Linear regularization methods

Rodrigo Gonzalez, PhD

Hello!

I am Rodrigo Gonzalez, PhD

You can find me at rodralez@gmail.com

1.
Linear regression

A review

Linear regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad (6.1)$$

Commonly used to describe the relationship between a response Y and a set of variables X1 , X2 , . . . , Xp .

The least squares fitting procedure estimates β0 , β1 , . . . , βp using the values that minimize

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

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

Linear regression

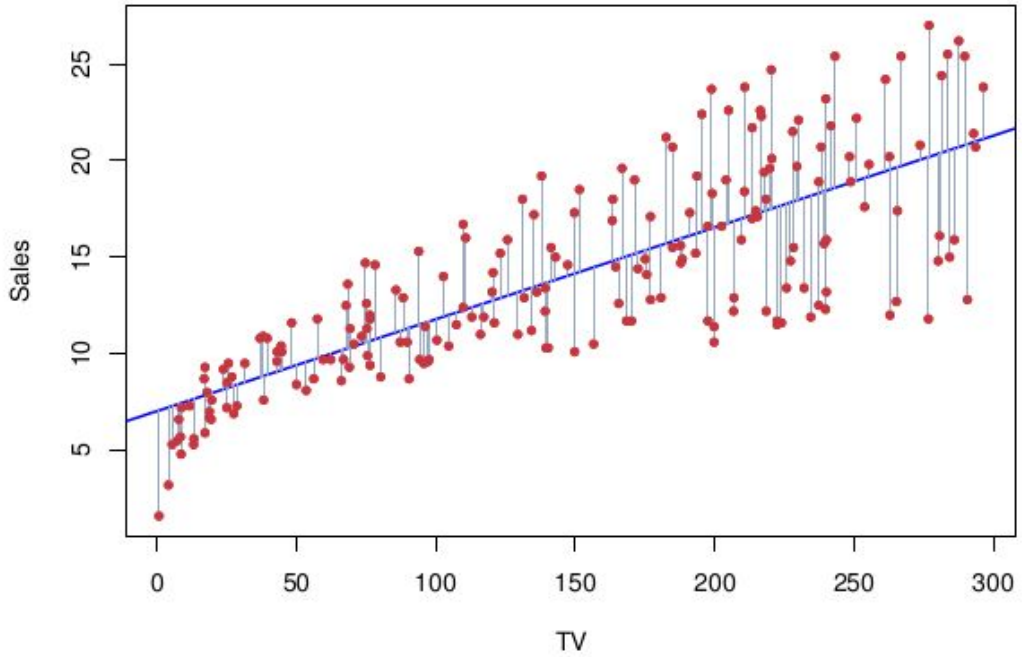


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression of* `sales` *onto* `TV` *is shown. The fit is found by minimizing the residual sum of squares. Each grey line segment represents a residual. In this case a linear fit captures the essence of the relationship, although it overestimates the trend in the left of the plot.*

Linear regression

How to evaluate a linear regression model?

- R2 statistic

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad \text{TSS} = \sum (y_i - \bar{y})^2$$

  - R2 compares the LR performance vs the mean performance.
  - The mean is your baseline solution!
- P-value
  - Is there no relationship between X and Y? (null hypothesis).
  - A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.
- MAE on testing data.

2.

Linear Model Selection and Regularization

Ridge and Lasso

Least squares fitting problems

Why might we want to use another fitting procedure instead of least squares?

- Prediction Accuracy
  - If *n* (the number of observations) is not much larger than *p* (the number of variables or predictor or features), then there can be a lot of variability in the least squares fit. Sometimes we need to shrink *p*.

- Model Interpretability
  - Including such irrelevant variables leads to unnecessary complexity in the resulting model. We see some approaches for automatically performing *feature selection* or variable selection.

Alternatives methods to linear regression

- Subset Selection
  - Identify a subset of the predictors that we believe to be related to the response.
  - 2^p possible models.
  - What if you have 30 predictors?

- Shrinkage
  - This approach involves fitting a model involving all p predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates.
  - Ridge regression and The Lasso

3.
Ridge regression

Ridge regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad (6.1)$$

Linear regression fitting

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

Ridge regression fitting

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2, \qquad (6.5)$$

$\lambda \geq 0$ is a tuning parameter

Tuning parameter

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{6.1}$$

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{6.5}$$

The second term is the *shrinkage penalty*. Not to the intercept β0.

λ ≥ 0 is a tuning parameter

- λ = 0, the penalty term has no effect.
- However, as λ → ∞, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.
- Selecting a good value for λ is critical.

Tuning parameter

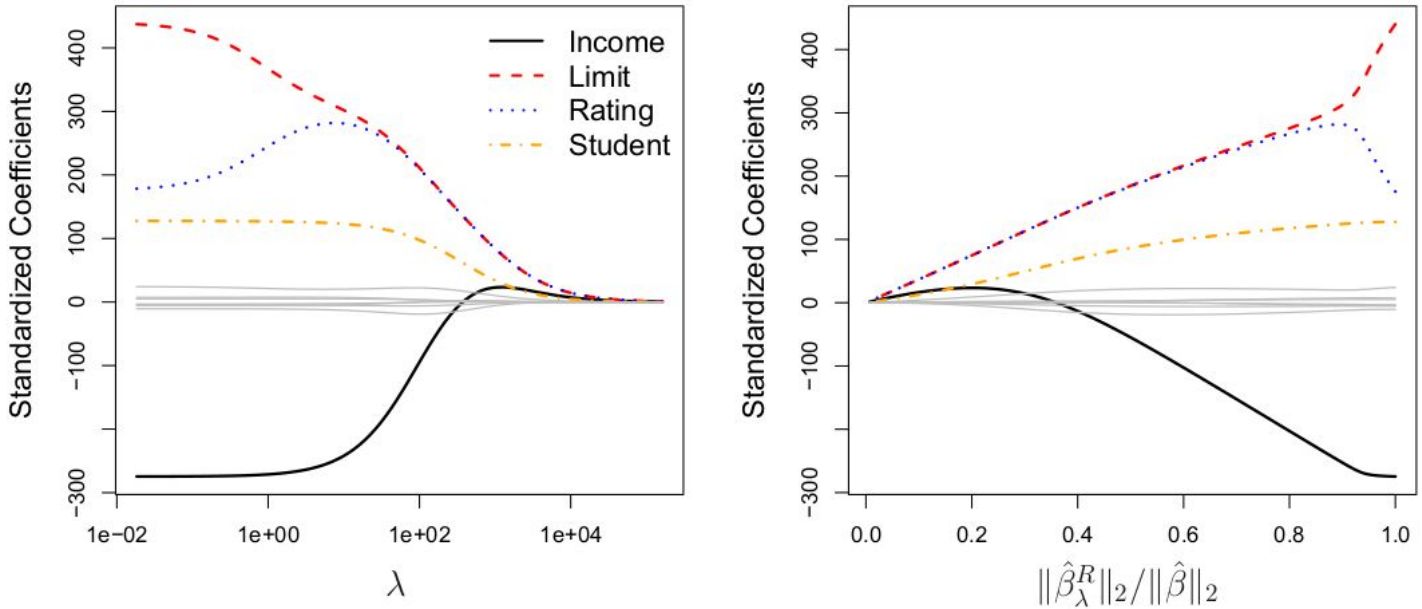


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

Ridge regression advantage over least squares

Bias-variance trade-off

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon). \qquad (2.7)$$

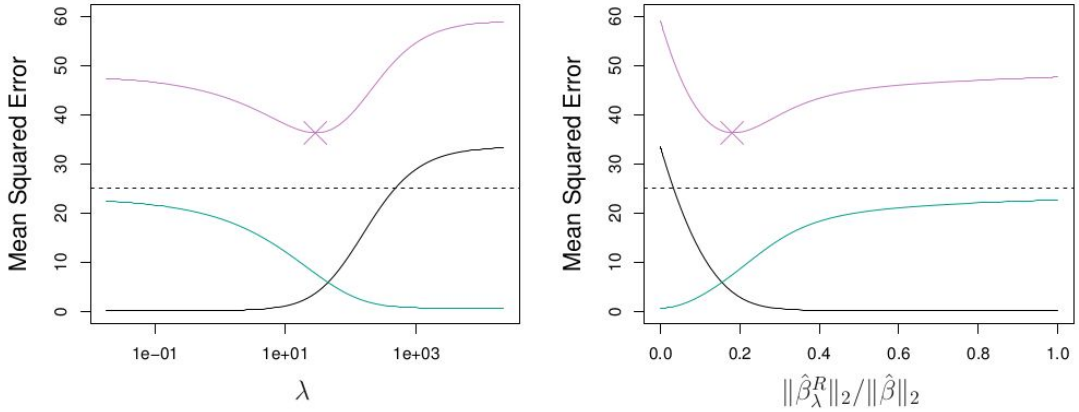


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

p = 45 predictors and n = 50 observations, The minimum MSE is achieved at approximately λ = 30

4.
The Lasso

The Lasso

Ridge regression always includes the *p* predictors.

The lasso forces some of the coefficient estimates to be exactly equal to zero.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad (6.1)$$

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|. \qquad (6.7)$$

$\lambda \geq 0$ is a tuning parameter.

The lasso performs feature selection.

The Lasso vs Ridge

In this example, all 45 predictors are related to the response.

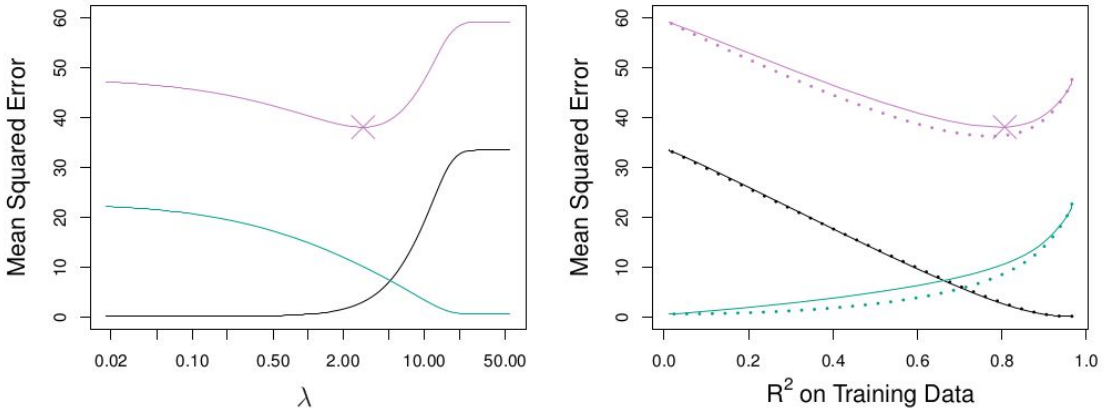


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance, and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

The Lasso vs Ridge

In this example, only 2 predictors are related to the response.

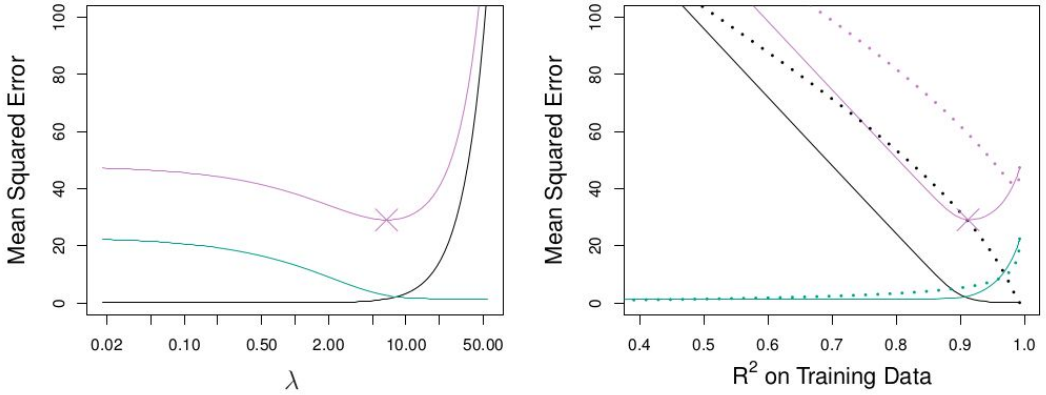


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance, and test MSE between lasso (solid) and ridge (dotted). 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.*

5.
Data standardization

Data standardization

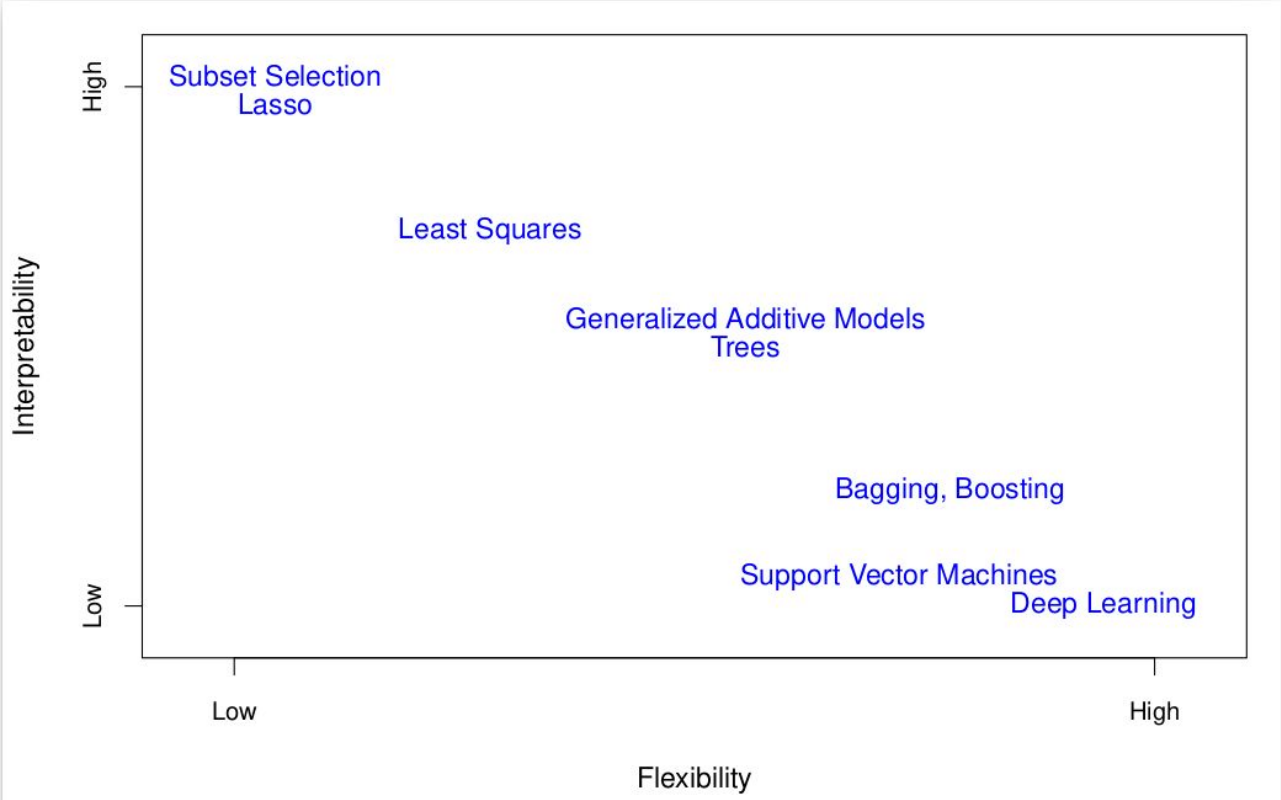
It is best to apply ridge regression or lasso regression after standardizing the predictors, using the formula:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}, \qquad (6.6)$$

so that they are all on the same scale. In (6.6), the denominator is the estimated standard deviation of the jth predictor. Consequently, all of the standardized predictors will have a standard deviation of one. As a result the final fit will not depend on the scale on which the predictors are measured.

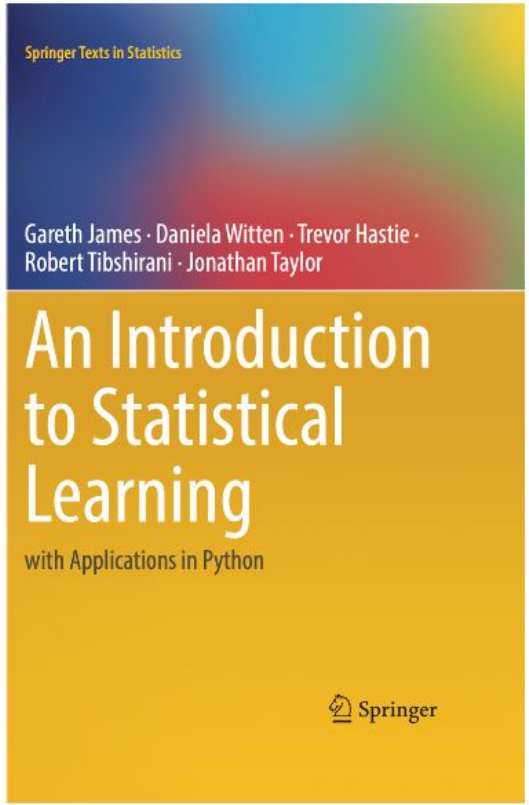
6.
Models interpretability

Tradeoff between flexibility and interpretability



7.
The book

An Introduction to Statistical Learning With Application in Python



- https://www.statlearning.com/

  Chapter 6

- https://hastie.su.domains/ISLP/ISLP_website.pdf

Thanks!

Any questions?

You can find me at:

rodrigo.gonzalez@ingenieria.uncuyo.edu.ar