S&DS 365 / 565

Data Mining and Machine Learning

Model Selection

① Find which model to pick from multiple models

use cross-validation

② For Linear Regression, find which features to pick

called Variable Selection

Linear regression
Logistic regression
KNN
k means clustering
Neural Network
Decision Tree
Random Forest
Boosting

Yale

Variable selection

Data: n observations, p predictors

- Use all predictors?

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

Methods

① Regularization

Lasso address this by l_2-norm regularizer

② enumerate all possible combinations of p predictors : 2^p

Computationally intractable for $p > 20$ or 30

Variable selection

Data: n observations, p predictors

- Use all predictors?

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

- Total number of possible subsets of variables to include: 2^p.

Variable selection

Data: n observations, p predictors

- Use all predictors?

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

Issues

- Total number of possible subsets of variables to include: 2^p.
- Bias-variance tradeoff in number of predictors included.

predictors ↑ Bias ↓ variance ↑

Variable selection

Data: n observations, p predictors

- Use all predictors?

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

- Total number of possible subsets of variables to include: 2^p.
- Bias-variance tradeoff in number of predictors included.
- More complex models are less interpretable.

For Neural Network ① keep # of edges too be small

② Big company has lots of CPU (computation power) training multiple Neural Network → pick a subset of best performance → randomly delete 90% edges → retrain → pick best one

Approaches to feature selection

1. Subset selection – use a "good subset" of the p predictors
2. Shrinkage – use all p predictors but encourage more coefficients to be near 0
3. Dimension reduction – condense the set of predictors by projecting to a lower subspace $\mathbb{R}^p \to \mathbb{R}^k$

Best-subset selection

Options range from null model $\mathcal{M}_0$ (no predictors) to full model $\mathcal{M}_p$ containing all p predictors.

Best-subset selection

Options range from null model $\mathcal{M}_0$ (no predictors) to full model $\mathcal{M}_p$ containing all p predictors.

- Fit $\mathcal{M}_0$.

Best-subset selection

Options range from null model $\mathcal{M}_0$ (no predictors) to full model $\mathcal{M}_p$ containing all p predictors.

- Fit $\mathcal{M}_0$.
- For $k = 1, 2, \ldots, p$, identify the best model $\mathcal{M}_k$ using k of the p predictors judged via training error.

Caveat: For k is large, pick best M_k is computationally intractable

Best-subset selection

Options range from null model $\mathcal{M}_0$ (no predictors) to full model $\mathcal{M}_p$ containing all p predictors.

- Fit $\mathcal{M}_0$.
- For $k = 1, 2, \ldots, p$, identify the best model $\mathcal{M}_k$ using k of the p predictors judged via training error.
  - e.g. for regression: use RSS, R^2

↑ Residual Sum of Square (Least Squares)

Best-subset selection

Options range from null model $\mathcal{M}_0$ (no predictors) to full model $\mathcal{M}_p$ containing all p predictors.

- Fit $\mathcal{M}_0$.
- For $k = 1, 2, \ldots, p$, identify the best model $\mathcal{M}_k$ using k of the p predictors judged via training error.
  - e.g. for regression: use RSS, R^2
  - e.g. for classification: use misclassification error, deviance

↑ logistic loss

Best-subset selection

Options range from null model $\mathcal{M}_0$ (no predictors) to full model $\mathcal{M}_p$ containing all p predictors.

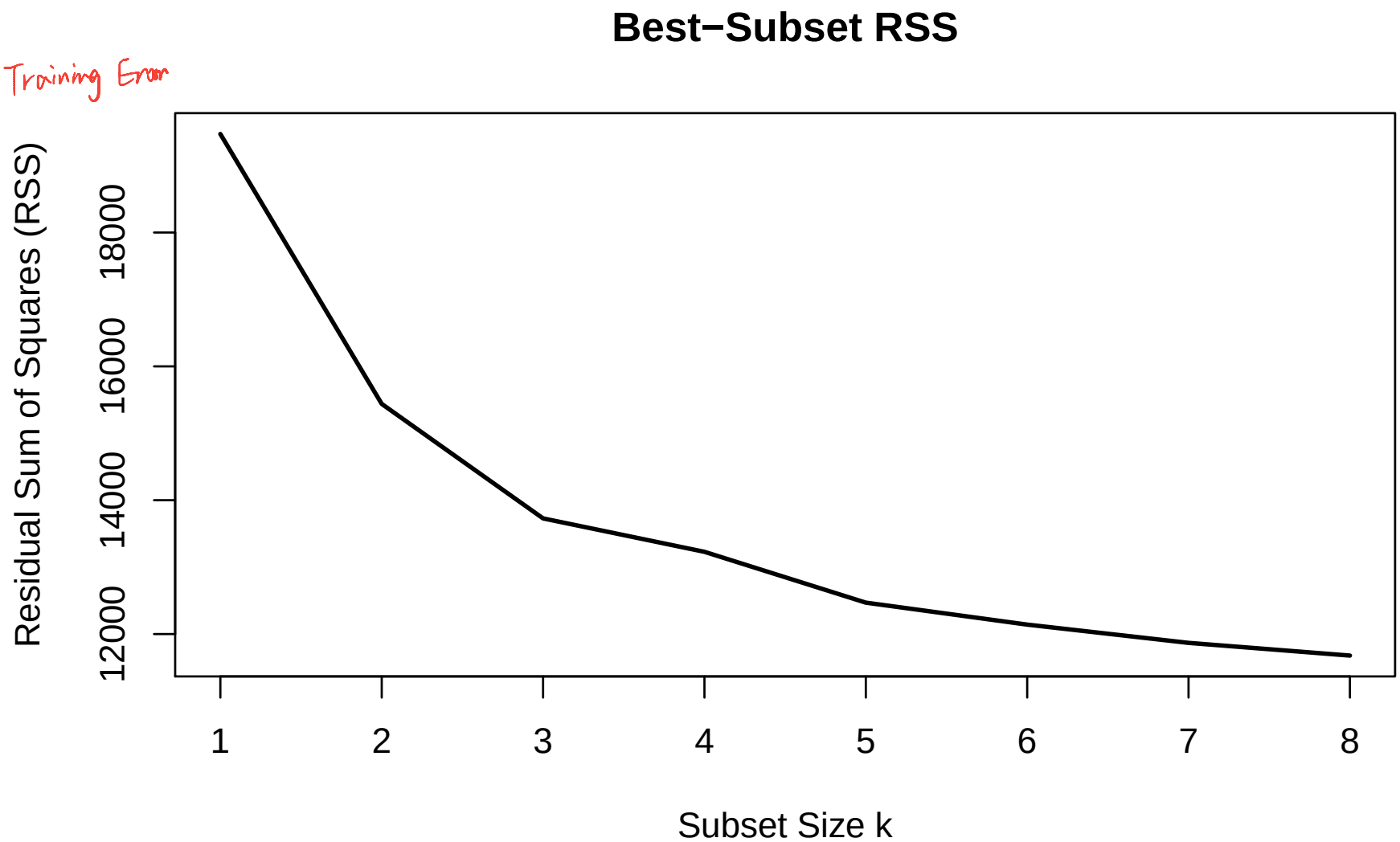
- Fit $\mathcal{M}_0$.
- For $k = 1, 2, \ldots, p$, identify the best model $\mathcal{M}_k$ using k of the p predictors judged via training error. → coz all models using k features
  - e.g. for regression: use RSS, R^2
  - e.g. for classification: use misclassification error, deviance
- Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ on basis of cross-validated prediction error. → can't use training error coz when # features ↑ traing error always ↓

Best-subset selection

Housing Price

Boston dataset:

Training Error



Best-subset selection

Not feasible when p is large. There are 2^p models to consider!

e.g. $p = 5 \Rightarrow 2^5 = 32$ models

$p = 10 \Rightarrow 2^{10} = 1024$ models

$p = 100$?

Subset selection: Stepwise selection

Recall Boosting "Forward Stagewise additive modeling"
can think Boosting as a type of linear regression model, where features are all possible weak learners

1. Forward stepwise selection

Starting from the null model, build an increasing sequence of *nested models*.

Subset selection: Stepwise selection

1. Forward stepwise selection

Starting from the null model, build an increasing sequence of *nested models*.

- Start with $\mathcal{M}_0$.
- For $k = 1, \ldots, p$, pick the best **one** of the remaining unused predictors to add to $\mathcal{M}_{k-1}$ to form $\mathcal{M}_k$.
- Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ on basis of estimated prediction error.

$\mathcal{M}_0$

based on training error

or cross validation

only one split

Note: different splits of train/test sets by cross validation may result in different best predictor, compute fraction of best predictor

$$\frac{\#\ X_k \text{ is the best predictor}}{\#\ \text{splits}}$$

→ choose best predictor with largest fraction

Subset selection: Stepwise selection

2. Backward stepwise selection

Starting from the full model, build a decreasing sequence of *nested models*.

Subset selection: Stepwise selection

2. Backward stepwise selection

Starting from the full model, build a decreasing sequence of *nested models*.

- Start with $\mathcal{M}_p$.
- For $k = p-1, p-2, \ldots, 0$, pick the worst **one** of the existing predictors to remove from $\mathcal{M}_{k+1}$ to form $\mathcal{M}_k$.
- Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ on basis of estimated prediction error.

Subset selection: Stepwise selection

2. Backward stepwise selection

Starting from the full model, build a decreasing sequence of *nested models*.

- Start with $\mathcal{M}_p$.
- For $k = p-1, p-2, \ldots, 0$, pick the worst **one** of the existing predictors to remove from $\mathcal{M}_{k+1}$ to form $\mathcal{M}_k$.
- Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ on basis of estimated prediction error.

Note:
Backward and forward stepwise selection are more computationally feasible than best subsets, but no guarantee they'll find the best subset of the p predictors to use.

Subset selection: Stepwise selection

3. Bidirectional stepwise selection

    - Start with either the null model or the full model

Subset selection: Stepwise selection

3. Bidirectional stepwise selection

   - Start with either the null model or the full model
   - At each step, consider the impact of adding or subtracting a predictor

Subset selection: Stepwise selection

3. Bidirectional stepwise selection

- Start with either the null model or the full model
- At each step, consider the impact of adding or subtracting a predictor
- Stop when no further improvements can be made

Subset selection: Stepwise selection

3. Bidirectional stepwise selection

- Start with either the null model or the full model
- At each step, consider the impact of adding or subtracting a predictor
- Stop when no further improvements can be made
  - Will not cycle because model must reduce RSS

Subset selection: Stepwise selection

3. Bidirectional stepwise selection

- Start with either the null model or the full model
- At each step, consider the impact of adding or subtracting a predictor
- Stop when no further improvements can be made
  - Will not cycle because model must reduce RSS
  - Will eventually stop because only finitely many models

Subset selection: Stepwise selection

3. Bidirectional stepwise selection (Forward–backward selection)

used for high dimension space eg. $n = 100$ $p = 1000000$ Full model is obviously useless
Design Matrix X not have full rank, the ordinary least square solution won't be unique

- Start with either the null model or the full model

Forward: add p features

- At each step, consider the impact of adding or subtracting a predictor

add a feature, training error decrease by Δ
then see if delete a feature, training error increase $< \frac{\Delta}{2}$

- Stop when no further improvements can be made
  - Will not cycle because model must reduce RSS
  - Will eventually stop because only finitely many models
  - Could run for more than $O(p)$ steps

Scoring metrics for final step

- *RSS* is a bad metric to use (as is multiple R^2). (Why?)

Scoring metrics for final step

- *RSS* is a bad metric to use (as is multiple R^2). (Why?)
- Cross-validated MSE is a good criterion, but is time consuming.

Scoring metrics for final step

Because big model has advantage

- *RSS* is a bad metric to use (as is multiple R^2). (Why?)
- Cross-validated MSE is a good criterion, but is time consuming.
- Other options? { Mallow's Cp, AIC, BIC }

cross validation

based on penalty : regularization

use penalty (regularizer) to account for the additional error you get by adding more features

Obviously add more feature will decrease training error, regularizer can find if the improvement come from overfitting to noise or actual signal gain

Scoring metrics for final step

often used for regression model { linear, logistic

- Mallow's C_p (regression only) — linear
- AIC (regression or classification) — Stat's module will give for linear regression
  Akaike info criterion
- BIC (regression or classification) — Bayesian info criterion

Mallow's C_p

$$C_p = \frac{1}{n}(RSS + 2p\widehat{\sigma}^2),$$

where p is the number of coefficients fitted and $\widehat{\sigma}^2$ is estimated error variance.

Mallow's C_p

pick model with smallest C_p

$$C_p = \frac{1}{n}(RSS + 2p\hat{\sigma}^2),$$

RSS: training error; $2p\hat{\sigma}^2$: penalty

$\hat{\sigma}^2 = \frac{RSS}{n-p_{full}} = \frac{1}{n-p_{full}}\sum_{i=1}^{n}(y_i - \hat{y}_{i,full})^2$

where p is the number of coefficients fitted and $\hat{\sigma}^2$ is estimated error variance. Derivation Setup:

unbiased estimate of noise

Data: (X, Y), X is $n \times p$ and Y is $n \times 1$

Fitted model: $\hat{Y} = X\hat{\beta}$

assume X is fixed

randomness of $\hat{\beta}$ just come from noise

new data: some X with new label $\tilde{Y}$ (new noise)

Consider how well our model predicts $(X, \tilde{Y})$, measured via out-of-sample MSE:

$$MSE_{OOS} = E\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_i - \hat{Y}_i)^2\right]$$

$\tilde{Y}_i$ and $\hat{Y}_i$ are independent

Mallow's C_p

$$MSE_{OOS} = E\left[\frac{1}{n}\sum_{i=1}^{n}(\widetilde{Y}_i - \widehat{Y}_i)^2\right]$$

For any i,

$$\begin{aligned} E[(\widetilde{Y}_i - \widehat{Y}_i)^2] &= Var(\widetilde{Y}_i - \widehat{Y}_i) + (E[\widetilde{Y}_i - \widehat{Y}_i])^2 \\ &= Var(\widetilde{Y}_i) + Var(\widehat{Y}_i) - 2Cov(\widetilde{Y}_i, \widehat{Y}_i) + [E(\widetilde{Y}_i) - E(\widehat{Y}_i)]^2 \end{aligned}$$

Mallow's C_p

Decompose MSE_{OOS}

$$MSE_{OOS} = E\left[\frac{1}{n}\sum_{i=1}^{n}(\widetilde{Y}_i - \widehat{Y}_i)^2\right]$$

pull out just focus on

For any i,

$Var(X) = E(X^2) - [E(X)]^2$

$$E[(\widetilde{Y}_i - \widehat{Y}_i)^2] = Var(\widetilde{Y}_i - \widehat{Y}_i) + (E[\widetilde{Y}_i - \widehat{Y}_i])^2$$
$$= Var(\widetilde{Y}_i) + Var(\widehat{Y}_i) - 2Cov(\widetilde{Y}_i, \widehat{Y}_i) + [E(\widetilde{Y}_i) - E(\widehat{Y}_i)]^2$$

$\widetilde{Y}_i$, $\widehat{Y}_i$ is independent

$Cov(\widetilde{Y}_i, \widehat{Y}_i) = 0$

Note that $Cov(\widetilde{Y}_i, \widehat{Y}_i) = 0$, so:

$$E[(\widetilde{Y}_i - \widehat{Y}_i)^2] = Var(\widetilde{Y}_i) + Var(\widehat{Y}_i) + [E(\widetilde{Y}_i) - E(\widehat{Y}_i)]^2$$

Mallow's C_p

In-sample (IS) MSE:

$$MSE_{IS} = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2\right]$$

Mallow's C_p

In-sample (IS) MSE:

$$MSE_{IS} = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2\right]$$

For any i,

$$\begin{aligned} E[(Y_i - \widehat{Y}_i)^2] &= Var(Y_i - \widehat{Y}_i) + (E[Y_i - \widehat{Y}_i])^2 \\ &= Var(Y_i) + Var(\widehat{Y}_i) - 2Cov(Y_i, \widehat{Y}_i) + [E(Y_i) - E(\widehat{Y}_i)]^2 \end{aligned}$$

Mallow's C_p

In-sample (IS) MSE:

$$MSE_{IS} = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2\right]$$

For any i,

$$\begin{aligned} E[(Y_i - \widehat{Y}_i)^2] &= Var(Y_i - \widehat{Y}_i) + (E[Y_i - \widehat{Y}_i])^2 \\ &= Var(Y_i) + Var(\widehat{Y}_i) - 2Cov(Y_i, \widehat{Y}_i) + [E(Y_i) - E(\widehat{Y}_i)]^2 \end{aligned}$$

dependent coz we train on Y_i to get $\widehat{Y}_i$

Note Y_i and $\widetilde{Y}_i$:

- are independent
- have the same distribution, e.g., $Var(Y_i) = Var(\widetilde{Y}_i)$ and $E(Y_i) = E(\widetilde{Y}_i)$

Mallow's C_p

i-th term in the summation of MSE_{OOS} again:

$$E[(\widetilde{Y}_i - \widehat{Y}_i)^2] = Var(\widetilde{Y}_i) + Var(\widehat{Y}_i) + [E(\widetilde{Y}_i) - E(\widehat{Y}_i)]^2$$

Mallow's C_p

i-th term in the summation of MSE_{OOS} again:

$$\begin{aligned} E[(\widetilde{Y}_i - \widehat{Y}_i)^2] &= Var(Y_i) + Var(\widehat{Y}_i) + [E(Y_i) - E(\widehat{Y}_i)]^2 \\ &= E[(Y_i - \widehat{Y}_i)^2] + 2Cov(Y_i, \widehat{Y}_i) \end{aligned}$$

Averaging over all i, we get:

$$\frac{1}{n} E\left[\sum (\widetilde{Y}_i - \widehat{Y}_i)^2\right] = \frac{1}{n} E\left[\sum (Y_i - \widehat{Y}_i)^2\right] + \frac{2}{n} \sum Cov(Y_i, \widehat{Y}_i)$$

Mallow's C_p

i-th term in the summation of MSE_{OOS} again:

$$\begin{aligned} E[(\widetilde{Y}_i - \widehat{Y}_i)^2] &= Var(Y_i) + Var(\widehat{Y}_i) + [E(Y_i) - E(\widehat{Y}_i)]^2 \\ &= E[(Y_i - \widehat{Y}_i)^2] + 2Cov(Y_i, \widehat{Y}_i) \end{aligned}$$

Averaging over all i, we get:

$$\frac{1}{n} E\left[\sum (\widetilde{Y}_i - \widehat{Y}_i)^2\right] = \frac{1}{n} E\left[\sum (Y_i - \widehat{Y}_i)^2\right] + \frac{2}{n} \sum Cov(Y_i, \widehat{Y}_i)$$

We can show that $\sum Cov(Y_i, \widehat{Y}_i) = \sigma^2 p$.

Mallow's C_p

i-th term in the summation of MSE_{OOS} again:

$$E[(\widetilde{Y}_i - \widehat{Y}_i)^2] = Var(Y_i) + Var(\widehat{Y}_i) + [E(Y_i) - E(\widehat{Y}_i)]^2$$
$$= E[(Y_i - \widehat{Y}_i)^2] + 2Cov(Y_i, \widehat{Y}_i)$$

Link out of sample error with in sample error

Averaging over all i, we get:

$$\frac{1}{n}E\left[\sum(\widetilde{Y}_i - \widehat{Y}_i)^2\right] = \frac{1}{n}E\left[\sum(Y_i - \widehat{Y}_i)^2\right] + \frac{2}{n}\sum Cov(Y_i, \widehat{Y}_i)$$

We can show that $\sum Cov(Y_i, \widehat{Y}_i) = \sigma^2 p$.

In summary,

C_p

test error = train error + adjustment

$$MSE_{OOS} = MSE_{IS} + \frac{2p\sigma^2}{n}$$

Mallow's C_p

We approximate MSE_{LS} using RSS/n and σ^2 using $\widehat{\sigma}^2$.

$$C_p = \frac{RSS}{n} + \frac{2p\widehat{\sigma}^2}{n}.$$

- Adjusts RSS with a penalty that depends on number predictors and variance of error term.
- If $\widehat{\sigma}^2$ is unbiased estimate of σ^2, then C_p is an unbiased estimate of test MSE.

Mallow's C_p

We approximate MSE_{LS} using RSS/n and σ^2 using $\widehat{\sigma}^2$.

$$C_p = \frac{RSS}{n} + \frac{2p\widehat{\sigma}^2}{n}.$$

- Adjusts RSS with a penalty that depends on number predictors and variance of error term.
- If $\widehat{\sigma}^2$ is unbiased estimate of σ^2, then C_p is an unbiased estimate of test MSE.

To summarize, choose model with lowest C_p.

AIC

Akaike Information Criterion (regression or classification):

$$AIC = -2 \log L + 2p,$$

where L is the likelihood of the model. p: # predictors

We can show for linear regression,

$$-2 \log L = \frac{RSS}{\widehat{\sigma}^2} + C,$$

for some constant C. Hence,

$$AIC = \frac{1}{n\widehat{\sigma}^2}(RSS + 2p\widehat{\sigma}^2).$$

Very similar to C_p

BIC

Bayesian Information Criterion (regression or classification):

$$BIC = -2 \log L + p \log(n).$$

For regression,

$$BIC = \frac{1}{n}(RSS + \log(n) p \hat{\sigma}^2).$$

How do AIC and BIC compare?

- Penalty on AIC: $2p$ constant penalty
- Penalty on BIC: $\log(n)p$ as $n \uparrow$ penalty $\uparrow$

BIC

Bayesian Information Criterion (regression or classification):

$$BIC = -2 \log L + p \log(n).$$

For regression,

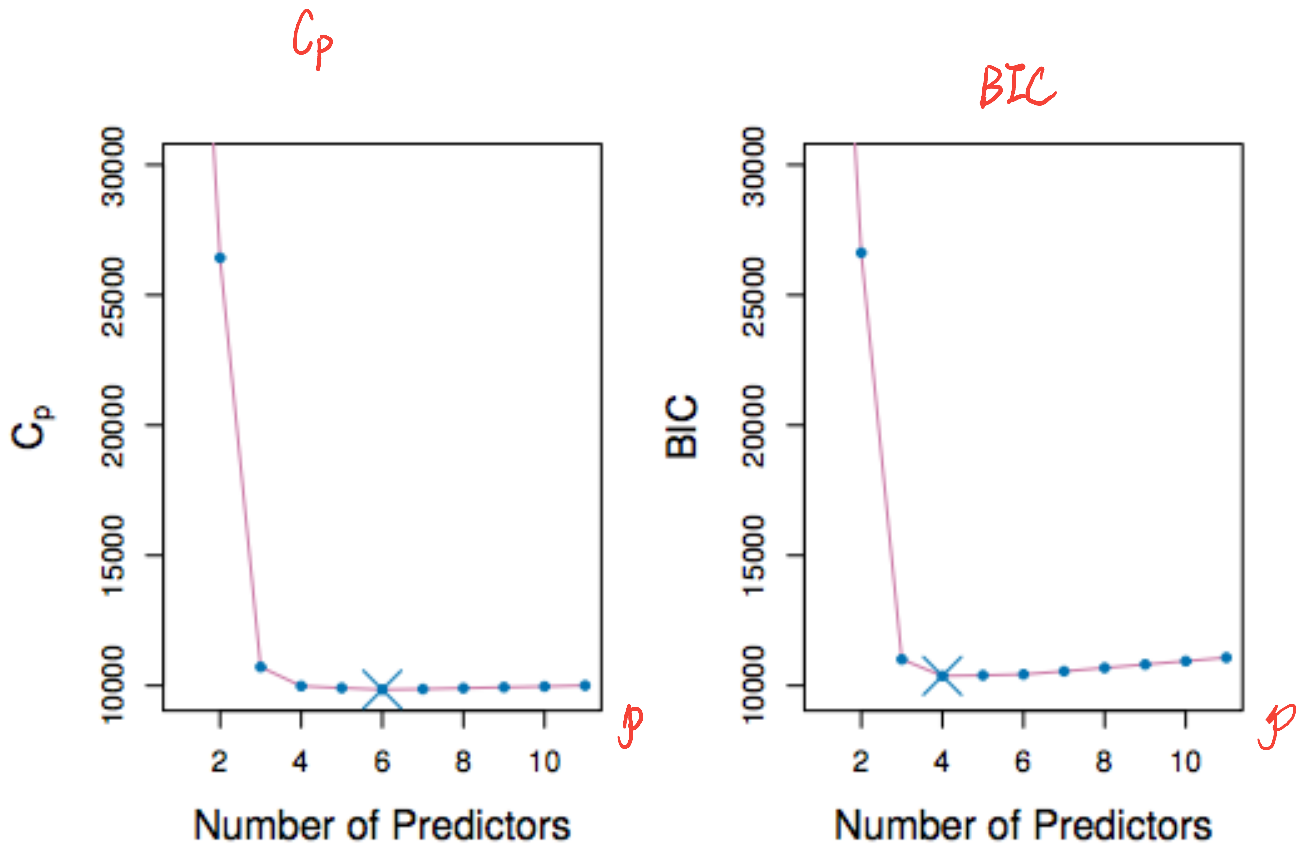
$$BIC = \frac{1}{n}(RSS + \log(n) p \widehat{\sigma}^2).$$

How do AIC and BIC compare?

- Penalty on AIC: $2p$
- Penalty on BIC: $\log(n)p$
- $\log(n) > 2$ for $n > 7$

BIC has heavier penalty on number of variables, produces smaller models.

Comparison


C_p
C_p
10000
15000
20000
25000
30000
2
4
6
8
10
Number of Predictors
p
BIC
BIC
10000
15000
20000
25000
30000
2
4
6
8
10
Number of Predictors
p

Summary

- We like models that minimize expected test error.
- Cross-validation is nice, but requires a lot of computation.
- Train/Test split is also nice, but when n is small, you are wasting data for test
- Stepwise model selection allows us to pick a model based on some measure of expected test error using in-sample measures like C_p, AIC, or BIC.
- After variable selection, must be careful about doing inference. See research on **post selection inference**