S&DS 355 / 365 / 565

**Data Mining and Machine Learning**

# Model Selection

Thursday, September 19th

Yale

# Outline for next topics

SDS 355 won't have to know some of this

- Model selection
- Ridge regression
- The lasso

# What did we talk about last time?

- Stochastic gradient descent is a first order (first derivatives) method that scales to large classification and regression problems
- Cross validation is a practical way of estimating the variability of test error.

# Variable selection

Data: $n$ observations, $p$ predictors

- Use all predictors?

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

# Variable selection

Data: $n$ observations, $p$ predictors / independent variables

- Use all predictors?

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

include / not include

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

↓

expensive

# 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.

you don't want to overfit, so we try to reduce the number of parameters in model

# 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.

# Approaches to feature selection

similar to HW

(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

# 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.

# 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, $R^2$

# 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

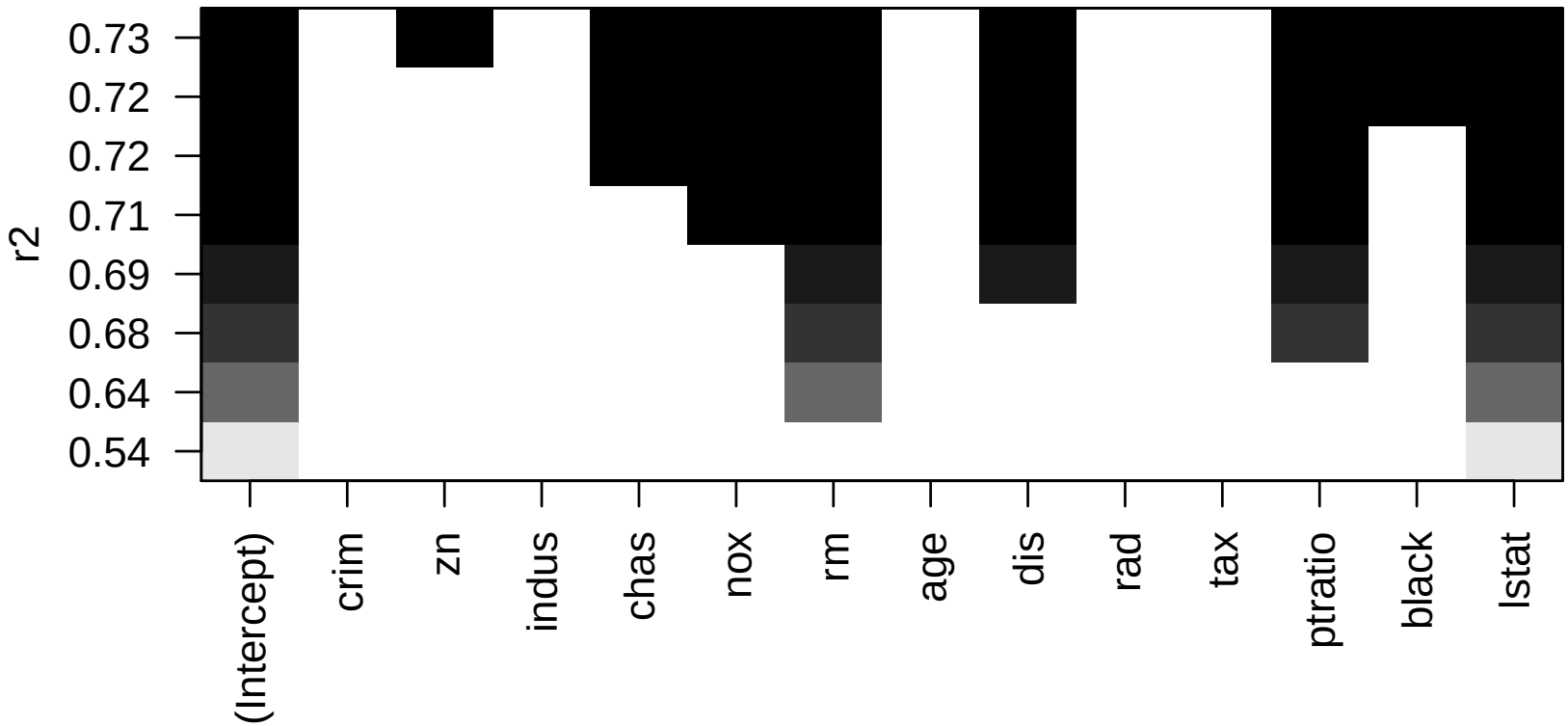
# Best-subset selection

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
- Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ on basis of cross-validated prediction error.

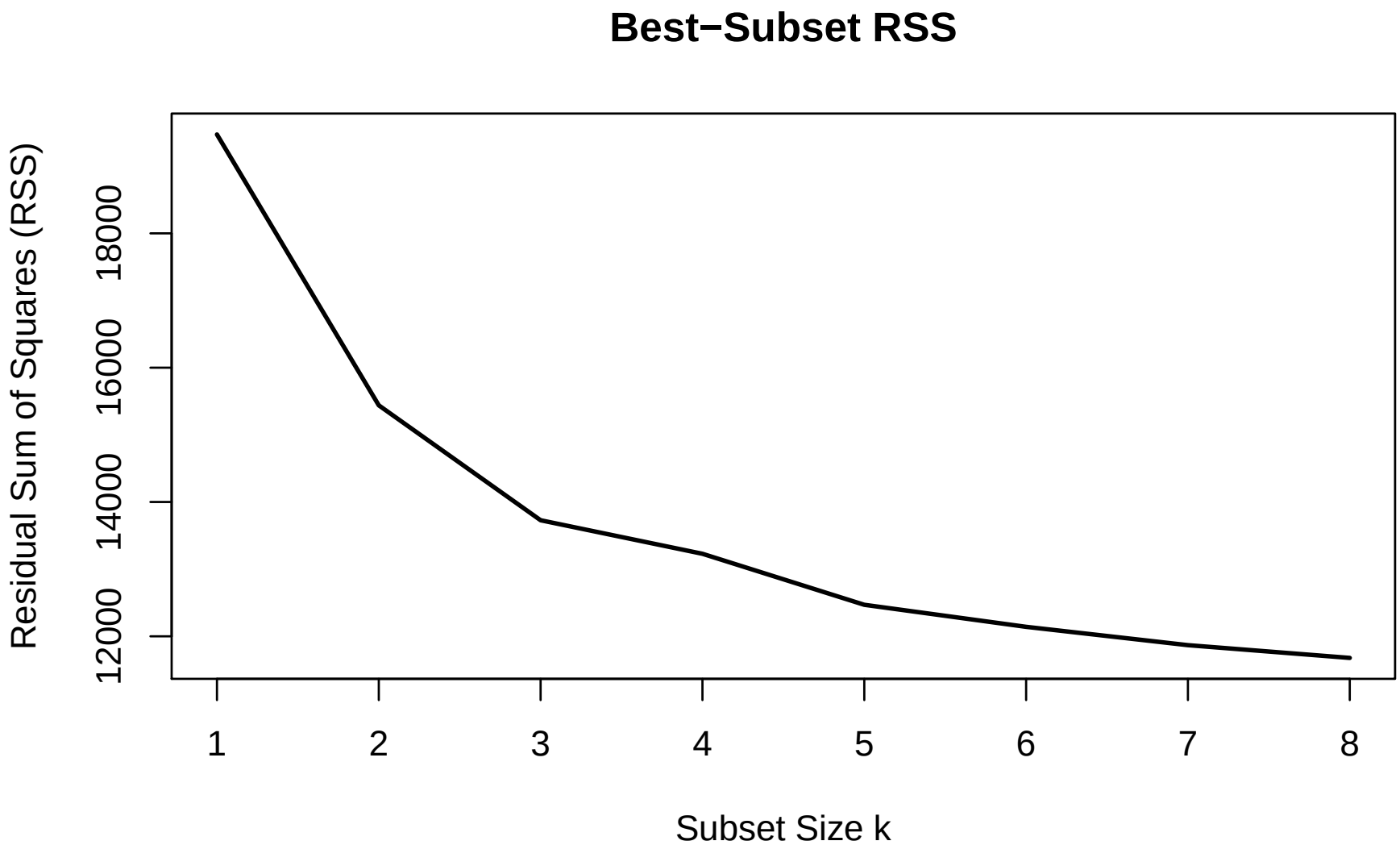
# Best-subset selection

```
library(MASS)
library(leaps)
m1 <- regsubsets(medv ~ ., data=Boston, nbest=1, method="exhausti
```



# Best-subset selection

Boston dataset:



# 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

Instead of computing all combinations, just take a greedy approach

1. Forward stepwise selection

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

greedy selection takes a locally optimal decision at every step but often this works out.

# Subset selection: Stepwise selection

pick the covariate that gives the best model, and keep doing this in a loop.

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.

# 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.

At the end

model with $p = 1$

# 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.

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.

because greedy choices are not necessarily globally optimal

# 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

- 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
  - Could run for more than $O(p)$ steps

# Scoring metrics for final step

i.e selecting between $M_1 \ldots M_p$

- *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

- *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?

# Scoring metrics for final step

- Mallow's $C_p$ (regression only)
- AIC (regression or classification)
- BIC (regression or classification)

# 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 $C_p$

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

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

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

Derivation not needed for 355

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

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

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

# 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 $C_p$

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

For any $i$,

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

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 $i$,

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

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:

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

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

In summary,

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

# Mallow's $C_p$

We approximate $MSE_{LS}$ using $RSS/n$ and $\sigma^2$ 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 $\sigma^2$, then $C_p$ is an unbiased estimate of test MSE.

# Mallow's $C_p$

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

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,

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

How do AIC and BIC compare?

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

# BIC

Bayesian Information Criterion (regression or classification):

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

For regression,

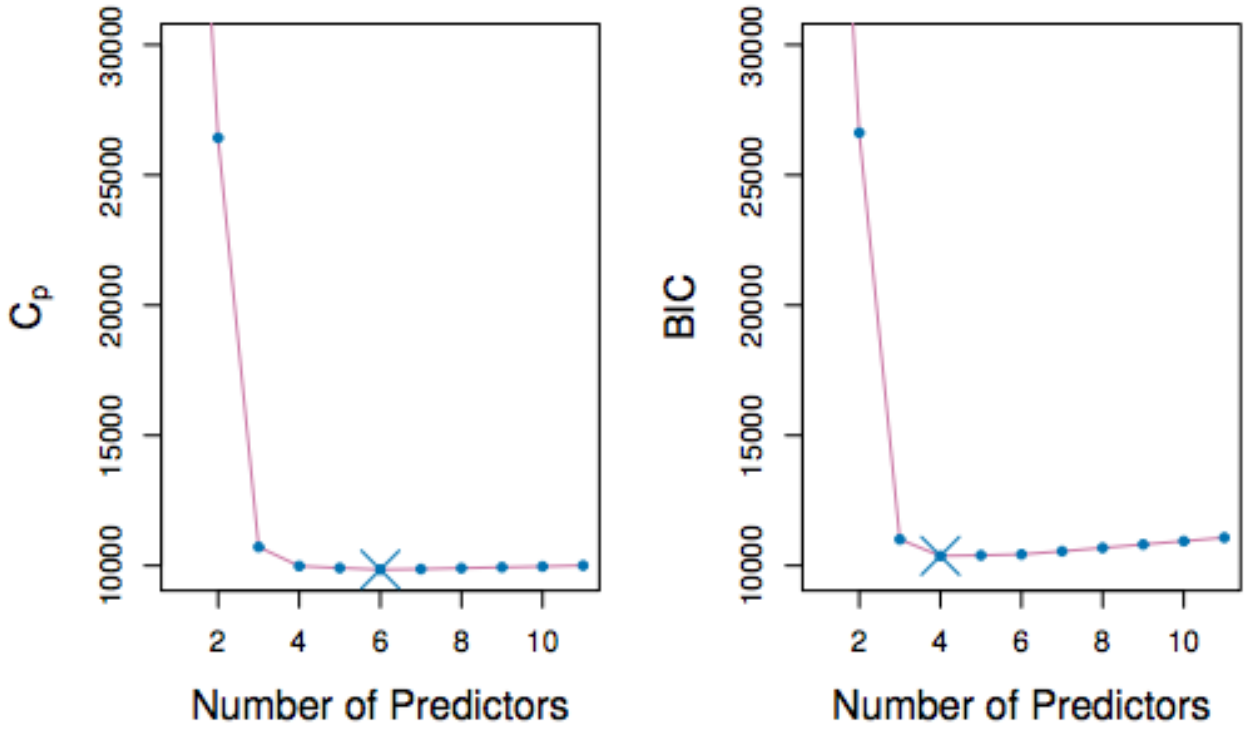
$$BIC = \frac{1}{n}(RSS + \log(n) p \widehat{\sigma}^2).$$

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


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

# Summary

- We like models that minimize expected test error.
- Cross-validation is nice, but requires a lot of computation.
- 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 $BIC$.

# Next up: Shrinkage and selection

An alternative to variable selection is to simply use all predictors, but impose a penalty on the magnitude of the coefficients.

- Ridge regression

Result is that the coefficients get *shrunken* towards 0 and standard error of coefficients is much lower.

The Lasso combines shrinkage and selection.