Regression Analysis

Model Selection

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Regularized Regression: Penalties

About This Lesson

Bias-Variance Tradeoff

Prediction Risk: Measure of the Bias-Variance Tradeoff

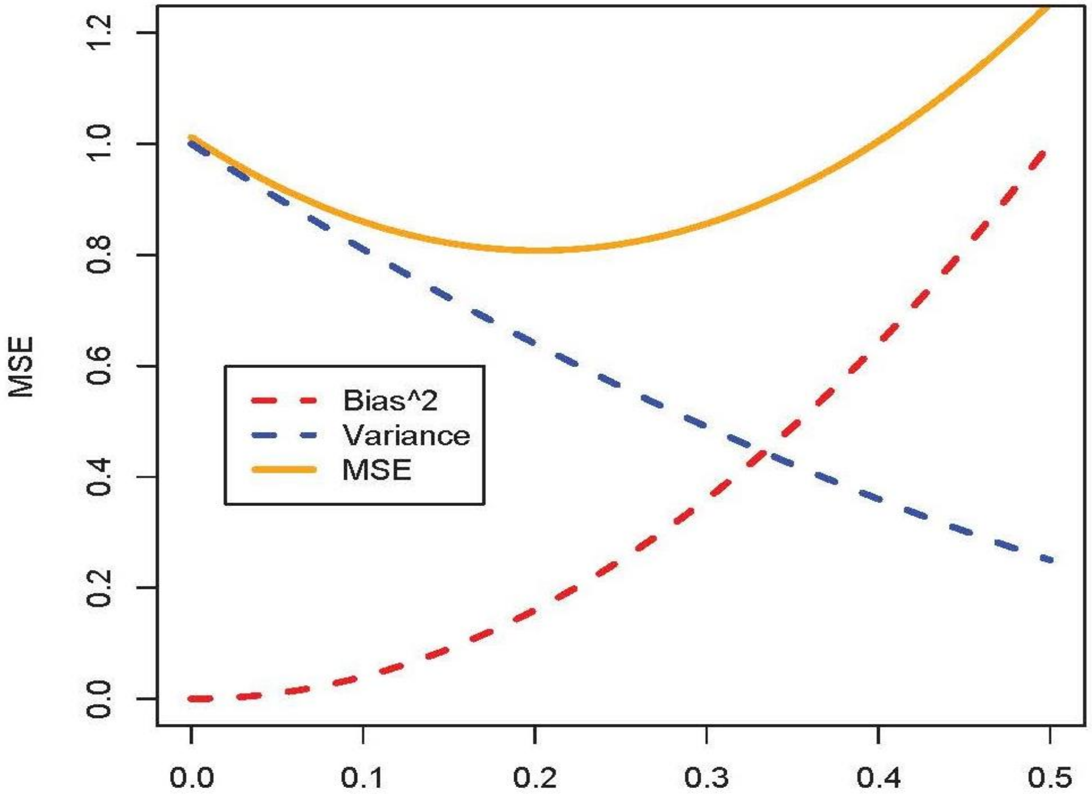
$$R(S) = \frac{1}{n}\sum_{i=1}^{n} E\left(\hat{\boldsymbol{Y}}_i(S) - \boldsymbol{Y}_i^*\right)^2$$

$$= \overbrace{V(\boldsymbol{Y}_i^*)}^{\text{Irreducible error}} + \overbrace{Bias^2\left(\hat{\boldsymbol{Y}}_i(S)\right) + \mathrm{V}(\hat{\boldsymbol{Y}}_i(S))}^{\text{Mean Square Error}}$$

for a submodel S, with $\hat{\boldsymbol{Y}}_i(S)$ the fitted response for model S and $\hat{\boldsymbol{Y}}_i^*$ the future observation.

- It is possible to find a model with lower MSE than the full model!
- It is “generic” in statistics: introducing some bias often yields in a decrease in MSE.

Bias-Variance Tradeoff



Biased Regression: Penalties

Not all biased models are better.

We need a way to find "good" biased models!

- Penalize large values of βs jointly
  - Should lead to "multivariate" shrinkage of the vector $\boldsymbol{\beta}$

- Goal is really to penalize "complex" models
  - Heuristically, "large" is interpreted as "complex model"
    - If truth really is complex, this may not work!
      - It will then be hard to build a good model anyways

Regularized Regression

Without Penalization

Estimate (β_0, β_1,…, β_p) by minimizing the sum of squared errors

$$\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})\right)^2$$

With Penalization

Estimate (β_0, β_1,…, β_p) by minimizing the penalized sum of squared errors

$$\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})\right)^2 + \lambda Penalty(\beta_1, \ldots, \beta_p)$$

The bigger l, the bigger the penalty for model complexity.

Regularized Regression (cont'd)

The penalized sum of squared errors:

$$Q(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \right) \right)^2 + \lambda Penalty(\beta_1, \ldots, \beta_p)$$

We consider three choices for the penalty:

$\boldsymbol{L_0}$ **penalty**

$||\beta||_0 = \#\{j\colon \beta_j \neq 0\}$ ⇒ Minimizing Q means searching through all submodels

$\boldsymbol{L_1}$ **penalty** (LASSO Regression)

$||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ ⇒ Minimizing Q forces many β_js to be zeros

$\boldsymbol{L_2}$ **penalty** (Ridge Regression)

$||\beta||_2 = \sum_{j=1}^{p} {\beta_j}^2$ ⇒ Minimizing Q accounts for multicollinearity

Comparing Penalties

- L_0 penalty
  - Provides best model given a selection criterion
  - Requires fitting all submodels
- L_1 penalty
  - Measures sparsity
- L_2 penalty
  - Easy to implement
  - Does not do variable selection

Example: Consider vectors $\boldsymbol{u} = (1, 0, \cdots, 0)$ and $\boldsymbol{v} = (\frac{1}{\sqrt{p}}, \cdots, \frac{1}{\sqrt{p}})$, both of length p.

Vector $\boldsymbol{u}$ is sparse, because it contains mostly zeros.

Using the $\boldsymbol{L}_1$ norm, we have $||\boldsymbol{u}||_1 = \sum_{i=1}^{p} |u_i| = 1$ and $||\boldsymbol{v}||_1 = \sum_{i=1}^{p} |v_i| = \sqrt{p}$.

Using the $\boldsymbol{L}_2$ norm, we have $||\boldsymbol{u}||_2 = \sum_{i=1}^{p} {u_i}^2 = 1$ and $||\boldsymbol{v}||_2 = \sum_{i=1}^{p} {v_i}^2 = 1$.

The $\boldsymbol{L}_1$ penalty rewards the sparsity of $\boldsymbol{u}$; the $\boldsymbol{L}_2$ penalty makes no distinction.

Summary