Regularized regression

Akitaka Matsuo

University of Essex

Regularized regression

- We saw linear regression in the previous lecture.
- Linear regression is BLUE for the train set, but might be overly-sensitive to the train data.
- We can adjust the problem by using penalized regression.
- Methods
  - Ridge regression
  - LASSO
  - (Elastic net)

Regularized regression, objective function

University of Essex

– Linear regression:

$$\text{argmin}_{\beta} \sum_{i} (Y_i - (\beta_0 + \sum_{j} \beta_j X_{ij}))^2$$

– Regularized regression:

$$\text{argmin}_{\beta} \sum_{i} (Y_i - (\beta_0 + \sum_{j} \beta_j X_{ij}))^2 + \lambda g(\beta_{-0})$$

– The shape of $g(\beta)$ is different across methods

Ridge regression

$$\text{argmin}_{\beta} \sum_{i} (Y_i - (\beta_0 + \sum_{j} \beta_j X_{ij}))^2 + \lambda \sum_{j} \beta^2$$

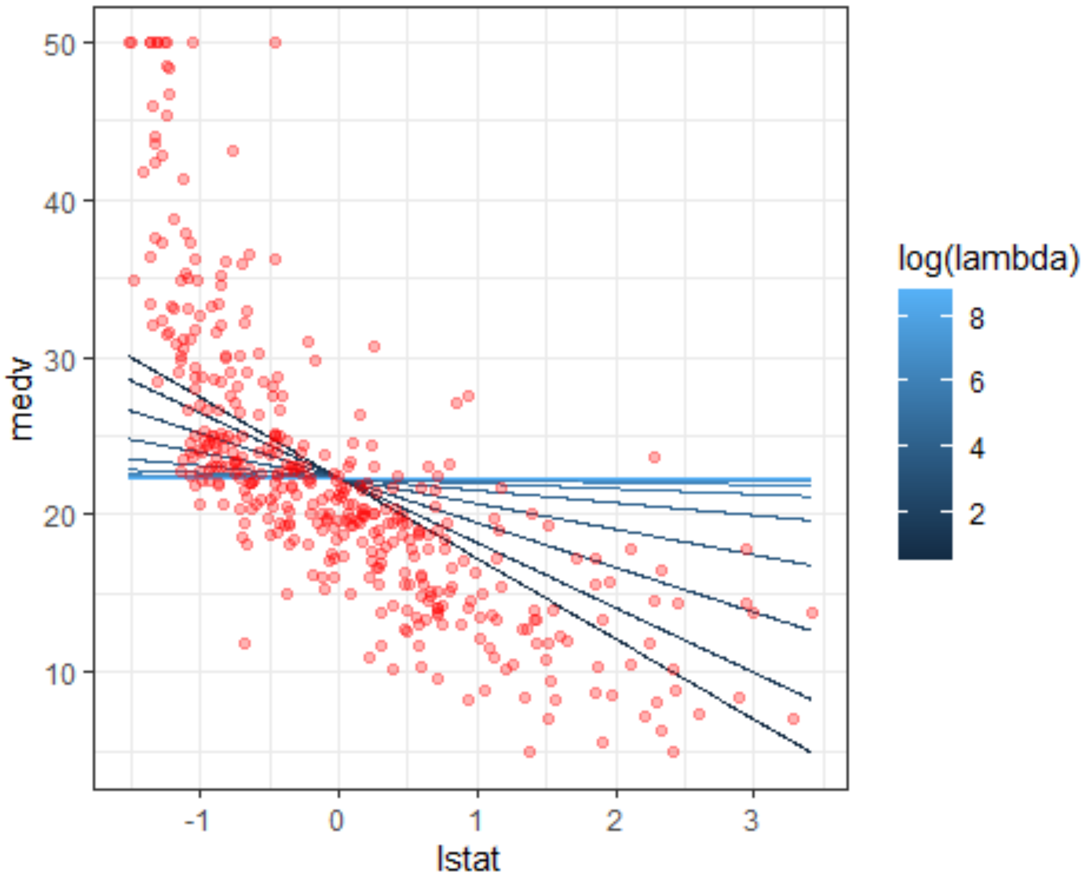
– $\lambda \sum_j \beta^2$ is the penalty term (i.e. shrinkage penalty)

- L2-penalty
- sum of the squared βs multiplied by λ
- $\lambda = 0$: OLS
- $\lambda = \infty$: completely shrunken β

– λ is an only tuning parameter in ridge regression

Ridge regression, diferent lambda

- This is an illustration of fitted line with different λ value
- When λ gets bigger, the line gets flatter
- The best λ value:
  - Enough shrinkage without too much bias (see example later)



LASSO Regression

The objective function is similar but slightly different

- LASSO

$$\text{argmin}_{\beta} \sum_i (Y_i - (\beta_0 + \sum_j \beta_j X_{ij}))^2 + \lambda \sum_j |\beta|$$

- Ridge regression

$$\text{argmin}_{\beta} \sum_i (Y_i - (\beta_0 + \sum_j \beta_j X_{ij}))^2 + \lambda \sum_j \beta^2$$

- LASSO penalty
  - L1-penalty
  - sum of the absolute value of βs multiplied by λ
  - $\lambda = 0$: OLS
  - $\lambda = \infty$: completely shrunken β
- λ is an only tuning parameter in LASSO regression

Similarity/Difference between Ridge and LASSO

Similarity

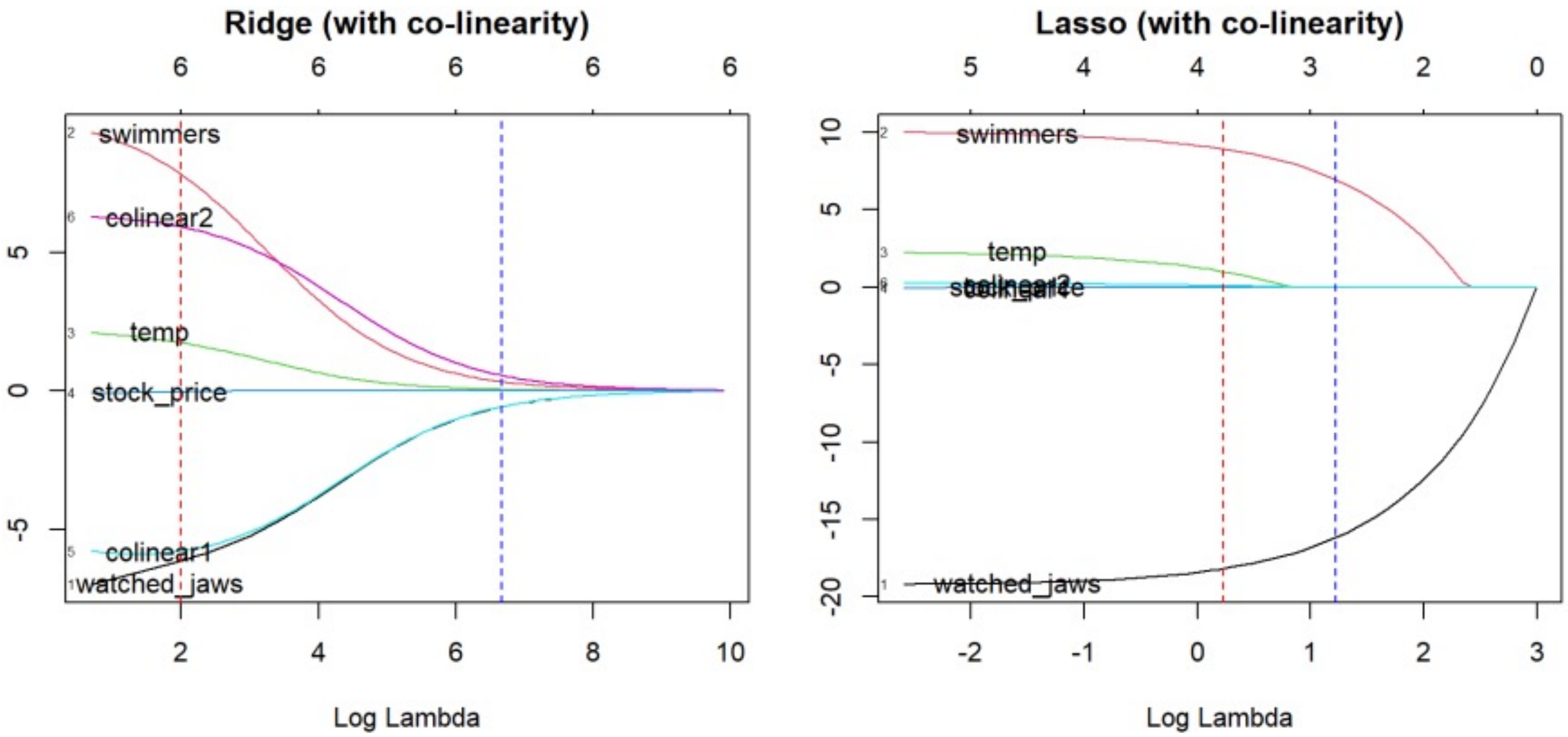
- Both penalize
- Can be estimated when more variables than observations

Differences

- The way of shrinking:
  - Ridge: Make all beta smaller but rarely gets to 0
  - LASSO: Quickly shrink β for meaningless variables to 0

So, Ridge is powerful when a lot of weak/meaningful predictors, while LASSO is useful when a lot of junk variables. That's why LASSO is used for variable selection.

Regularization paths for LASSO and Ridge



From: https://www.r-bloggers.com/2020/06/understanding-lasso-and-ridge-regression/

Elastic net

- Elastic net is the combination of Lasso and ridge regression with both L1 and L2 norm
- One formulation is:

$$\text{argmin}_{\beta} \sum_{i} (Y_i - (\beta_0 + \sum_{j} \beta_j X_{ij}))^2 + \lambda(\alpha \sum_{j} |\beta| + (1-\alpha) \sum_{j} \beta^2)$$

- Two tuning parameters:
  - α: weight of L1 and L2
    - $\alpha = 1$: LASSO
    - $\alpha = 0$: Ridge regression
- If tuned well, could perform the best

Summary

- Regularized regression: Methods to reduce model variance
- Two methods:
  - Ridge regression
    - Shrink everything smaller, basically keep all variables
  - LASSO regression
    - Variable selection