Data Science for Public Policy

Aaron R. Williams - Georgetown University

Regularization

Linear regression

Ordinary least squares (OLS) linear regression generates coefficients that minimize the sum of squared residuals.

$$\min(SSE) = \min\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)$$

It is impossible to solve for OLS when $p > n$ where p is the number of predictors and n is the number of observations. Additionally, OLS is unstable when data contain multicollinearity.

Regularization

为什么要做这个？要查阅前文

Regularization/Penalization: To reduce the magnitude of parameters (coefficients).

Regularization, or penalization, allows linear regression to work with very wide data, to generate stable estimates for data with multicollinearity, and to perform feature selection.

For regression, the idea is to add a penalty P to the optimization routine:

$$\min(SSE + P) = \min\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + P\right)$$

三种不同的penalty

Ridge Regression

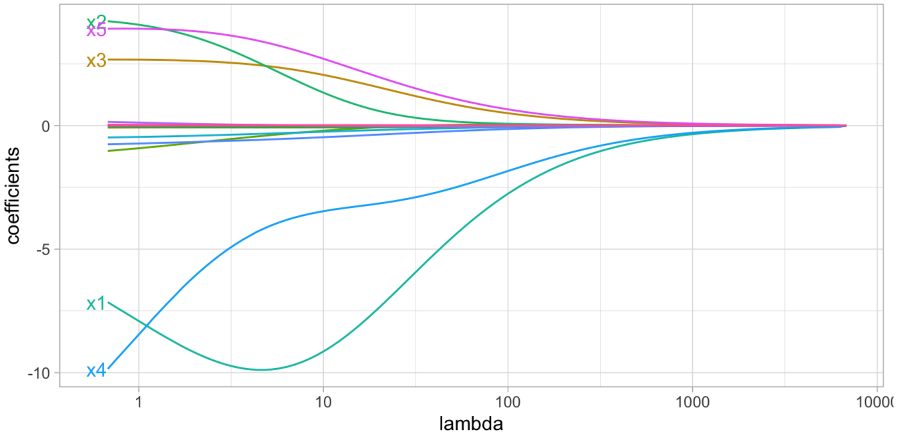
$$\min(SSE + P) = \min\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right)$$

Ridge regression adds an L2 penalty to the optimization routine. The model has one hyperparameter, λ, which determines how much penalty to add. There is no penalty when $\lambda = 0$ (just OLS).

All variables should be centered and scaled (standardized) before estimation. Thus the coefficients will be in standardized units.

Ridge regression reduces coefficients but it does not eliminate coefficients.

Ridge regression reduces but does not eliminate coefficients



Source: Hands on Machine Learning with R

LASSO Regression

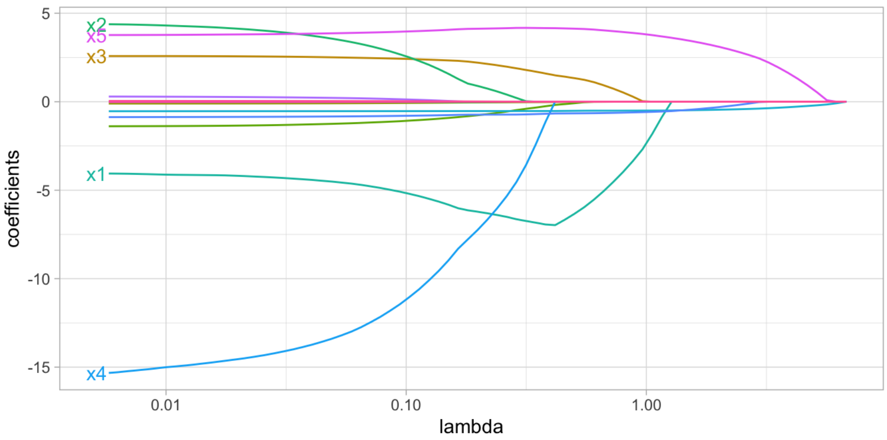
$$\min(SSE + P) = \min\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right)$$

Least Absolute Shrinkage and Selection Operator (LASSO) regression adds an L1 penalty to the optimization routine. The model has one hyperparameter, λ, which determines how much penalty to add. There is no penalty when $\lambda = 0$ (just OLS).

All variables should be centered and scaled (standardized) before estimation. Thus the coefficients will be in standardized units.

LASSO regression can regularize coefficients all the way to zero.

LASSO regression eliminates coefficients



Source: Hands on Machine Learning with R

Elastic Net Regression

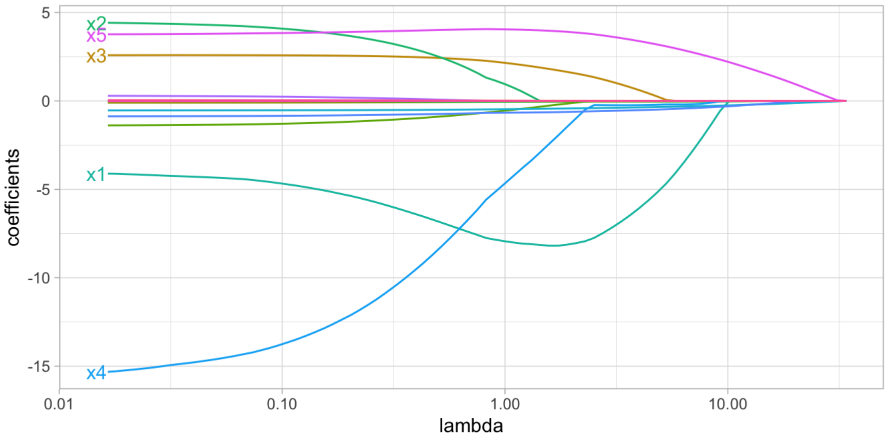
$$\min(SSE + P) = \min\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p}\beta_j^2 + \lambda_2 \sum_{j=1}^{p}|\beta_j|\right)$$

Elastic net regression combines ridge regression and LASSO regression. It has two hyperparameters, λ_1 and λ_2. Sometimes the hyperparameters are λ and mixture, which determines how much of λ to apply to each penalty (i.e. mixture $= 0$ is ridge regression and mixture $= 1$ is LASSO regression).

All variables should be centered and scaled (standardized) before estimation. Thus the coefficients will be in standardized units.

Elastic net regression can perform feature selection, but in a less dramatic fashion than LASSO regression.

Elastic net blends Ridge regression and LASSO regression



Source: Hands on Machine Learning with R