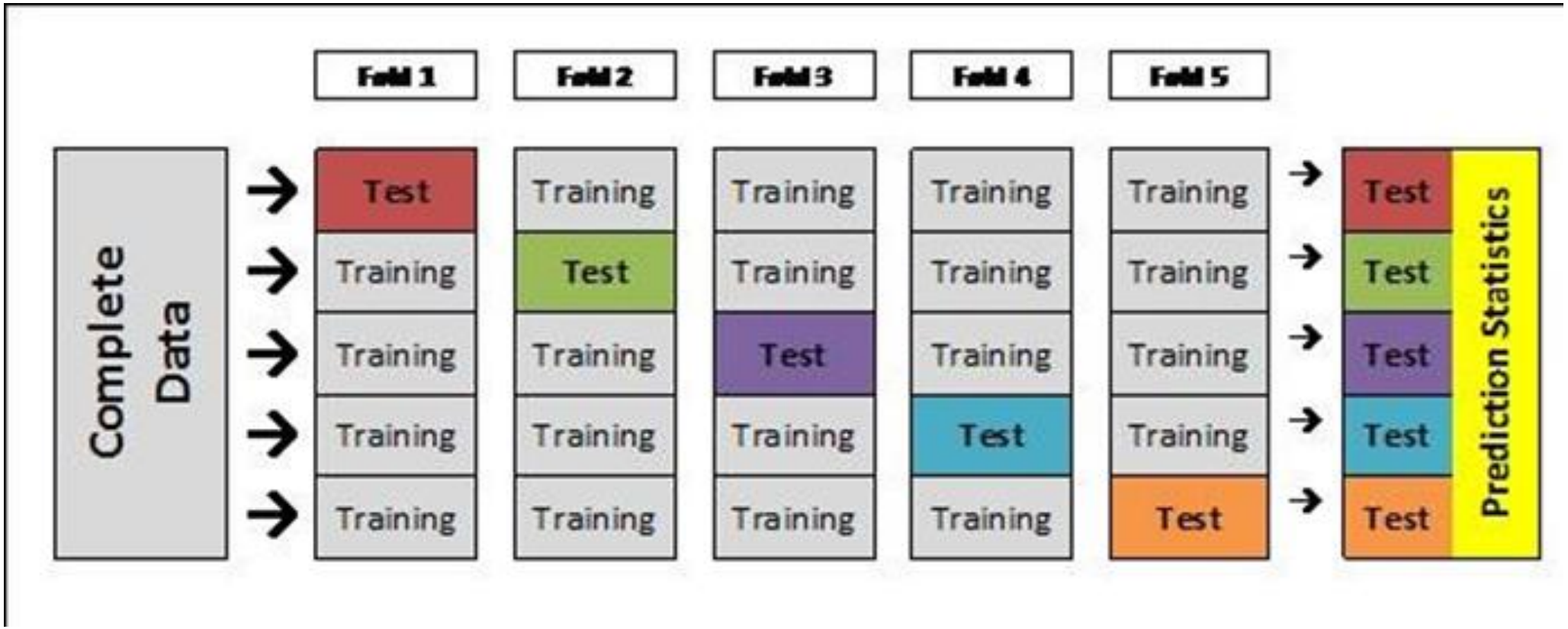
K-fold Cross-Validation (CV)

- What is k-fold cross validation?
- Divide data into k (equal-sized) folds
  - Can also do 80/20
- Use each one as a validation set, average the MSE* across sets
- Common k values: 5 and 10 (because is computational advantageous)

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

*Note: This is the test or prediction error. 'Best' is the one that generates the smallest CV.

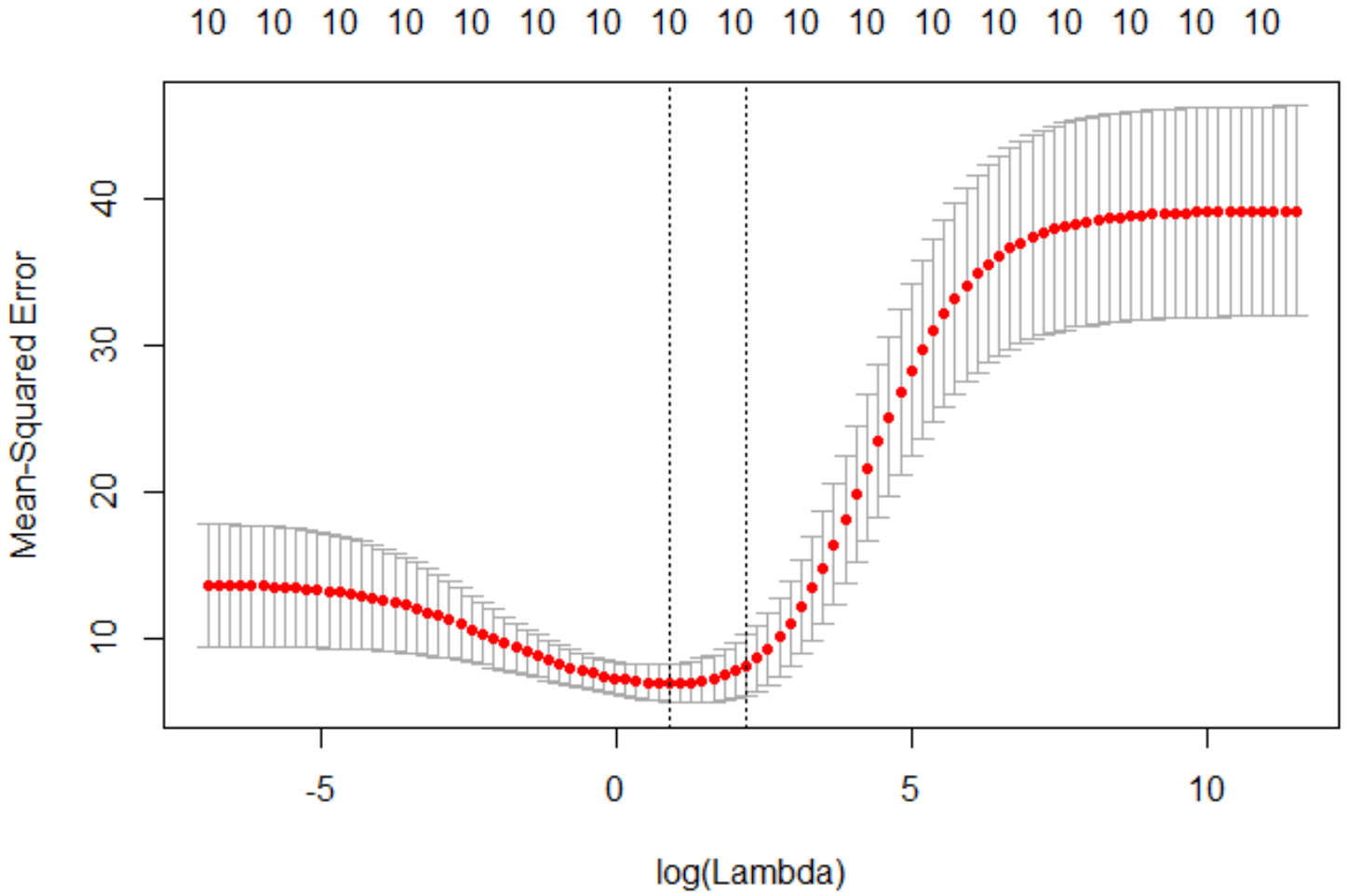
5-fold Cross-Validation (CV)



N-fold Cross-Validation

- This is a special case of k-fold CV (k=n)
- Also called Leave-one-Out Cross Validation (LOOCV)
- LOOCV works by deleting the i^{th} case, estimate the regression using n-1 observations and then obtain the fitted value of the i^{th} case
- Computational expensive, but it will give approximately unbiased estimates of the test error
- Low bias comes at the expense of higher variance: trade-off

Cross Validation: Optimal Interval



Elastic Net

- Elastic net aims at minimizing the following loss function:

$$\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\left(\sum_{j=1}^{p}\frac{(1-\alpha)}{2}{\beta_j}^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right) \quad (6)$$

- For $\alpha = 1$, elastic net is the same as Lasso
- As $\alpha \to 0$, elastic net approaches Ridge regression
- The two terms of the penalty function play two complementary roles:
    - ℓ_1 part (absolute value) performs automatic variable selection
    - ℓ_2 part (square) stabilizes the solution paths and improves the prediction

Elastic Net: Tuning Parameters

- There are two tuning parameters in the elastic net, thus we need to use cross-validation in a two-dimensional surface - Easy!
- Use R to tune λ and α
- The R package ***glmnet()*** allows to tune λ via cross-validation for a fixed α, but it does not support α-tuning
- The R package ***caret()*** is able to tune both parameters for Elastic Net

When is Elastic Net preferred?

- In simulations, it has been shown that elastic net often outperforms Lasso in terms of prediction accuracy
- Lasso regression performs better than Ridge in scenarios with many noise predictors and worse in the presence of correlated predictors
- Elastic net, is a hybrid of the two, and performs well in all these scenarios, especially when the number of observations is larger than the number of predictors

Other interesting questions…

Penalized regression output show only the coefficient estimates, but not test statistics and/or p-values

- There are tools for post-selection inferences to use with Lasso (Gaussian, logistic, and Cox models)

- Adjusted inferences by conditional probability (in general, the adjusted p-values will be much larger than the 'unadjusted' ones); R package ***selectiveInference()***

- Standard errors not straightforward to obtain, but you can use resampling-based methods, or Bayesian Lasso

- Other ideas: "A significance test for the Lasso" (Lockhart *et al*, 2014)

Other interesting questions…

What results/coefficients are we going to report? Can we re-fit a linear regression with the non-zero predictors found in Lasso?!

- Typically leads to exaggerated effect sizes, invalid p-values and confidence intervals with below nominal coverage

Is the goal to reduce bias?

- Lasso coefficients are biased due to soft thresholding
- Reduce bias by performing least-squares on a subset of variables corresponding to non-zero Lasso regression coefficients
- Alternatively, use the Adaptive Lasso

Some References

James G, Witten D, Hastie T, Tibshirani R,
An Introduction to Statistical Learning (with R applications)

https://www.springer.com/us/book/9781461471370

Molnar C, *Interpretable Machine Learning*

https://christophm.github.io/interpretable-ml-book/

THANK YOU