Regularized Regression

DS 301

Iowa State University

announcements: HW 6 due Wed.
find team for final project (3 is ok).

Model Selection Summary

\# of predictors (p)

$p < 30$

subset selection

(exhaustive search)

$M_1, M_2, M_3, \ldots, M_p$

$p \geq 30$

forward, backward, stepwise

(greedy algorithms)

$M_1, M_2, M_3, \ldots, M_p$

(1) indirect estimate test MSE

(2) directly estimate test MSE.

Model Selection Summary

(1) indirect estimate of test MSE

AIC, BIC, adjusted R^2, Mallow's Cp.

↳ functions of RSS + penalty / weight for predictors in your model.

(2) direct estimate of test MSE.

- validation set approach

  train → model selection

  test → evaluate $M_1, M_2, \ldots, M_p$ (and compute test MSEs)

- K-fold CV → optimal model size.

Regularized Regression (Shrinkage Methods)

'Modern' Regressions

lm(·).

- So far, when we fit a model we use the least squares approach.
- Alternatively, we might want to use another fitting procedure instead of least squares.
- These procedures can yield better prediction accuracy and model interpretability.

Least squares estimation

Recall for least squares, we are trying to find:

$$\hat{B}_{LS} = \min_{B} \sum_{i=1}^{n} \left(y_i - (B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_p X_p) \right)^2$$

Advantages of least squares:

- easy to implement
- analytical solution
- inference is well studied.
- unbiased estimates.
- easy to extend/understand for more complicated settings. (non-linear)

Alternative to least squares

- As an alternative, we can fit a model containing all p predictors using a technique that **constrains** or **regularizes** the coefficient estimates.
- In effect, this technique will **shrink** some of the coefficient estimates towards zero. (introducing bias)
- The two best-known techniques for shrinking the regression coefficients towards zero are:
    1. Ridge regression
    2. The lasso

Intuition behind shrinking regression coefficients

We know from the bias/variance decomposition that low bias situations lead to high variance. This in turn, can lead to a high test MSE.

↳ introduce a little bias
to see if we can decrease our variance.

↳ if decrease in variance is larger
than our increase in bias
→ lower test MSE.

This is motivation for shrinkage methods
(regularized regression).

Ridge regression

We want to find $\hat{\beta}^R$ that minimizes

$$\hat{B}^R = \min_{B} \left(\sum_{i=1}^{n} (y_i - (B_0 + B_1 x_{i1} + \cdots + B_p x_{ip}))^2 + \underline{\lambda \sum_{j=1}^{p} B_j^2} \right)$$

1st term: RSS (assessing quality of fit)

2nd term: $\underbrace{\text{shrinkage}}_{l_2 \text{ penalty}}$ penalty

it has the effect of shrinking B's toward 0.

↳ controlled by $\lambda > 0$.

Tuning parameter

$$\hat{\beta}^R = \min_{\beta} \left(\overset{(1)}{\sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p))^2} + \lambda \overset{(2)}{\sum_{j=1}^{p} \beta_j^2} \right), \quad \lambda > 0$$

- λ is referred to as our tuning parameter.
- It controls the relative balance of these two terms: modulates the tradeoff between fit and shrinkage.
- $\lambda = 0$: $\hat{B}^R$ just defaults back to least squares.
- $\lambda \to \infty$: penalty term will dominate
  $\to \hat{B}^R \to 0$.

◦ every value of λ will give you different $\hat{B}^R$.

Tuning parameter

- Each value of λ will give you a different set of coefficient estimates.
- Selection good λ is critical to ridge regression.

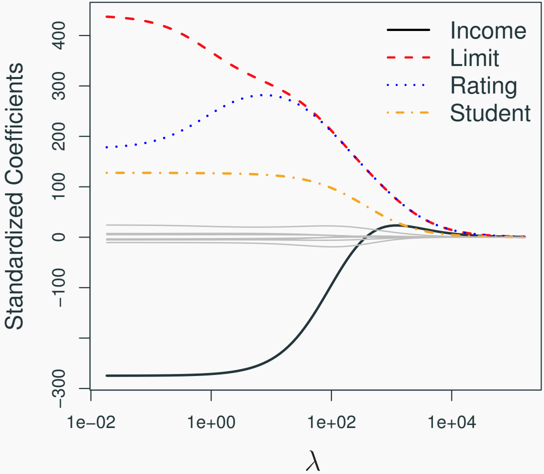
Some things to keep in mind

- The penalty term is applied to the regression coefficients $(\beta_1, \beta_2, \beta_3, \ldots, \beta_p)$ but not the intercept.
  - We want to shrink the estimated association of each predictor with the response; however, we do not want to shrink the intercept.
- Before applying ridge regression, we must standardize the predictors so they are all on the same scale.
- As we increase λ, does the model become more flexible or less flexible?

$\lambda \uparrow \rightarrow$ bias $\uparrow \rightarrow$ less flexible model.

How to choose λ?

Each λ will give us a different set of regression coefficients.



How to choose λ?

Use cross-validation:

1. Choose a grid of λ values. 0, . . . 10,000 .
2. Compute the cross-validation error for each value of λ.
3. Select the value of λ for which the cross-validation error is the smallest.
4. Refit the model using all data and the selected λ. This is your final model.

ridge regression

See R script: `shrinkage_methods.R`