Day 6: Model Selection II

Lucas Leemann

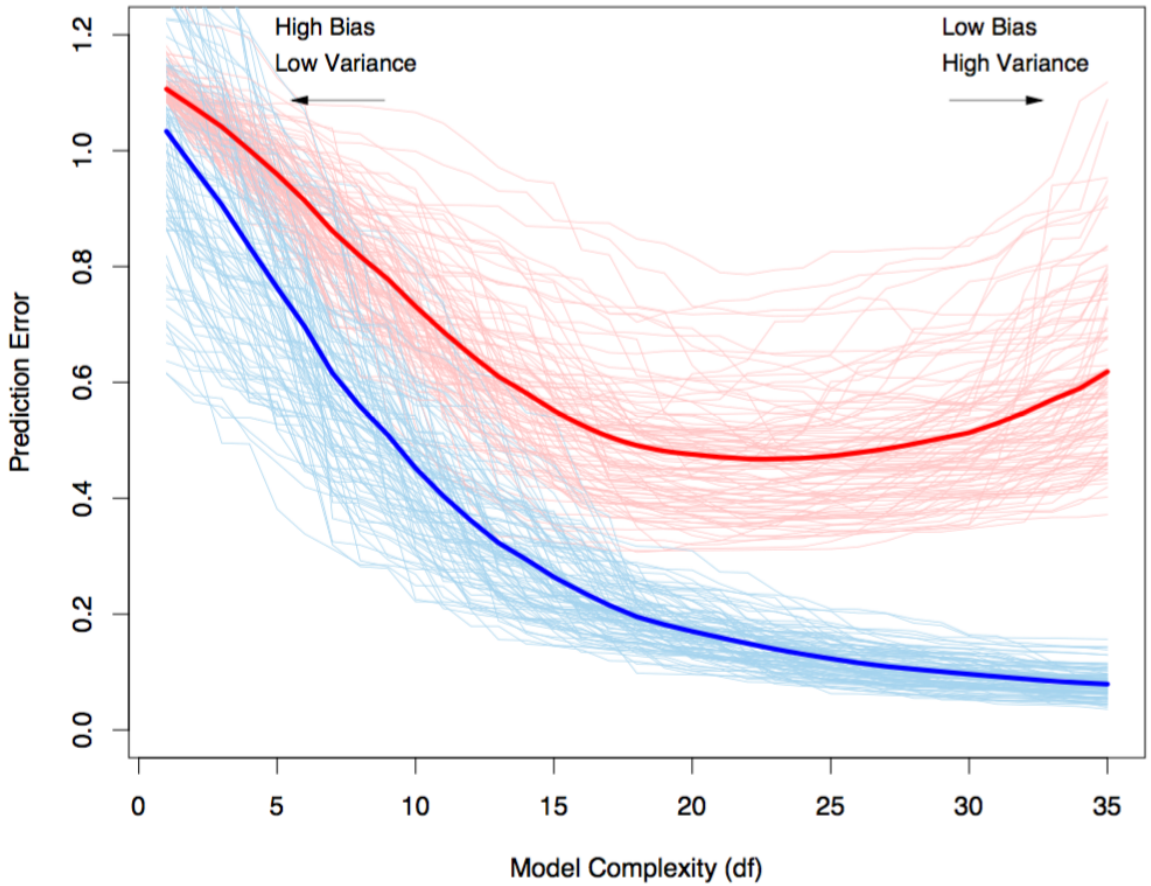
Essex Summer School

Introduction to Statistical Learning

1. Repetition Week 1

2. Regularization Approaches
   - Ridge Regression
   - Lasso
   - Lasso vs Ridge

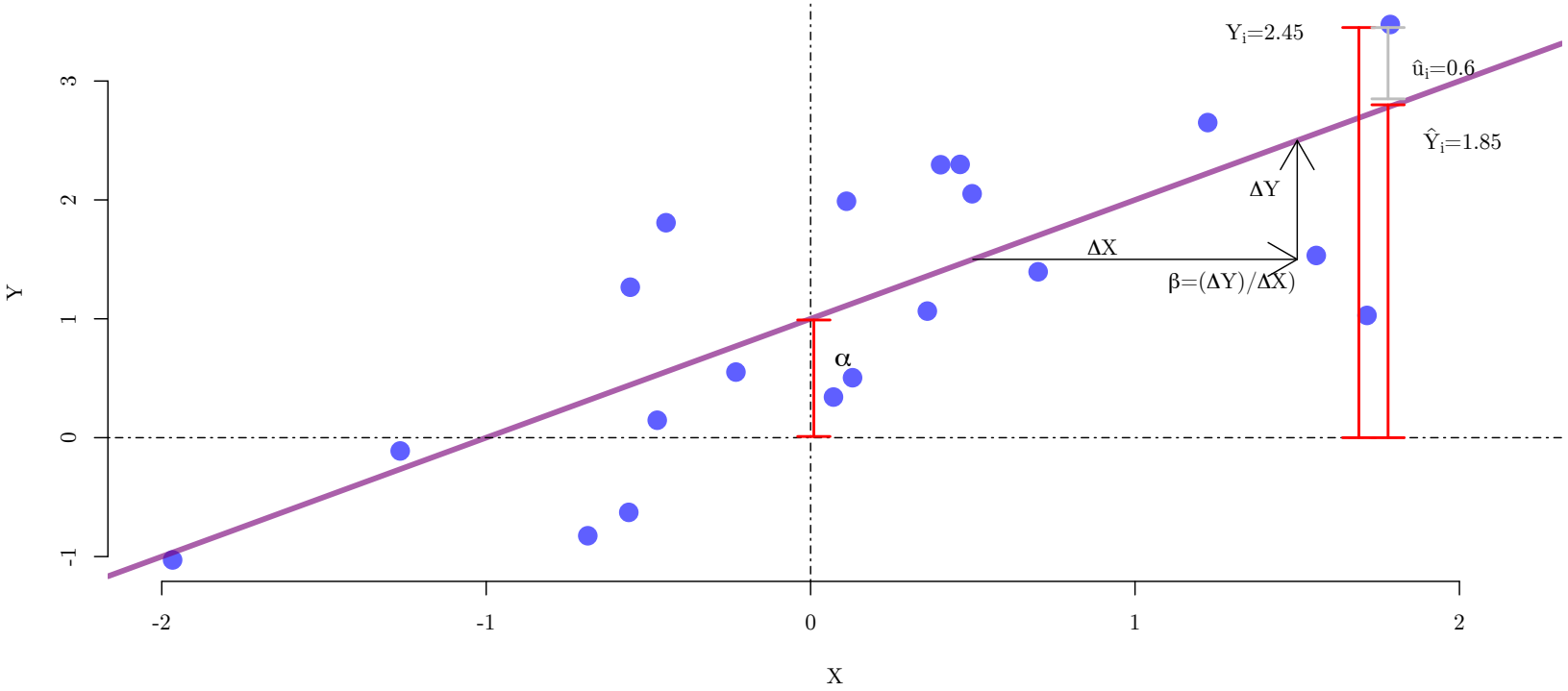
Repetition: Fundamental Problem



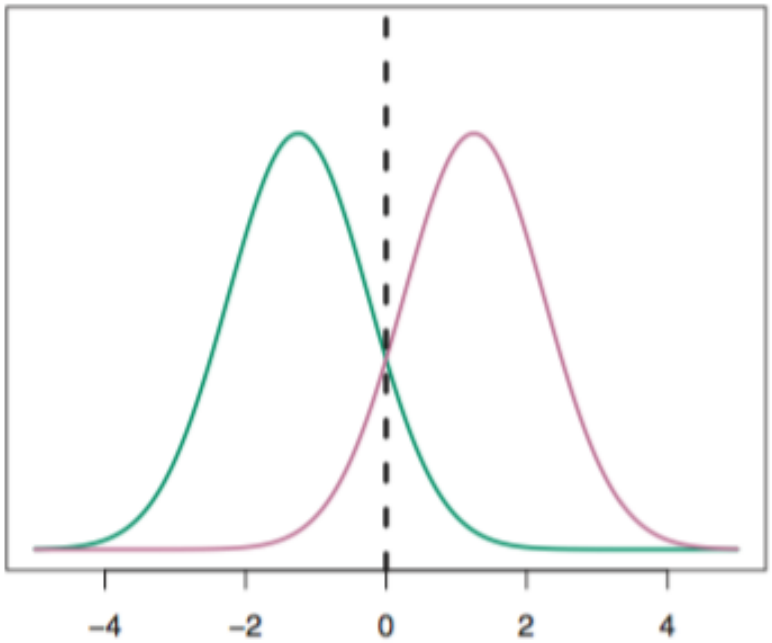
Red: Test error.
Blue: Training error.

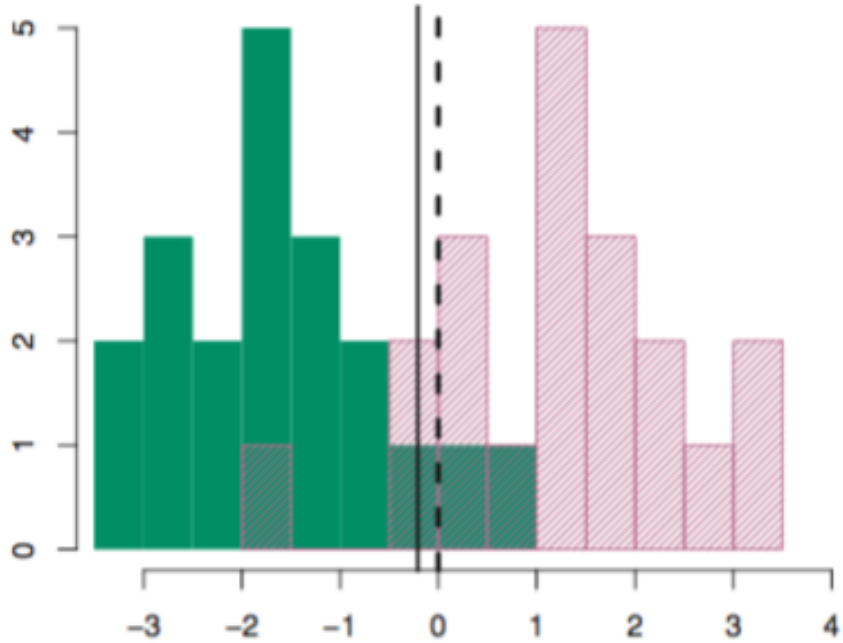
(Hastie et al, 2008: 220)

Tuesday: Linear Models



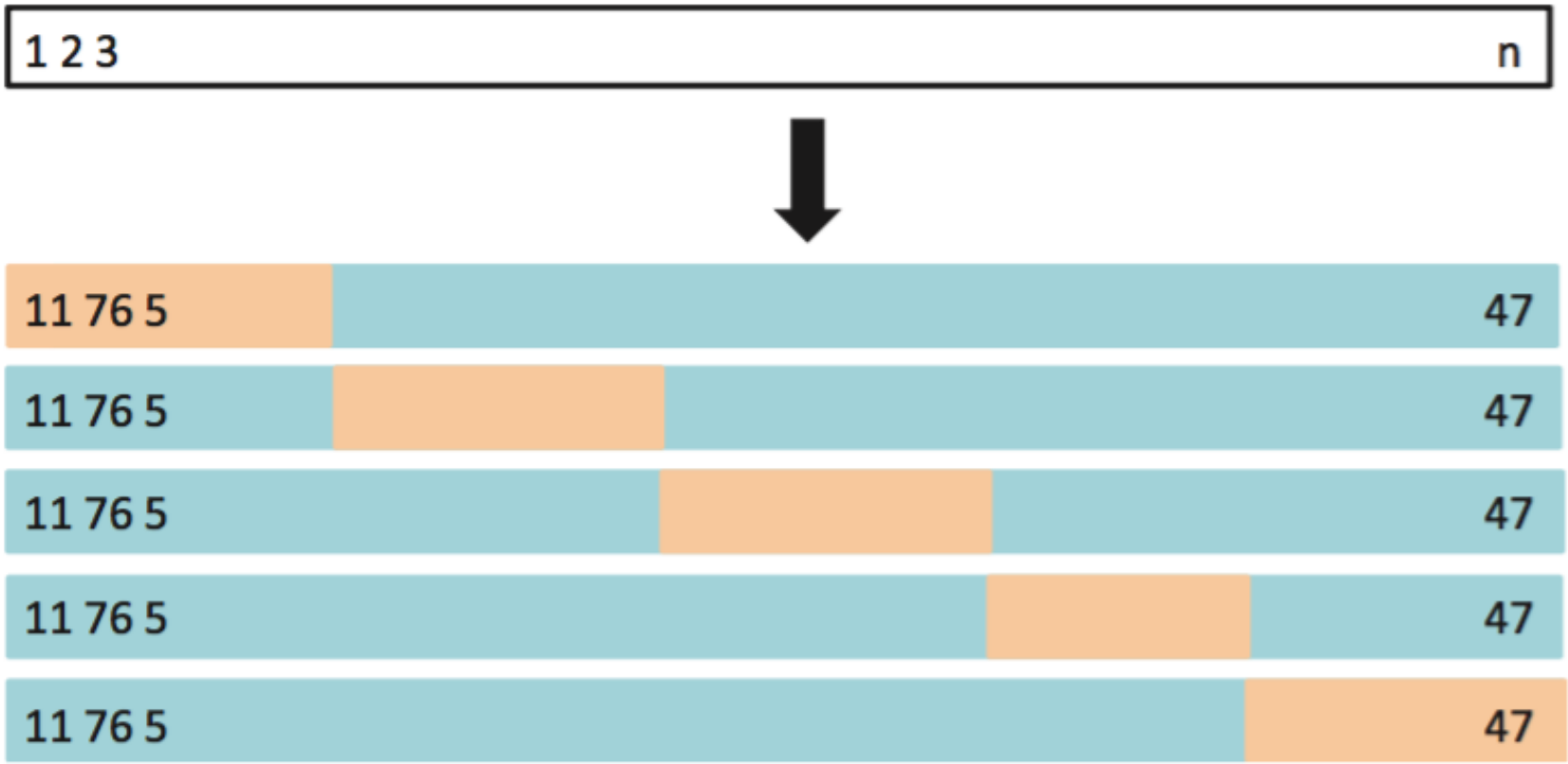
Wednesday: Classification





(James et al, 2013: 140)

Thursday: Resampling



(James et al, 2013: 181)

Friday: Model Selection I

Subset Selection:

1. Generate an empty model and call it $\mathcal{M}_0$
2. For $k = 1....p$:
   i) Generate all $\binom{p}{k}$ possible models with k explanatory variables
   ii) determine the model with the best criteria value (e.g. R^2) and call it $\mathcal{M}_k$
3. Determine best model within the set of these models: $\mathcal{M}_0$, ..., $\mathcal{M}_p$ - rely on a criteria like AIC, BIC, R^2, C_p or use CV and estimate test error

Regularization Approaches

Shrinkage Methods

Ridge regression and Lasso

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can 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.
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

Regularization

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 \quad = \quad RSS$$

- In contrast, the regularization approach minimizes:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda f(\beta_j) \quad = \quad RSS + \lambda f(\beta_j)$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately.

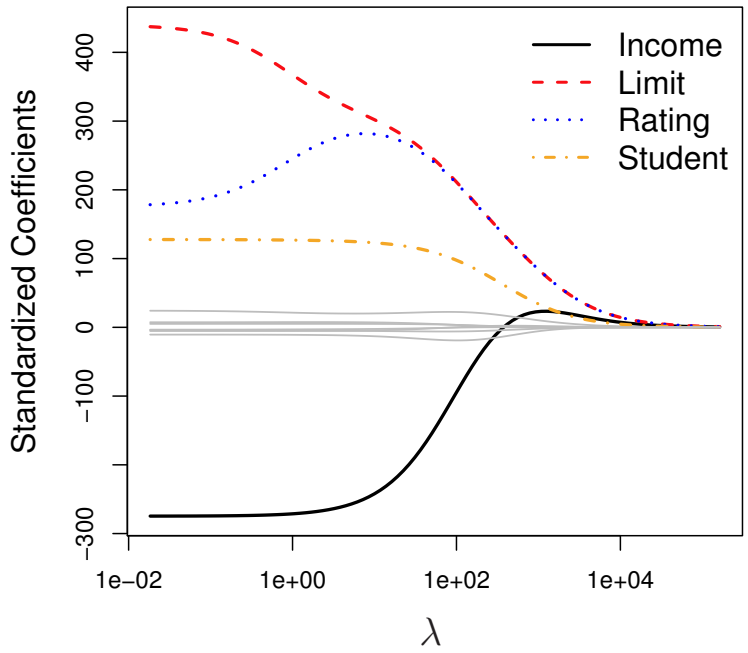
Ridge Regression

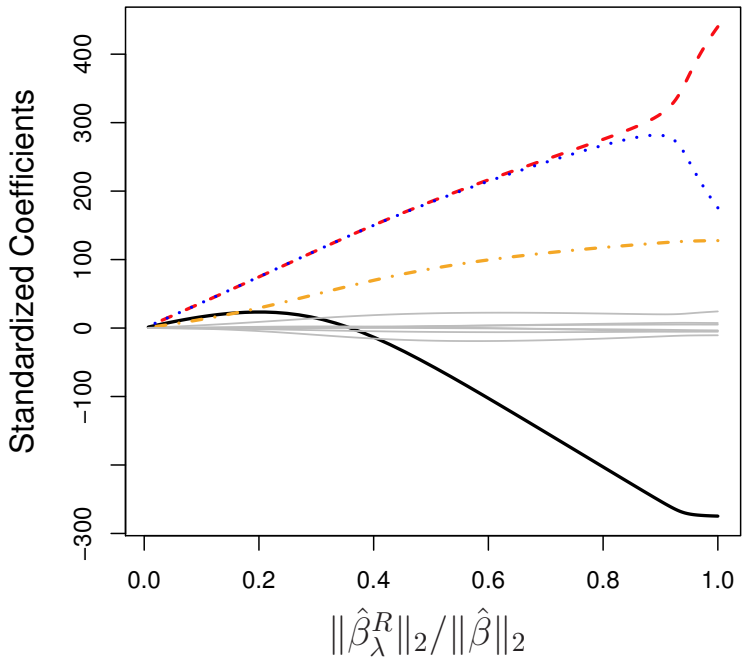
- Ridge Regression minimizes this expression:

$$\underbrace{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{J}\beta_j x_{ij}\right)^2}_{standard\ \ OLS\ \ estimate}+\underbrace{\lambda\sum_{j=1}^{J}\beta_j^2}_{penalty}$$

- λ is a tuning parameter, i.e. different values of λ lead to different models and predictions.
  - When λ is very big the estimates get pushed to 0.
  - When λ is 0 the ridge regression and OLS are identical.
- We can find an optimal value for λ by relying on cross-validation.

Example: Credit data





$$||\hat{\beta}||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$$

(James et al, 2013: 216)

Ridge Regression: Details

- Shrinkage is not applied to the model constant β_0, model estimate for conditional mean should be *un-shrunk*.
- Ridge regression is an example of ℓ_2 regularization:
  - $\ell_1 : f(\beta_j) = \sum_{j=1}^{J} |\beta_j|$
  - $\ell_2 : f(\beta_j) = \sum_{j=1}^{J} \beta_j^2$

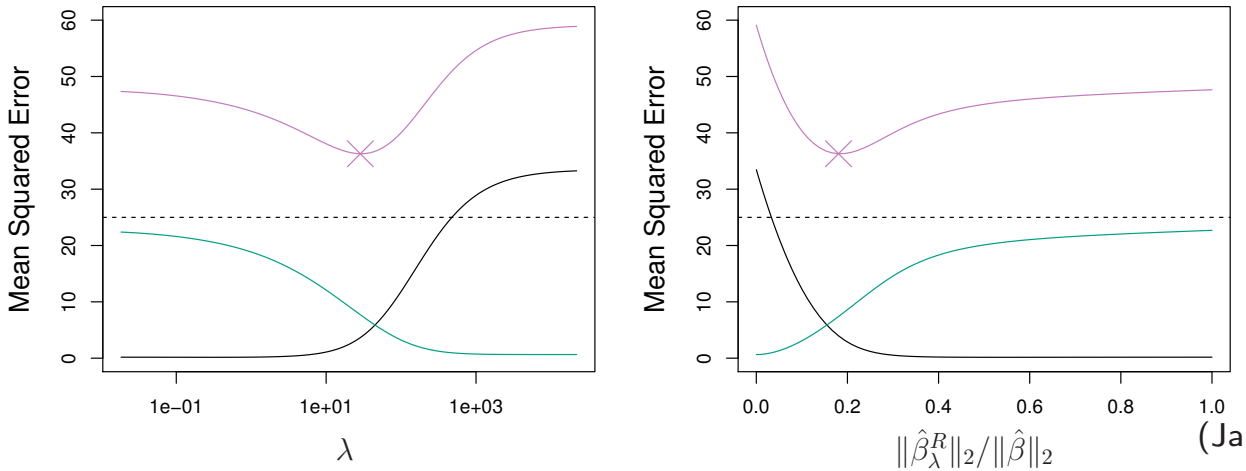
$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

Ridge regression: scaling of predictors

- The standard least squares coefficient estimates are scale equivariant: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. In other words, regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after standardizing the predictors, using the formula

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

Why Does Ridge Regression Improve Over Least Squares?



(James et al, 2013: 218)

- Simulated data with $n = 50$ observations, $p = 45$ predictors, all having nonzero coefficients.
- Squared bias (black), variance (green), and test mean squared error (purple).
- The purple crosses indicate the ridge regression models for which the MSE is smallest.
- OLS with p variables is low bias but high variance - shrinkage lowers variance at the price of bias.

The Lasso

- Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.
- The Lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}^L_\lambda$, minimize this quantity

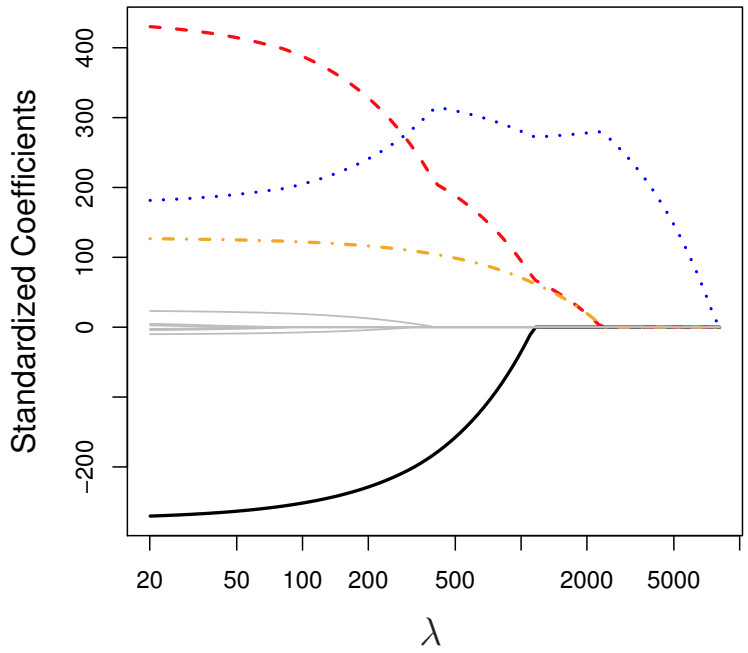
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

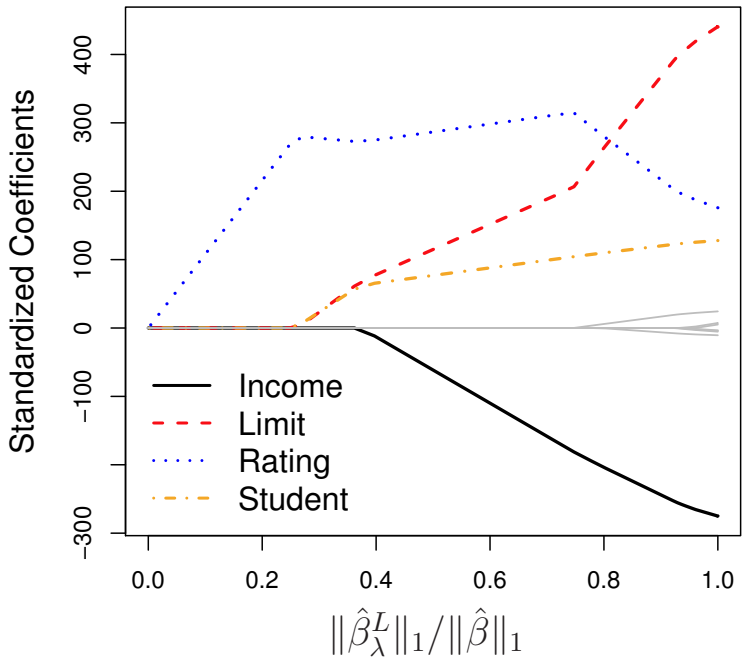
- In statistical parlance, the lasso uses an ℓ_1 (pronounced "ell 1") penalty instead of an ℓ_2 penalty. The ℓ_1 norm of a coefficient vector β is given by $\|\beta\|_1 = \sum|\beta_j|$.

The Lasso: continued

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Hence, much like best subset selection, the lasso performs variable selection.
- We say that the lasso yields sparse models – that is, models that involve only a subset of the variables.
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice.

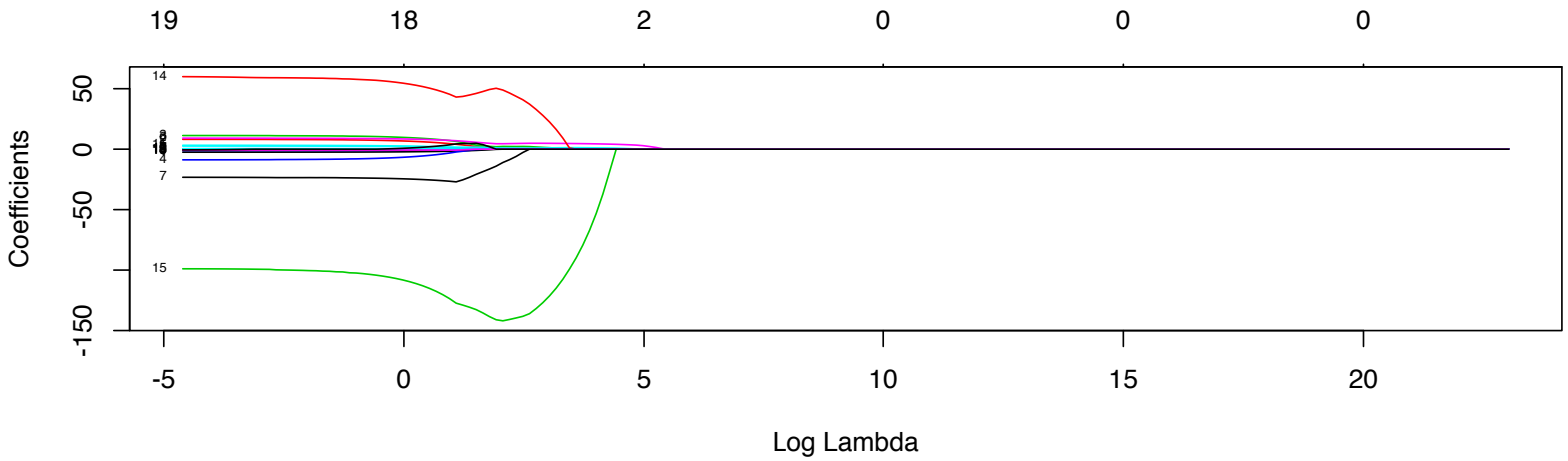
Example: Credit data

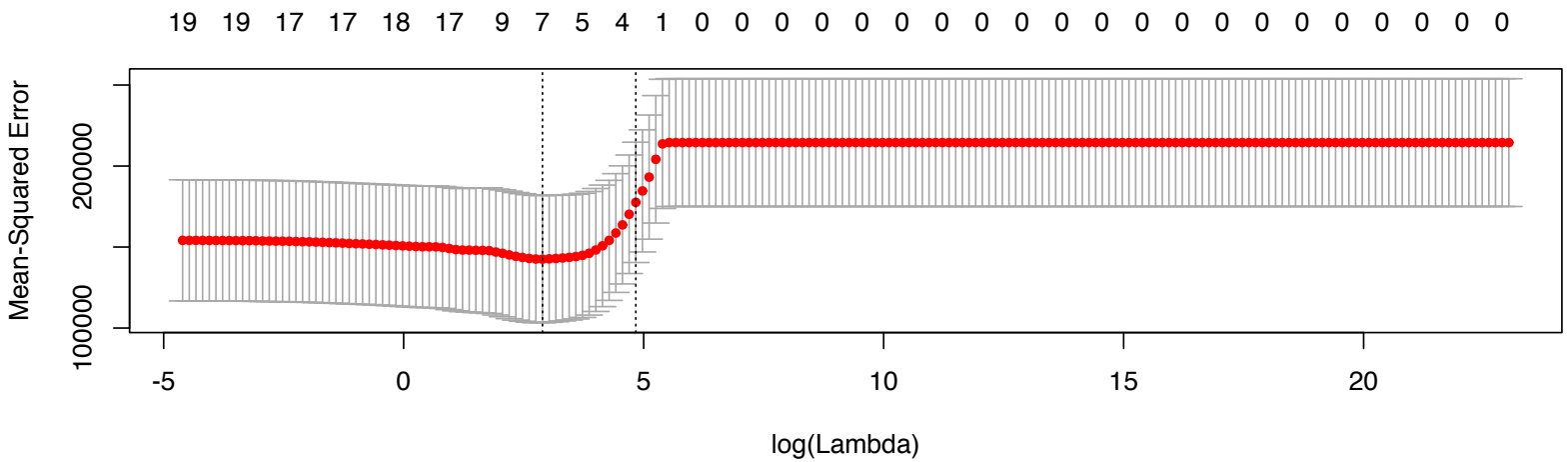




(James et al, 2013: 220)

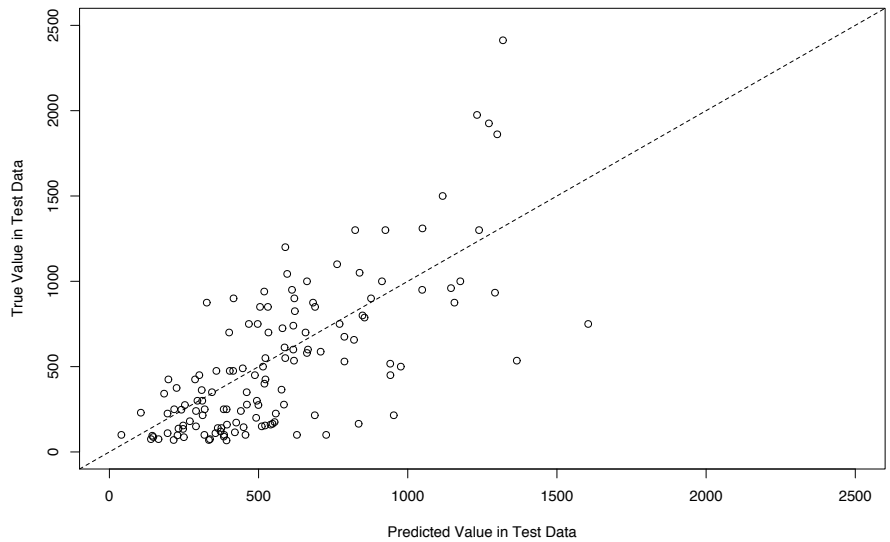
Example: Baseball Data



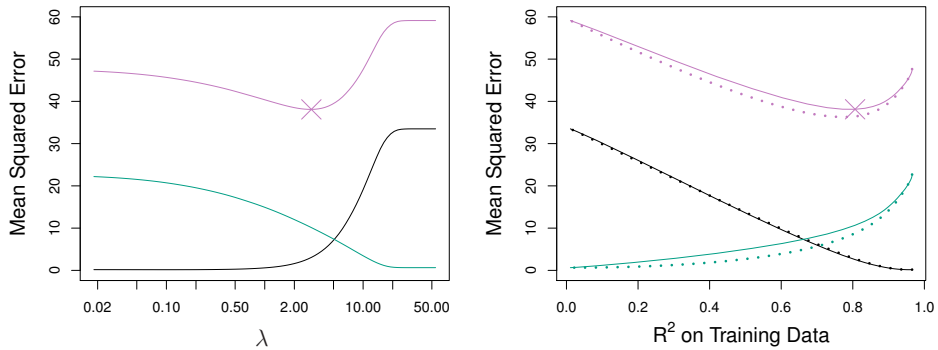


Lasso Example 4

```
> lasso.pred <- predict(lasso.mod, s = log(cv.out$lambda.1se), newx = x[test, ])
> plot(lasso.pred, y[test], ylim=c(0,2500), xlim=c(0,2500), ylab="True Value in Test Data",
   xlab="Predicted Value in Test Data")
> abline(coef = c(0,1),lty=2)
```

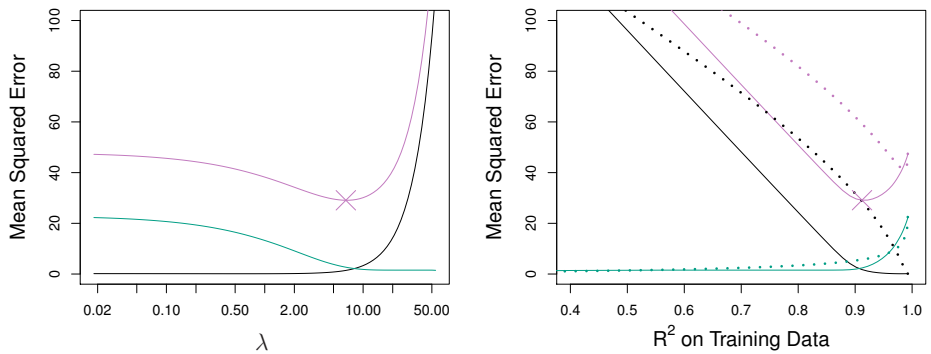


Comparing the Lasso and Ridge Regression



- Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on simulated data set.
- Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed).
- Both are plotted against their R^2 on the training data, as a common form of indexing.
- The crosses in both plots indicate the lasso model for which the MSE is smallest.

Comparing the Lasso and Ridge Regression: continued



- Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that previous slide, except that now only two predictors are related to the response.
- Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed).
- Both are plotted against their R^2 on the training data, as a common form of indexing.

Why does Lasso shrink to exactly 0?

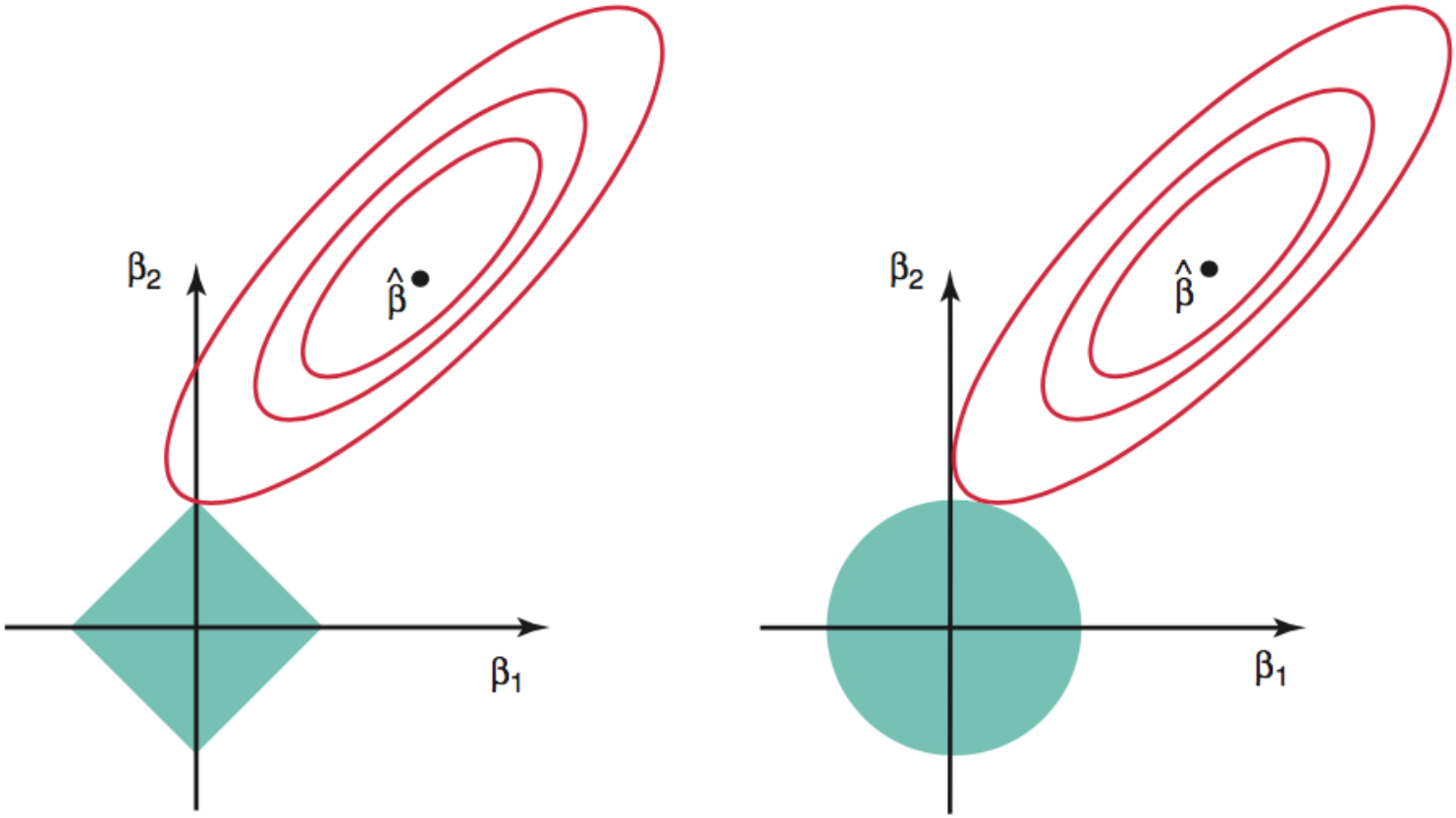


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Ridge vs Lasso

- Ridge is preferred when some features are (strongly) correlated – Lasso may only pick one.
- Elastic net: Combining Lasso and Ridge:

$$\tilde{\beta} \quad = \quad \text{argmin}\Big(RSS - \lambda \sum_{j=1}^{J} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|\Big)$$

  we now have two tuning parameters: α and λ
- Details: Hastie et al. 2008. *The Elements of Statistical Learning*. Springer.

Take away message

- The two examples illustrate that neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors.
- However, the number of predictors that is related to the response is never known *a priori* for real data sets.
- A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.
- Ridge can be expected to work better than Lasso if some features are highly correlated.

Selecting the Tuning Parameter for Ridge Regression and Lasso

- For ridge regression and lasso we require a method to determine which of the models under consideration is best.
- That is, we require a method selecting a value for the tuning parameter λ.
- Cross-validation provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ.
- We then select the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.